Evolution and Human Behavior 36 (2015) 475–479

Contents lists available at ScienceDirect

Evolution and Human Behavior

journal homepage: www.ehbonline.org

Original Article

Third-party monitoring and sanctions aid the evolution of language

Robert Boyd *, Sarah Mathew

School of Human Evolution and Social Change, Arizona State University

ARTICLE INFO

Article history:
Initial receipt 9 March 2015
Final revision received 18 June 2015

Keywords:
Language
Third-party sanctions
Reciprocity
Honest signaling
Reputation

ABSTRACT

The control of deception is an important problem in the evolution of all communication systems including human language. A number of authors have suggested that because humans interact repeatedly, reputation can control deception in human language. However, there has been little work on the theory of repeated signaling. This lacuna is important because unlike many others forms of defection, lies are not easily detected, and attempts to determine the truthfulness of signals can lead to false accusations of deception. Here we modify a standard model of animal signaling, the Sir Philip Sidney Game, to allow for repeated interactions between pairs of individuals. We show that unless it is easy to detect lies, communication is unlikely to be evolutionarily stable. However, third-party monitoring of pairwise interactions and sanctioning of dishonesty increases the range of conditions under which cheap talk can evolve, a finding that suggests that cooperation enforced by third-party monitoring and punishment may have predated the evolution of language.

The control of deception is an important problem in the evolution of communication systems. Maynard Smith and Harper (Maynard Smith & Harper, 2003) define a signal as a behavior that evolved because it affects the behavior of others who have evolved to respond to the signal. When the interest of the individual who generates the signal conflicts with the interest of the receiver, selection will favor deceptive signals that decrease the fitness of receivers unless such deceptive signals entail sufficient cost to deter defection. This problem is particularly acute in the case of human language because words can be recombined to generate an almost unlimited range of meanings; the possibilities for deception are endless (Lachmann, Szamado, & Bergstrom, 2001). The handicap principle is one well-known solution to this problem (Godfray, 1991; Grafen, 1990; Maynard Smith & Harper, 2003; Zahavi, 1975), although its empirical significance is debated (Számadó, 2011). Other solutions include indices, pooling and hybrid equilibria, and repeated interactions (Számadó, 2011; Zollman, 2013).

Of these, repeated interactions coupled with punishment of deception is the most plausible mechanism for maintaining honesty in human language (Lachmann et al., 2001; Scott-Philips, 2014). Contingent behavior of this kind is thought to play a key role in many forms of human cooperation (Axelrod & Hamilton, 1981; Nowak & Sigmund, 1998; Panchanathan & Boyd, 2004; Trivers, 1971; van Veelen, Garca, Rand, & Nowak, 2012). However, the evolution of honest communication differs from other forms of reciprocity. In most forms of reciprocity the failure to provide help is easily detected. The recipient does not receive aid and thus knows that her partner defected and can immediately stop helping the partner. Lies are different. The liar knows the truth, but the listener can detect the lie only by making use of other evidence—the false signal itself is not sufficient.

There is little formal theory of repeated signaling among known individuals when lies are hard to detect. Silk, Kaldor, and Boyd (2000) study a model of repeated interaction among cercopithecine primates in which individuals signal their intent to peacefully approach with a quiet grunt. In this model, honest signaling is an equilibrium if receivers do not believe signals from particular partners once they have been deceived by that partner. Rich and Zollman (2015) investigate the repeated discrete Sir Philip Sidney game (Maynard Smith, 1991) where individual lies are never detected because the signaler's condition is known only to the signaler. They show that a strategy that keeps track of the rate at which an individual signals need, and stops transferring when that rate is too high can support honest signaling. These two models focus on two ends of an important continuum: in Silk *et al.* lies are always detected immediately, while in the model of Rich and Zollman individual lies are never detected, but dishonest behavior can be detected statistically over the long run. Catteeuw, Han, and Manderick (2014) study a one shot version of the Sir Philip Sidney game, but with punishment. Because it is a one shot model, and signals are costly, and signaler and receiver are related, this model does not capture the key features of the evolution of human language.

We examine a revised version of the Sir Philip Sidney game Maynard Smith (1991) in which interactions are repeated, and take place between unrelated individuals. Because only the speaker knows whether he lied or not, the listener has to use other evidence to decide whether her partner lied or not. The listener correctly detects a lie with some probability, and erroneously thinks a true statement is a lie with some probability. Thus the less trusting they are, the more likely they are to detect lies, but also the more likely they are to erroneously perceive a lie. This means there will be a high rate of perception errors (Nowak & Sigmund, 2005)—the speaker spoke the truth but the listener

* Corresponding author. 900 Cady Mall, Arizona State University, AZ, 85287. Tel.: +1 480 965 7671.
E-mail address: robert.t.boyd@gmail.com (R. Boyd).

http://dx.doi.org/10.1016/j.evolhumbehav.2015.06.002
1090-5138/© 2015 Elsevier Inc. All rights reserved.

erroneously believes it is a lie. While there has been some study of the evolution of direct reciprocity when rates of perception errors are very low, little is known about higher rates of perception errors, or the effect of such errors on indirect reciprocity.

1. The model

There is a large population of individuals, and individuals are paired with unrelated partners. One individual is a signaler and the other is a receiver. Signalers are "deserving" with probability p or "undeserving" with probability 1–p. Receivers can perform a costly action that benefits the signaler. The payoffs are given in Table 1.

Receivers do not know the signaler's state, but signalers can signal their state at zero cost. After a signal and a helpful act, receivers assess the truthfulness of the signal. The receiver correctly identifies false signals with probability e, and incorrectly identifies a truthful signal as false with probability ε. Each time period the interaction continues with probability w. Individuals alternate roles, so if an individual is a receiver in a one time period, she is a signaler in the next time period and so on. The expected number of interactions for each individual in a given social role is $T = \frac{1}{1-w^2}$.

