Mechanisms for Enforcing Honest Signaling

Christoher Ahern Robin Clark
Department of Linguistics
University of Pennsylvania
Philadelphia, PA 19104

September 21, 2012

1 Introduction

Central to the study of meaning is the assumption that speakers and hearers prefer truth. Utterances totally unhinged from their truth conditions would seem to alter language beyond recognition or, for that matter, use. Grice (1975) goes so far as to enshrine this preference in his Maxim of Quality:

(1) **Quality**
Try to make your contribution one that is true.

1. Do not say what you believe to be false.
2. Do not say that for which you lack adequate evidence.

Yet, Grice's formulation assumes that agents behave as thought their interests were totally aligned. Is there, however, any compelling reason to believe this is so? The evolution of cooperation has been a perennial question in biology and the social sciences for many years (Nowak (2006); Bowles and Gintis (2011), and citations). If we presume that language is inherently cooperative with regard to truthfulness we mask the distinction between two problems for the evolution of stable communication (Scott-Phillips, 2008, 2010): *reliability*, and *honesty*.

Signals are *reliable* if they are correlated with some aspect of the signaler or the environment, such that this correlation provides information that benefits the receiver. Signals can become reliable as part of a convention (Lewis,

1969). Once such a conventional meaning is established, they can also be used to reliably convey implicatures in a given context (Parikh (2001); Benz et al. (2006); Parikh (2010); Clark (2012), among others). However, it is by no means guaranteed that interests are aligned in such a way as to allow reliability to be usefully exploited.

This brings us to the problem of *honesty*. In cases where interests are not perfectly aligned, it will happen that one agent has reason to misinform another—either about the state of the world or his intended actions. Excluding pathological behavior, the agent providing the false information presumably benefits from the other agent believing it. In these cases, we wouldn't expect the agent to be much interested in obeying the Maxim of Quality. Cooperation would seem to require that signalers be not just reliable, but honest, particularly where interests are not perfectly aligned. We take honesty to be absence of deception and adopt the following definition (Searcy and Nowicki, 2005):

(2) Deception occurs when a Sender sends a signal to a Receiver,

1. the receiver responds in a way that benefits the sender,
2. the response is appropriate if the signal reliably indicates a situation, X, and
3. it is not the case that X obtains.

How do we know whether our interests are aligned in any particular instance? Once language becomes suspect in some instances, shouldn't the suspicion spread so that language eventually becomes useless as a signaling device (Zahavi, 1993). Without honesty, signaling would be counter-selected for. If receivers could not rely on signals—if signals were often deceptive—they would tend to ignore signals, so attending to them would incur a cost. If receivers do not attend to signals, then signalers should not bother sending them, since sending them would incur a cost without benefit to them. Thus, without honesty, it would be surprising to find signaling systems at all. Communication via conventional signaling should not exist in populations with divergent interests. In this sense, dishonesty would seem to lead to a collapse in the meaning market (Akerlof, 1970). Insofar as we are by and large honest and language is meaningful, we might ask what allows it to be so.

In this paper, we will consider cases where rational agents might have incentive to be less than honest and ask how truthfulness—in particular,

the first Submaxim of Quality—fares under such circumstances. In the next section we consider signaling in games of partial common interest and show how meaning is undermined by dishonesty. We then turn to mechanisms for enforcing honest signaling. We find that a simple sort of memory might allow for the enforcement of the first submaxim, while something like gossip might suffice in more complicated and realistic situations. Throughout, we will aim to define the limits of the different cognitive and social mechanisms, pushing towards their limits, and where they yield insight into how honest signaling might have arisen and been maintained.

2 Signaling with Partial Common Interest

Common to all organisms is the problem of efficiently allocating resources. Here we consider how an agent might choose to allocate resources based on signals received from other agents in a population with only partial common interests (Rabin and Sobel, 1996; Blume et al., 2001). That is, we examine a situation where some agents have perfectly coincident interests, but others do not. We show how meaning is undermined in this context.

2.1 Signaling in a Population

To begin we will take the viewpoint of a single receiver in a population with possibly divergent interests. Let us call him Ralph. Ralph has a certain amount of energy to go looking for and find food. He must use these resources wisely to guarantee his continued survival. Now, suppose another agent approaches Ralph. We will call her Sally. Sally suggests one of two things. The first being that the two of them scrounge around in some bushes to find fruit that has dropped off of trees. It takes two to gather the fruit efficiently, one pulling back the branches of the underbrush and the other grabbing the fruit. It is tedious work, but not too bad as they take turns moving branches and eating what they can reach. The pieces of fruit that fall to the ground are not the best quality, so Sally can also suggest that Ralph let her stand on his shoulders to grab the good pieces higher up in the trees. It takes more effort to lift Sally up than to pull back the branches, but the food in the tree is far tastier than that on the ground. At this point Ralph has two options: he can ignore the suggestion and continue searching for food on his own, or he can attend to the signal.

Suppose that Sally has certain intentions or abilities. For example, she

might only be asking Ralph to lift her up so that she can stuff her face without dropping any down for him. She might also be limited in her abilities to effectively gather fruit for both of them, e.g. she might have short arms. She might also be able to and intend to get as much fruit as possible for both of them. We will refer to the former case as a dishonest Sally, and the latter as an honest Sally. When it comes to foraging, she cannot abuse Ralph's trust in the same way. They take turns getting fruit and pulling back branches. Sally, honest or dishonest, behaves in exactly the same way.

So, how should Ralph respond to Sally's signal? We gain further insight when we consider the preferences of both. If Sally is honest and able, then Ralph prefers to lift her up into the tree as it guarantees better fruit for both of them. It would also be preferable to forage with Sally if she is honest rather than ignore her entirely. This could be the case for two reasons. First, if Ralph forgoes interacting with an honest Sally, he might end up interacting with a dishonest Sally. Second, he might observe her working with someone else to get the tasty fruit, and regret his decision. An honest Sally has exactly the same preferences as Ralph. If Sally is dishonest, then Ralph would prefer foraging with her to lifting her up into the tree. But, he would ultimately prefer to ignore her entirely. This could be due to two things. First, by ignoring a dishonest Sally, Ralph leaves himself available to work with an honest Sally. Second, Ralph might also be relieved that he chose not to work with a dishonest Sally if he sees her duping another agent. A dishonest Sally would prefer joint foraging to being ignored. If a dishonest Sally is lifted up into a tree, she gets to eat some fruit, but there are risks. If Ralph discovers her dishonesty and drops her, or otherwise retaliates against her, she might be worse off despite getting some fruit. This would seem reasonable if Ralph actively monitors her progress.

2.2 Signaling Games

To formalize these notions we introduce signaling games and a utility structure that conforms to our discussion above. A signaling game is a tuple of the following form $\langle \{S, R\}, T, \delta, M, A, U_S, U_R \rangle$. There are two roles in the game: S is the sender, and R is the receiver. T is a set of states and δ is a probability distribution over those states, $\delta \in \Delta(T)$. As per our discussion above, there are two types of signalers. Honest signalers, t_h will gather fruit from the tree for both. Dishonest signalers, t_d, will eat without sharing any benefits. M is the set of messages available to the sender. A is the set of

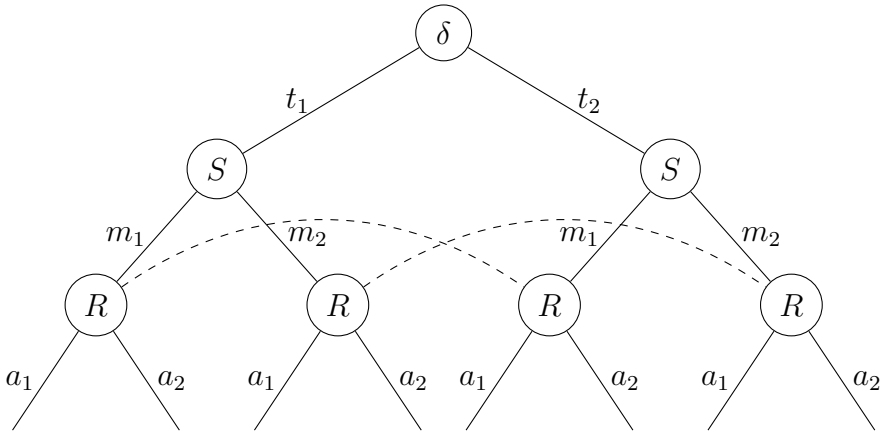

Figure 1: Structure of Signaling game with two states, messages, and actions.

actions available to the receiver. Receivers can ignore the signal entirely, or they can invest some effort to forage in the bushes or a great deal to lift the sender up to get fruit from the tree. U_S and U_R are the payoff functions of the Sender and Receiver, respectively, which define a preference over the possible outcomes of combined strategies.

The game proceeds as follows. First, the sender observes a state $t \in T$, determined by the probability distribution δ. The state here is the type that the sender happens to be. The sender then chooses a message $m \in M$ based on a strategy $s \in [T \to M]$ which is then transmitted to the receiver. Finally, the receiver takes an action $a \in A$ based on the message received and a strategy $r \in [M \to A]$. The payoffs for the Sender and Receiver are determined by the type t, message sent $s(t)$, and action taken $r(s(t))$.

The structure of a signaling game with two states, messages, and actions is shown in Figure 1. The root node represents the probability distribution over the states as determined by nature. The sender then chooses a message at one of the nodes labeled S. Finally, the receiver chooses an action at one of the nodes labeled R. The dashed lines indicate R's uncertainty as to what state of affairs actually holds, that is, the type of the sender.

In what follows it will be cumbersome to deal with the full expressions of utility. To aid discussion, we use numbers that satisfy the preferences

outlined above, without any commitment to the particular numbers beyond their ordinal ranks.[1]

	a_0	a_{pe}	a_e
t_h	0,3	5,5	8,8
t_d	0,6	5,5	3,0

Figure 2: Payoffs in a Game of Partial Common Interest

2.3 Pooling and Meaning

Now we are in a position to consider signaling behavior. There are a limited number of initial states from which a population could begin. If both types send the same signal, then receivers should respond with action a_{pe}.[2] If the types send different signals, then receivers should respond to the signal sent by type t_h with a_e and the signal sent by t_d with a_0. That is, receivers should work with honest senders and ignore dishonest senders. However, in response to this dishonest senders should adopt the same signal as honest senders. Once both types send the same signal, receivers do best by taking action a_{pe}. In fact, this state of affairs where both types send the same message and are met by an intermediate effort is a a *pooling* equilibrium.[3] The signals from all types are indistinguishable, and no agent can benefit from unilateral deviation.

Signaling in this equilibrium is not dishonest *per se*, but it is the direct result of dishonesty when starting from states where the different types send

[1]For a more detailed development of the payoff structure, see the appendix.

[2]The expected utility of a_0 decreases with the proportion of honest signalers in the population, the expected utility of a_{pe} remains constant, and the expected utility of a_e increases. Thus, if the proportion of honest signalers in the population exceeds a certain threshold then it is always best for receivers to take action a_e. Analogously, if the proportion of honest signalers falls below a certain threshold then it is best for receivers to always play a_0. In what follows, we will assume populations composed in such a way that a cautious intermediate investment is always the best pure strategy without specific knowledge.

[3]In fact, it is a set of pooling equilibria, one for each message in M. In what follows we will use the singular when referring to this set.

different signals. Moreover, in exactly these cases dishonesty undermines the initial reliable correlation between the signals and an attribute of the sender. In this sense, dishonesty undermines meaning. At best, signals come to carry only the vacuously true"I am honest or dishonest". This result seems dismal. Not only does honest signaling never arise, but a great deal of potential utility is lost. Yet pooling is certainly not the last word. In the next section we turn to the sorts of mechanisms that might enforce honest signaling.

3 Possible Mechanisms to Maintain Honest Signaling

Here we will examine the impact of neologisms, punishment, memory, and gossip on the maintenance of honest signaling with the payoff structure outlined above. We wish to determine if any of these additional socio-cognitive mechanisms might keep signaling from the pooling equilibrium. Of central concern will be the payoffs of the different roles in comparison to the pooling state. We are concerned with both the qualitative and quantitative effect of the different mechanisms on honest signaling.

3.1 Neologisms

It might be that, once a signal has been undermined by dishonest signalers, the honest signalers could, as a group, come to use a new form that would unambiguously signal their (honest) intentions. This form could be a "neologism" or it could be an old, disused signal that is co-opted for the purposes of honest signaling. This would fix the problem, at least until the dishonest signalers noticed what the honest signalers were doing and began doing likewise, thus undermining the new system. It's easy to see that the solution is a temporary one at best, but it's always worth showing this more formally. In this subsection, we'll attack the problem from two directions; we will give a formal proof, using Markov analysis, and we will show the result using simulations. We'll turn to the simulations first.

We start with a population of 10,000 agents arranged on a lattice wrapped around a torus. In our simulation, each agent plays with eight other agents drawn at random from the population. Each agent plays twice as a sender and once as a receiver. That is, each agent has a strategy that specifies what to do if they are of type t_h as well as what to do if they are of type t_d. Each agent adopts the best sender and receiver strategies from the set of agents it interacts with, ties being broken randomly. The process of copying strategies

involves a certain amount of error. For example, if an agent observes another agent doing better by sending m_2 when t_d, then the observing agent should copy this behavior. With some probability the observing agent will make a mistake in copying this new behavior. This error rate will result in new signals becoming available.