This game structure can represent a number of different situations. For example, mutual aid is common in human societies (Sugiyama, 2004). When an individual is sick or injured, she requests help from others. Later when others are sick or injured, she returns the aid. It also represents many situations in which the signaler has an obligation that may be avoided under some circumstances. In the classic John Hughes film, *Ferris Bueller's Day Off*, Ferris is obliged to go to high school but feigns illness so that his mother will let him stay home. But instead Ferris ditches class and enjoys a day gadding about Chicago. More generally, the model applies to any circumstance in which the receiver is motivated to perform a costly act benefiting the signaler in one state of the world but not others, and only the signaler knows whether that state of the world is correct.

There are two strategies for each social role: Signalers can be:

Honest (H) Signals when deserving, does not signal when undeserving.
Dishonest (D) Always signals.

Receivers can:

Respond (R) Help a signaler who is in good standing when own standing is good; otherwise do not help. An individual begins in good standing and falls out of good standing if (1) as a signaler she has been identified as giving a false signal, (2) she did not help the last time she received a signal of need from an actor in good standing, or (3) she helped after receiving a signal from someone in bad standing.
Never Respond (N) Ignore the signal and never help.

Of course, many other strategies are possible. In particular, with these strategies once an individual falls into bad standing she can never get back into good standing. Dealing with this problem is an important, but difficult problem when lies and false accusations of lying cause different actors to have different beliefs about what has occurred. We focus on the strategies described above because they are the simplest that capture the essential features of the problem.

When signals are honest, individuals who respond to the signals of need can resist invasion by those who do not respond when

$$wT_H(a - wc) > c \qquad (1)$$

where $T_H = \frac{1}{1-w^2(1-p\varepsilon)}$ is the waiting time until an honest signaler is falsely accused. (See Supplementary Information for proofs, available on the journal's website at www.ehbonline.org.) The right hand side of (1) is the cost of providing help during the first interaction on which individuals hear a signal. The left hand side is the long-term advantage of receiving help when in need minus the cost of helping. This is multiplied by the number of time periods in which honest players are in good standing. Thus, increasing w makes it more likely that responders are favored. Allowing $\varepsilon > 0$ will reduce the range of conditions under which providing help is favored, but this increment will be small if ε is small. We will assume that (1) is satisfied.

Table 1
Fitness effects during one time period contingent on receiver behavior and signaler state.

	Do not help	Help
Receiver	1	1–c
Undeserving signaler	1–b	1
Deserving signaler	1–a	1

If responders are common, honest signalers can resist invasion by rare liars if

$$\frac{pa}{pa + (1-p)b} > \frac{T_L}{T_H} \qquad (2)$$

where $T_L = \frac{1}{1-w^2(1-p\varepsilon-(1-p)e)}$ is the waiting time until a liar is exposed.

The left hand side of (2) is the ratio of the incremental fitness effect of signaling only when needy to the incremental fitness of always signaling. If the benefit of lying (b) is large or there are many opportunities for lying (1–p), then the ratio is small. When lies are not particularly beneficial or the opportunity to benefit from a lie is rare, the ratio will be close to one. The right hand side gives the ratio of the time until a liar falls into bad standing to the time until an honest signaler falls into bad standing. This ratio is always less than one. Thus the stability of honest signaling depends on the relative difficulty of detecting lies and the propensity to mistake honest signals as lies.

Because identifying lies requires other evidence, it should be thought of as a signal detection problem. After the donor gets the signal she decides whether the signal was truthful. This decision is based on cues that she observes. Ferris fakes a fever, and then joins his friends for a day on the town. If his mother had happened to see him later in the day, she would know he had lied. Of course, such cues can mislead. Ferris might really be home in bed, but his mother sees a Ferris-look-alike driving downtown and concludes that Ferris faked his symptoms. To avoid such mistakes, actors can impose a higher burden of proof before they conclude someone has lied. Ferris's mother could try to make sure that she got close enough to be sure it was Ferris on the float, but this increased burden of proof will reduce the probability of detecting true lies.

We model this tradeoff using a signal detection model (McNicol, 2005). The cue is a normally distributed random variable, as shown in Fig. 1. Actors infer that a lie has occurred if the cue value is greater than d, and therefore $e = 1 - F(d|\text{lie})$ and $\varepsilon = 1 - F(d|\text{truth})$. As d is increased both e and ε decrease.

This model suggests that the stability of honest signaling is sensitive to the probability that lies are detected. In Fig. 2, we plot $\frac{T_L}{T_H}$ as a function of e for three values of M, the difficulty of distinguishing lies from true

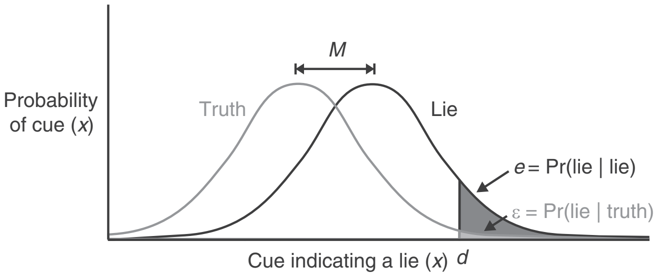

Fig. 1. The Gaussian probability densities of cue values conditioned on a truthful signal (grey) and a lie (black). Actors infer that a lie has occurred if the observed cue value is greater than d. The probability of detecting a lie, e, is always greater than the probability of falsely attributing a lie to an honest signal, ε. Increasing the value of d, decreases both e and ε but decreases the ratio ${}^{e}/_{\varepsilon}$, so errors become relatively less likely.