Dishonest signalers have no reason to innovate; they do best when they are mistaken for honest signalers, so they prefer to use the "known" signal for honesty. Honest signalers will do well at first—the system will start out in a separating state—but once the dishonest signalers have gamed the system, they will receive only the pooling payoff; the system will fall into the pooling state. Honest signalers can do much better by adopting a new signal that comes to be associated with honesty. By mutation, such a signal will eventually be found and because the honest signaler does better, that signal will begin to propagate through the honest population: The system returns to the separating state. But, of course, the dishonest signalers will eventually discover the new signal and the system will eventually fall back into the pooling state.

Figure 3 shows the messages used over time by honest signalers (the top graph in Figure 3) and by dishonest signalers (the bottom graph in Figure 3) in a single representative simulation. Notice that the honest signalers begin by using message m_1 (red); dishonest signalers increase their use of message m_1, but the honest agents quickly abandon m_1 in favor of m_3 (blue). The dishonest agents quickly notice that honest signalers are doing better using m_3 and follow them. The honest signalers then abandon m_3 in favor of m_1 adn the process continues. This prompts the dishonest signalers to do so a few steps later. In other words, the bottom graph, aside from its initial segment, is just the top graph displaced in time. This is especially clear at the end of the graph, where the honest signalers have begun to use a new signal and the dishonest signalers have not yet picked it up. The dishonest signalers are chasing the honest signalers through the message space.

Figure 4 shows the receivers' responses to the various messages over time. Recall that honest signalers began by using message m_3; the appropriate response to this is action a_1, which is, indeed, how receivers respond after a brief lag (bottom panel). We can follow the fate of honest signalers by following the red line across the three different message graphs. Message m_3 is the initial winner, but it soon fades in favor of message m_1. As we saw in Figure 3, honest signalers stop using message m_3 and begin using message m_1. The dishonest signalers soon adopt m_1 and we see that receives

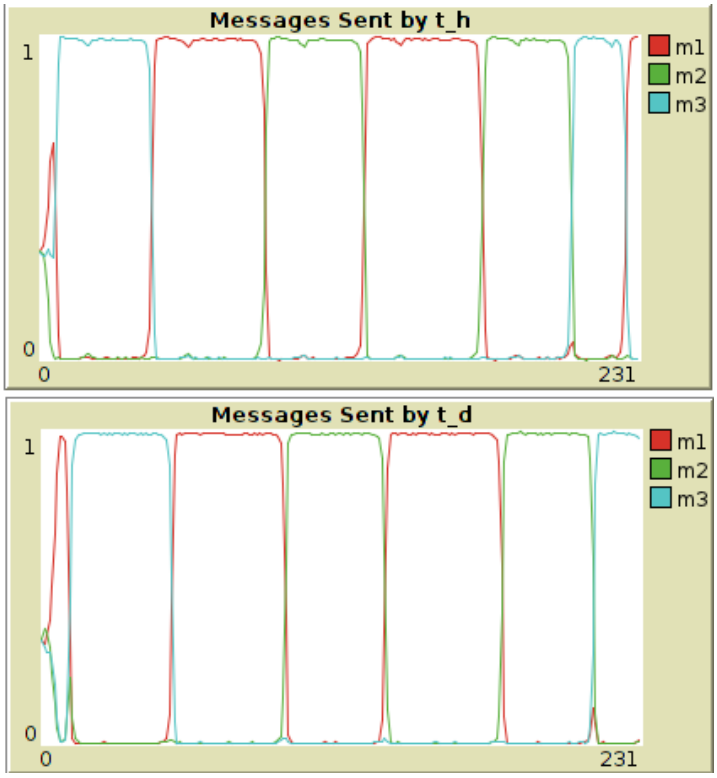

Figure 3: Messages sent over type by honest (t_h) and dishonest (t_d) agents

stop responding to it with action a_1 (and turns to a_3, the pooling action). Honest signalers move on to m_2, which receivers respond to with action a_1. In general, the red line (the optimal response to honest signalers) is followed by the blue line (the pooling response) and then followed by the green line (the optimal response to dishonest signalers). Throughout, we see the separating state collapsing into the pooling state and then reemerging in a different form. This can be seen most clearly in Figure 5.

Figure 5 is the proportion of honest signalers using a given message minus the proportion of dishonest signalers using that same message. The peaks in the graph correspond to times when honest signalers are easily distinguished from the dishonest signalers—that is, a separating state. The flat valleys

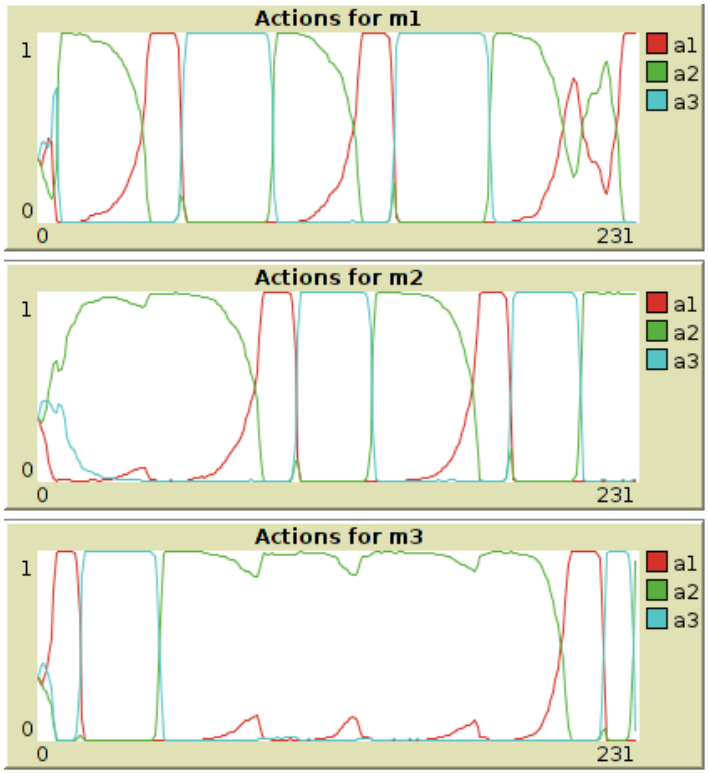

Figure 4: Receiver responses (actions) to signals

correspond to times with the honest signalers cannot be distinguished from the dishonest signalers—that is, the pooling state. Notice the initial peak occurs when the honest signalers are using message m_3, shown in green. This collapses into a flat valley; the honest signalers then use message m_1, shown in red, while the dishonest signalers are still using message m_3. This new separating state eventually collapses as well. Eventually, the honest signalers adopt message m_2, in green. All the messages are used in turn; they periodically fall into disuse and then reemerge as the signal used to signal honestly, and are subsequently undermined. Notice that at the end we see m_1 emerging as the signal of honesty.

It's easy to see that the system will wander between separating states

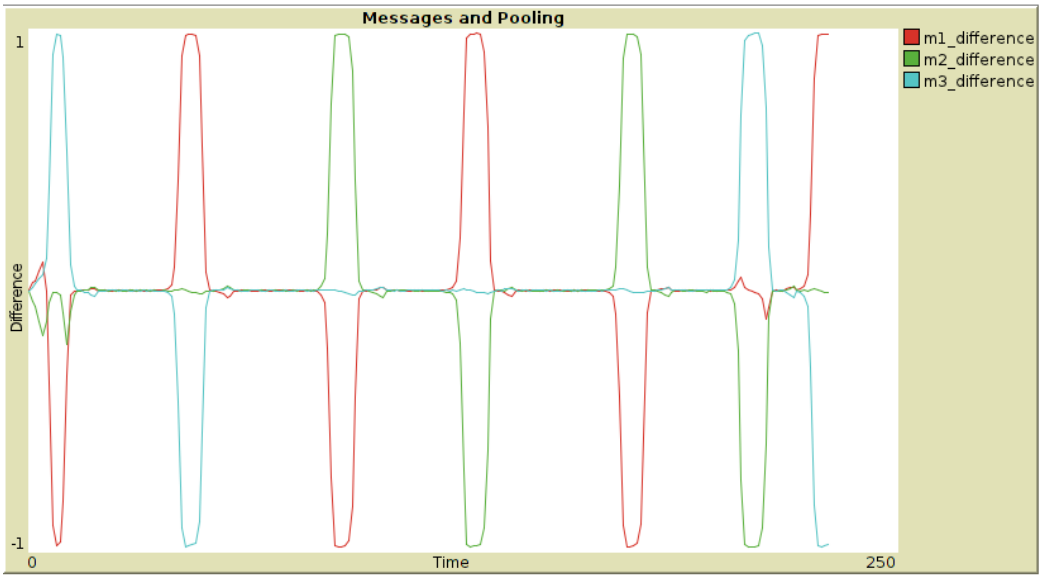

Figure 5: Separating and pooling states over time

and pooling states. In itself, this might not be so bad, if the system spends enough time in separating states to make up for the utility lost while in pooling states. The question, then, is how much time is spent in the two states; our simulations do not really answer this question. To do so, we turn to a Markov analysis of the system.

While the introduction of new signals, neologisms, into the game destabilizes the pooling equilibria, it does nothing to ensure honest signaling; it takes quite a bit of running to come back to the same place.[4] We first consider why the pooling equilibria are not stable, and what predictions this allows us to make as to the state of signaling at any given time. Even under generous assumptions regarding the ability of honest types and receivers, we find that honest signaling is fleeting.

As it stands, neologisms destabilize the pooling equilibria, but we wish to know where things go from there; do they allow the system to escape into a separating states? We characterize the dynamics of the game as a Markov

[4]None of the pooling equilibria are neologism proof in the sense of Farrell (1993).

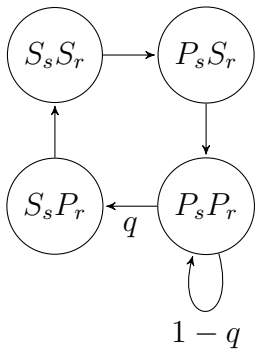

Figure 6: State Diagram for Markov Process with Neologisms

process and note that the system spends the majority of time in pooling states. To see this clearly, first note that states of the sender population can be grouped into two sets: *separating* states where the types send different messages (S_s) and *pooling* states where the send the same message (P_s). Similarly, we might classify the receiver strategies into those that take unique actions for each message used in a separating state (S_r) and those that take the identical action for two or more distinct messages (P_r). Define the states of the Markov process as combinations of the different sender and receiver strategies.

Now, let q be the probability that an honest type will employ a neologism when in the pooling state. Assume that if an agent can benefit materially from recognizing and exploiting a change in the environment, that it will.[5] We can thus represent the process as in Figure 6. Imagine starting in the pooling state, P_sP_r. With probability q the honest types start sending a previously unused message; that is, the system transitions to the state S_sP_r. When this new signal is used receivers recognize it and act accordingly, returning the system to a separating state; the system returns to the state S_sS_r. However, this return is temporary. In the separating state the dishonest type benefits from noticing the new signal and the process starts anew.

[5]More generally we might assume that the probability of noticing is a monotonic function of the possible increase in payoff. This leaves open the relative speed of recognition for the types. These issues aside, we note that once a new signal is used, the receiver will

	S_sS_r	P_sS_r	P_sP_r	S_sP_r
t_h	8	8	5	5
t_d	0	3	5	5
R	$8 \times \delta(t_h) + 6 \times \delta(t_d)$	$8 \times \delta(t_h)$	5	5

Figure 7: Utilities of agents in different states

Intuitively, when q, the probability of senders adopting a neologism, is low the amount of time spent in the pooling state will be the highest. The rest of the time will be spent evenly in the other states. As q increases, the time spent in the pooling state decreases. We can find the stationary distribution of the process, which when paired with the payoffs received by different agents in the different states yields the expected utility after a certain period of time for both types and the receiver. The payoffs for agents in the various states of the process can be seen in Figure 7. Remember that $\delta(t_h)$ is the proportion of honest senders in the population and $\delta(t_d)$ that of dishonest senders; the row for R indicates the payoffs to the receiver in each of the four states.

The expected utility for the honest type and the receiver increase with q, and the expected utility of the dishonest type decreases. As q approaches 1, the amount of time spent in each state approaches $\frac{1}{4}$, and the expected utility of the various roles can be calculated.

$$\begin{aligned} EU_{t_h} &= \frac{1}{4}(8+8+5+5) = 6.5 \\ EU_{t_d} &= \frac{1}{4}(0+3+5+5) = 3.25 \\ EU_R &= \frac{1}{4}(16 \times \delta(t_h) + 6 \times \delta(t_d) + 5 + 5) = 2.5 \times \delta(t_h) + 4 \end{aligned} \tag{3}$$

We can compare these expected utilities to those received in the pooling equilibrium, column P_sP_r in Figure 7. We can also compare these to the expected utility of agents with completely honest signaling, column S_sS_r in Figure 7. Honest senders do better with neologisms than they do in the pooling equilibrium. Yet, they do not do as well as they would if signaling were

eventually recognize it and act accordingly, returning the system to a separating state.

completely honest. In contrast, dishonest signalers do worse than they would in the pooling equilibrium, but better than with honest signaling. Receivers do better with neologisms than the pooling equilibrium if $\delta(t_h) > .4$, but always do worse than with honest signaling. Note, however, that these are extreme values. If q is lower, as we might expect it to be, then the expected utilities of agents will move towards those of the pooling equilibrium.

As the system exits P_sP_r, the honest signalers do better, but there is a constant drag provided by the dishonest signalers who are pulling the system back to P_sP_r. The dishonest signalers will copy the honest signalers, and do a bit better than they would otherwise The receivers will come to ignore the signals and respond with the pooling action; P_sP_r is an attractor state from which the system cannot permanently escape. Neologisms allow honest signalers to outrun the dishonest signalers on average, but not by much. Insofar as they recapture some of the utility lost to dishonesty, it is only a slight and temporary fix. Agents do not refrain from saying what they know to be false. We continue on to examine other mechanisms.

3.2 Punishment

Experimental results suggest that punishment can support the maintenance of cooperation in social dilemmas (Fehr and Gächter (2000), Fehr and Gächter (2002), Fischbacher and Gächter (2010)). The idea that punishment encourages cooperation is appealing because punishment can be used as a deterrent to defection; in our case, dishonest signalers can be punished if their dishonesty is discovered and the punishment, if sufficiently costly, would deter dishonest signaling. As we will show, punishment, in the end, is not necessarily sufficient to enforce honest signaling.

To begin, let's suppose that when a receiver is tricked by a dishonest sender it has the option of expending some amount of effort to punish the receiver. We can suppose that the dishonest signaler is fined c; imposing this fine, of course, will cost the punisher (the wronged receiver) some amount of effort, which we can represent as a scalar on the punishment, bc. We can show how this modifies the payoff structure in Figure 2 to yield the modified payoff structure in Figure 8

Now, if $c > 3$ then punishment creates a new equilibrium. Namely, it creates a separating equilibrium where signalers of both types truthfully reveal themselves. To make this clear, we calculate the expected utility of a sender of type t_d considering deviating from the separating equilibrium. If all of

	a_0	a_{pe}	a_e
t_h	0,3	5,5	8,8
t_d	0,6	5,5	$3-c$,$-bc$

Figure 8: Payoffs in a Game of Partial Common Interest

the receivers in the population are willing to punish, and this punishment is greater than the benefit to a t_d receiver when the receiver plays a_e, then signaling will be honest. More generally, we can determine the number of punishers needed for a given level of punishment. To do this we determine when the expected utility of t_d lying is lower than the separating payoff of a_0. If $P(R_c)$ represents the proportion of punishing receivers in the population, this is true when:

$$P(R_c) > \frac{3}{c} \tag{4}$$

As c increases, the proportion of punishers required decreases. As c decreases, the proportion of punishers required increases.

However, the stability of punishment in a population over time is problematic. To see this, consider the payoff of both punishing and non-punishing receivers when the proportion of punishers is sufficient to create a separating equilibrium. Both types of receivers obtain the same payoff, and are thus neutrally stable. If the proportion of punishers drifts below the threshold required to maintain the equilibrium, then dishonesty will reemerge. Further, once the proportion of punishers falls below this threshold, it will not climb back. This is because, once dishonest signaling resumes, punishers always do worse than non-punishers. Thus, the proportion of punishers will continue to decrease. This means that even when starting with a sufficient proportion of punishers, we do not see the maintenance of honest signaling.

However, expending this extra effort is counter-selected for. If all receivers punish then deceit will be curbed, but a single receiver has an incentive to free-ride on the effort expended on punishment by other receivers. Given this, receivers might wish to distribute the costs of punishment, signaling their

intention to join in punishing. Yet, as before, signaling offers no solution. All receivers will give the signal that they will join in, but this just pushes the question of honesty to a higher order. The problem of higher-order free-riding renders punishment untenable.

Higher-order free-riders undermine cooperation at the level of punishment, which in the long run undermines the cooperation of individuals in terms of honesty. Thus, punishment cannot enforce honest signaling.

3.3 Memory

Receivers equipped with the ability to discover and remember the types of the senders with which they have interacted should be able to sort the types accurately under certain conditions. While this mechanism is inherently receiver-centric, it also impacts the payoffs of other agents. That is, as receivers learn to sort senders accurately, the whole system moves towards the payoffs of the separating state. Here we explore those conditions, focusing on the effects of population size, cost, and capacity on the expected utility of the receiver and other agents, and how this compares to the expected utilities of each in the pooling state and a system with neologisms.

To consider the effect of different parameters on the expected utility of a receiver with memory over time, we must first consider the means by which such a receiver might discover the types of senders. Imagine a receiver in a pooling equilibrium. Upon receiving a signal, suppose that this receiver decides to trust it with probability q. This receiver might be totally skeptical ($q = 0$), or particularly inquisitive ($q = 1$). Now, upon deciding whether to trust an incoming message, a sender must take an action besides the pooling action a_{pe}. Given that both types prefer to be identified as t_h, we might expect the receiver to experiment with a_e. If a receiver plays a_e and gets a payoff of 8, then it remembers that the sender is honest, if it gets a payoff of 0 that the sender is dishonest. The next time a receiver encounters the sender, it can act accordingly, receiving its preferred payoff.

We might ask how long it takes for this process to yield an expected utility greater than a certain baseline threshold. To find this, we first find the expected utility of a receiver with memory. Let K_t be the set of known senders at a given time t, and K'_t be the set of unknown receivers. Further, let $P(K_t) = \frac{|K_t|}{N}$ be the probability that a sender chosen at random from a population of size N at a given time is known to the receiver, and $P(K'_t) = \frac{|K'_t|}{N}$ the probability that a sender chosen at random is unknown. If the sender

is known, then the receiver is guaranteed the preferred payoff, which can be given as $a = 8\delta(t_h) + 6\delta(t_d)$. If the receiver is not known, then based on q, the receiver can expect to get a payoff of $b = q[8\delta(t_h) + 0\delta(t_d)] + (1 - q)5$. The expected utility of a receiver at a given point in time is then:

$$EU_R^t = P(K_t)(a) + P(K_t')(b) \tag{5}$$

Now, we might ask when the expected utility of a receiver with memory that acts in such a manner exceeds a given threshold. Let B be the baseline threshold of interest. The expected utility of a receiver with memory exceeds that of the baseline just when:

$$\begin{aligned} EU_R^t &> B \\ P(K_t) &> \frac{B - b}{a - b} \end{aligned} \tag{6}$$

Suppose that there is an even split between honest and dishonest senders in a population, $\delta(t_h) = \delta(t_d) = \frac{1}{2}$, where $N = 20$. The expected utility of a receiver with memory and $q = 1$ exceeds the baseline threshold of the pooling equilibrium, $B = 5$, when it discovers the type of seven agents.

$$|K_t| > \frac{20}{3} \tag{7}$$

Note that the number of agents that must be known depends on the population. For example, in a population of 100 the number of known agents would have to exceed 34. This requirement increases linearly with the size of the population. Given this threshold, we might also determine how much time is required to reach it. The amount of time required comes into play when we consider the lifespan of the agent. If it takes too long for memory to yield a cumulative advantage, then it will not grow in a population. If it yields such an advantage rapidly, then we might expect it to flourish.

Starting in an initially unknown population, where $q = 1$, there is no chance of encountering a known agent upon the first interaction. Thus, after the first interaction $|K_t| = 1$. Following the first interaction, the possibility of encountering a known sender arises and the expected amount of time for a receiver to discover the type of all senders is:

$$E = \sum_{i=0}^{N-1} \frac{N}{N - i} \tag{8}$$

Thus, in a population of 20 agents, split evenly between the types, the first sender will be revealed after 1 interaction, the second on average in 1.05 interactions, and so forth. Now, if a receiver is such that $q < 1$, then there will be breaks between these investigations. That is, if a receiver is not as aggressive in discovering the types of other agents then it will take longer to do so. The amount of time between discoveries is $\frac{1}{q}$. Thus, the expected time to the discovery of all types for any value of q can be given as:

$$E = \frac{1}{q} \sum_{i=0}^{N-1} \frac{N}{N-i} \tag{9}$$

That is, lower values of q extend the amount of time taken to find all of the dishonest agents. Figure 9 shows the initial and long term results of simulations of this process for various values of q with a population of $N = 350$. Note that a receiver with $q = 1$ has an initial expected utility of 4 before interacting with any senders, but this more aggressive approach quickly yields results that are better than either the pooling payoff or that of neologisms. Moreover, for all rates of q this is the case. Thus, we might expect that a receiver with memory, even a very short-lived one, would do better than one without such capacities.

We can determine the amount of time, measured in interactions, that will yield a higher cumulative expected utility than a given baseline. Let T be the amount of time. The cumulative expected utility of a receiver with memory exceeds a baseline when:

$$T > \frac{(b-a)}{(b-B)} \sum_{i=0}^{T} P(K_i) \tag{10}$$

Intuitively, we are simply adding up the area under the curves. Note the effect of large populations on this process. Namely, as we increase the population size, we increase the amount of agents that need to be discovered to surpass a given threshold and thus the amount of time to do so. Similarly, as we decrease the rate of exploration q, the amount of time increases, also necessitating a longer lifespan for memory to spread in a population.

Thus far, we have not directly incorporated the cost of memory. However, there are many ways to implement the notion of cost. Memory might involve some fixed cost, marginal costs, or some combination of the two. A rough estimate on the upper limit is the difference between the maximum payoff of

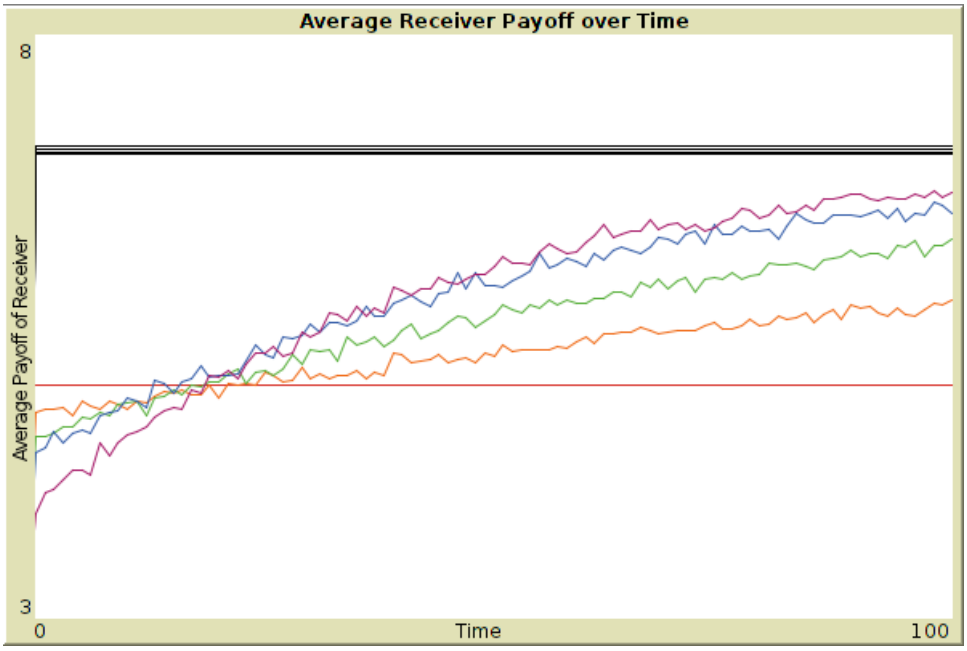

Figure 9: Expected Utility of Receiver with memory over time with different q. $N = 350$.

perfect information regarding types and the payoff of the pooling equilibrium. In an evenly split population, the payoff of perfect knowledge is $7 = \frac{1}{2} \times 8 + \frac{1}{2} \times 6$, and the pooling payoff is 5. Thus, if c is the cost of memory, then it must be the case that $c < 2$ for memory to yield any benefit. If this is not the case then a memoryless receiver that only ever played the pooling response a_{pe} would necessarily do better than a receiver with memory. The effect of these costs is a combination of vertical translation and scaling, which amount to a change in the number of interactions required for memory to convey some benefit.[6]

We have laid out the general effect of population size and cost. In cer-

[6]So far, we have only considered those cases where memory is perfect; receivers never forget an interaction and have, in effect, an unbounded memory. An obvious extension would be to consider cases where memory is bounded by either a moving window (where the last n agents are recalled) or probabilistic decay (where the nth previous interaction is forgotten with some fixed probability).

tain cases memory allows for what amounts to honest signaling. From the viewpoint of an agent that functions as both a sender and a receiver this can be made clear. A dishonest sender will only ever try to trick an unknown receiver, or one that they do not remember. This does not render signaling totally honest. Receivers can still be tricked by unknown receivers, but the likelihood of this can be decreased given a reasonable cost for memory. This, in effect, allows for the first submaxim of quality.

3.4 Gossip

It seems useful to think of gossip as the externalization of memories of social interactions. That is, agents might signal information about other agents to other agents. Thus, a system where the agents do not gossip is equivalent to a system where each agent relies on its own memory to track honest and dishonest agents. We explore the effect of gossip in populations of varying sizes, focusing in particular on the difference between local and non-local interaction. We find that where the usefulness of memory deteriorates quickly in a larger population, gossip allows agents to sort the honest from the dishonest.

To begin, we consider the limiting case of local interaction. Suppose that each agent only interacts with the neighborhood of eight agents surrounding it. Note that this is equivalent to being in a population of size $N = 8$. As noted in the previous section, the size of the population determines the amount of time it takes to discover the types of all agents. All else being equal, the types of agents will be discovered more quickly in a small population compared to a large one. In these cases, where memory is efficient, we do not expect a significant benefit from gossip. Instead, we look to those circumstances where memory breaks down.

When the population is large enough, and interaction is not constrained to a local neighborhood, memory ceases to convey as great a benefit. We can see this by comparing the expected utility of a receiver with local interaction to that of a receiver with non-local random interaction in a population. This comparison can be seen in Figure 10. The benefit of memory is quickly swamped in a larger population.

In a small enough population gossip yields no benefit over and above memory. But, as the population grows, we begin to see the benefit of gossip. If agents gossip about each other as time passes, we expect the reputation of honest and dishonest agents to spread more quickly through the population. Specifically, suppose that at each time step agents share a certain amount

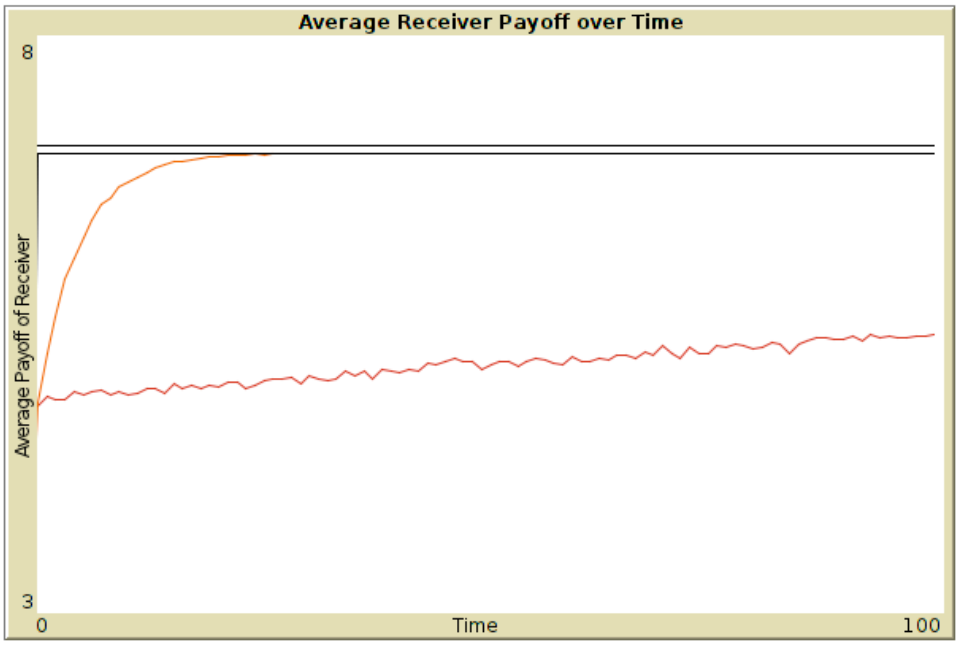

Figure 10: Average receiver payoff with local (orange) versus non-local (red) interaction. N = 500, q = .2

of information with each other. The case where the amount of information is 0 thus corresponds to memory, but as the amount is increased we see an increase in the average payoff of receivers even in large populations with random interaction. the results of simulation with various degrees of gossip are shown in Figure 11.

Here what we see is that as the amount of information conveyed via gossip increases, the average payoff of receivers increases. The difference between memory and a single piece of gossip is dramatic. The effect of gossip is to make a large population with random interactions much more local. In this sense, gossip lead to public reputations, which in turn lead to honest signaling.

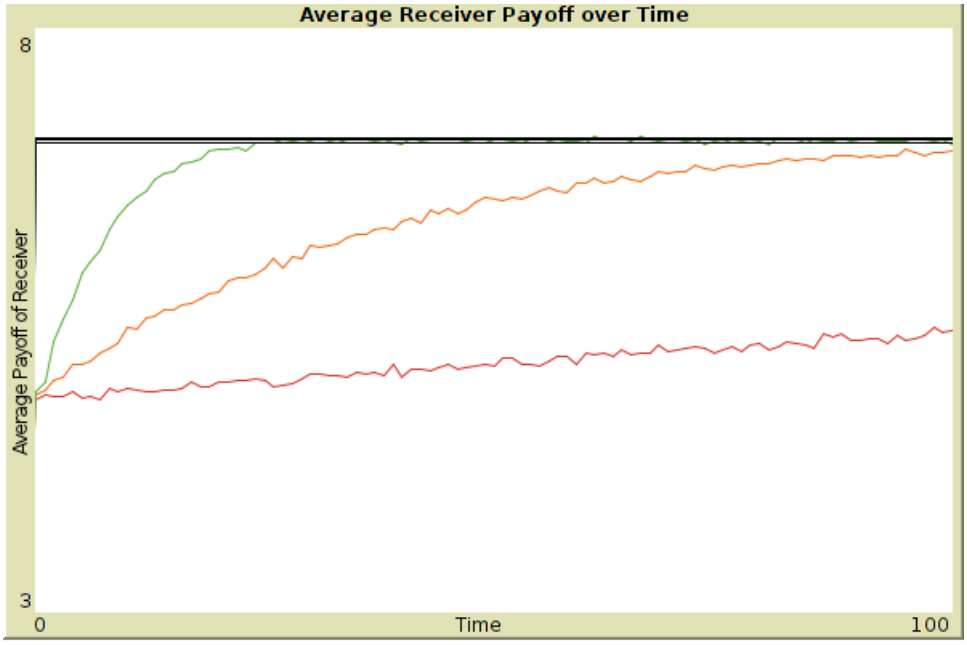

Figure 11: Average receiver payoff with no gossip (red), 1 piece of information (orange), 8 pieces of information (green), random interaction. N = 500, q = .2

4 Conclusion

We have examined how different mechanisms might support honest signaling as it relates to Grice's Maxim of Quality. Importantly, we have moved beyond games with perfect incentive alignment and considered how simple mechanisms such as neologisms, memory, and gossip might allow for the maintenance of honest signaling in cases of imperfect incentive alignment. Working out the interaction between the details of these different mechanisms is a goal for future research.

A Appendix: Payoff Structure

We adapt the general form of Trust Games (Berg et al. 1995) to capture the payoff structure discussed above and give form to U_S and. U_R. In Trust Games there are two roles, an Investor and a Trustee. The Investor begins with an Initial endowment, which he can keep or invest. If he invests the endowment with the Trustee it grows by some amount, and the Trustee must then decide what amount, if any, to return to the Investor. Let e be the initial endowment of energy, c be the cost of attending to a signal, $p > 0$ be the proportion of the endowment invested, m be the multiplier of growth, and r be the proportion returned by the sender. As suggested above, the multiplier and the rate of return can vary based on the type of the receiver and the amount invested by the receiver. We will think of the actions available to the receiver as varying degrees of energy investment. The receiver can ignore the sender and invest no effort, a_0, put forth some effort, a_{pe}, or put forth the maximal degree of effort, a_e. Finally, let γ be a function of the type of the sender that describes the effect of observing the sender interacting with another agent. If the sender is honest, then this will be some negative amount reflecting the lost opportunity on the part of the receiver. If the sender is dishonest this will be some positive amount reflecting the relief the receiver feels from avoiding being swindled. Thus, the payoff for the receiver can be given as:

$$U_R(t, a_i) = \begin{cases} e + \gamma(t) & \text{if } i = 0 \\ (1-p)(e-c) + p(mr(e-c)) & \end{cases} \tag{11}$$

For the sender, if the receiver ignores her signal, then she receives nothing from the interaction. If the receiver attends to her signal, then she receives some benefit based on the amount of effort invested by the receiver, and the amount given back:

$$U_S(t, a_i) = \begin{cases} 0 & \text{if } i = 0 \\ (1-r)(mpe) & \end{cases} \tag{12}$$

Thus we can represent the payoffs of both sender and receiver in strategic form in Figure 12. For reasons of space, let $\alpha = (1-p)(e-c) + p(mr(e-c))$.

We consider the preferences of each role in turn. Both types of senders want receivers to listen; they necessarily do better when this is the case. Both

	a_0	a_{pe}	a_e
t_h	0 , $e + \gamma(t)$	$(1-r)(mpe)$, α	$(1-r)(mpe)$, $mr(e-c)$
t_d	0 , $e + \gamma(t)$	$(1-r)(mpe)$, α	$(1-r)(mpe)$, $mr(e-c)$

Figure 12: Sender and Receiver Payoffs

types of senders do the same when it comes to an intermediate amount of investment. It does not take much skill or long arms to forage in the bushes. However, honest types prefer to have the receiver invest as much as possible as it results in a better payoff than an intermediate investment. If Sally has long arms or truly intends to gather a whole bunch of fruit, then it is a better outcome to get lifted into the trees. In contrast, a dishonest sender does not prefer to be lifted into the trees. That is, she may be able to grab some fruit, but not that much, and she risks the physical harm of getting dropped by the receiver.

The preferences of the receiver are conditional on the type of the sender. If the sender is honest, then the receiver prefers to exert the maximum amount of effort and lift her up into the tree to get the best fruit. That being said, a receiver prefers putting some effort into an interaction with an honest sender to ignoring her. This just means that the remorse over a lost opportunity is sufficient to render ignoring an honest sender less preferred than putting forth an intermediate effort. If the sender is dishonest, then the receiver would prefer to not get swindled. Moreover, it would be in the best interest of the receiver to just ignore the dishonest receiver as this would leave him available to listen to other possibly honest senders. For this to be the case, it suffices that the benefit from ignoring a dishonest sender and the cost of attending a signal outweigh the return on an intermediate effort.

These general preferences of senders and receivers can be summarized in the following:

1. $U_S(t_h, a_0) < U_S(t_h, a_{pe}) < U_S(t_h, a_{pe})$
2. $U_S(t_d, a_0) < U_S(t_d, a_e) < U_S(t_d, a_{pe})$
3. $U_R(t_h, a_0) < U_R(t_h, a_{pe}) < U_R(t_h, a_e)$

4. $U_R(t_d, a_e) < U_R(t_d, a_{pe}) < U_R(t_d, a_0)$

References

Akerlof, George A. 1970. The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics* 84(3): 488–500.

Benz, Anton, Gerhard Jäger, and Robert van Rooy. 2006. An introduction to game theory for linguists. In *Game theory and pragmatics*, ed. Anton Benz, Gerhard Jäger, and Robert van Rooy, 1–82. Palgrave studies in pragmatics, language and cognition, Basingstoke, UK: Palgrave Macmillan.

Blume, Andreas, Douglas V. DeJong, Yong-Gwan Kim, and Geoffrey B. Sprinkle. 2001. Evolution of communication with partial common interest. *Games and Economic Behavior* 37:79–120.

Bowles, Samuel, and Herbert Gintis. 2011. *A cooperative species: Human reciprocity and its evolution*. Princeton, NJ: Princeton University Press.

Clark, Robin. 2012. *Meaningful games: Exploring language with game theory*. Cambridge, MA: The MIT Press.

Farrell, Joseph. 1993. Meaning and credibility in cheap-talk games. *Games and Economic Behavior* 5:514–531.

Fehr, Ernst, and Simon Gächter. 2000. Cooperation and punishment. *American Economic Review* 90(4):980–994.

———. 2002. Altruistic punishment in humans. *Nature* 415:137–140.

Fischbacher, Urs, and Simon Gächter. 2010. Social preferences, beliefs, and the dynamics of free-riding in public good experiments. *American Economic Review* 100(1):541–556.

Lewis, David. 1969. *Convention: A philosophical study*. Cambridge, MA: Harvard University Press.

Nowak, M. A. 2006. Five rules for the evolution of cooperation. *Science* 314(5805):1560–1563.

Parikh, Prashant. 2001. *The use of language*. Stanford, CA: CSLI Publications.

———. 2010. *Equilibrium semantics*. Cambridge, MA: The MIT Press.

Rabin, M., and J. Sobel. 1996. Deviations, dynamics and equilibrium refinements. *Journal of Economic Theory* 68:1–25.

Scott-Phillips, Thomas. 2008. On the correct application of animal signaling theory to human communication. In *The evolution of language*, ed. S. D. M. Smith, K. Smith, and R. Ferrer I Cancho, 275–282.

———. 2010. Evolutionarily stable communication and pragmatics. In *Language, games and evolution*, ed. Anton Benz, Christian Ebert, Gerhard Jaeger, and Robert van Rooij, 117–133. Springer.

Searcy, William A., and Stephen Nowicki. 2005. *The evolution of animal communication: Reliability and deception in signaling systems*. Monographs in Behavior and Ecology, Princeton, NJ: Princeton University Press.

Zahavi, Amotz. 1993. The fallacy of conventional signaling. *Philosophical Transactions of the Royal Society B* 340(1292):227–230.