Honesty through repeated interactions

Patricia Rich and Kevin J. S. Zollman

February 14, 2015

Technical Report No. CMU-PHIL-192

Philosophy

Methodology

Logic

Carnegie Mellon

Pittsburgh, Pennsylvania 15213

Honesty through repeated interactions

Patricia Rich Kevin J. S. Zollman*

Abstract

The handicap principle – which posits that signals must be costly in order to be reliable – was the paradigmatic theory for understanding the apparent stability of communication between organisms in settings where deception would otherwise be beneficial. More recently it has been shown that marginal cost, rather than actual cost, is the essential element for honest signaling. In this paper, we show how marginal cost can arise endogenously in repeated interactions between individuals. Utilizing the Sir Philip Sidney game as an illustrative case, we demonstrate that repeated interactions can sustain honesty even when deception cannot be directly observed. We provide a number of potential experimental tests for this theory which would distinguish it from the available alternatives.

Keywords: Handicap theory, costly signaling, Sir Philip Sidney game, reputation

*To contact the authors please write to: Department of Philosophy, Baker Hall 135, Carnegie Mellon University, Pittsburgh, PA 15217, USA. Or email: kzollman@andrew.cmu.edu

1 Introduction

In many cases of signaling in nature, there is honest communication of information between individuals. This occurs even when individuals' selfish motives would seem to make deception profitable. The handicap principle was initially formulated by Zahavi to explain this apparent paradox (Zahavi, 1975; Zahavi and Zahavi, 1997). The basic insight was that if signaling carries a cost such that dishonesty is prohibitively expensive but honest signaling worthwhile, signalers do best by signaling honestly, and receivers do best to make use of the accurate information carried by signals. Formal models showing that such a cost structure indeed makes honest signaling evolutionarily stable, e.g. those by Grafen (1990), Godfray (1991) and Maynard Smith (1991), were used to support Zahavi's claim that the handicap principle is uniquely able to account for reliable signaling in nature.

The handicap principle has become ubiquitous, and is often treated as the only potential explanation for the stability of honest communication. Zahavi's description of the principle, and the early models of it, suggest honest signaling in the wild should come with high, observable costs to the signalers. However, the failure to find sufficiently high signal costs in a number of experiments has led to an interest in alternative theories (see, e.g. Maynard Smith and Harper, 2003; Searcy and Nowicki, 2005). It has since been shown that actual cost is not necessary to sustain honesty (Hurd, 1995; Számadó, 1999; Lachmann et al., 2001; Számadó, 2011b). Instead, it is marginal cost – the cost of deception – which is critical. Honesty can be free, so long as lying is costly. For example, Hurd (1997) showed that honest communication of fighting ability was possible with very low observed cost. This is achieved by postulating that the cost to a weak individual who imitates a strong one is sufficiently high to deter deception – a plausible assumption in animal contests.

Marginal costs, like these, might not be observed in systems in equilibria, and therefore could only be found by empirical investigation into how the system behaves outside of its natural state. While the theoretical correctness of this claim has been known for some time, there are relatively few biologically plausible methods for creating this marginal cost provided in the literature (for other example, see Lachmann and Bergstrom, 1998; Bergstrom and Lachmann, 1998; Johnstone, 1999; Silk et al., 2000; Számadó, 2008, 2011a; Catteeuw et al., 2014). This paucity of models makes empirical investigation into the possibility of marginal cost difficult.

This paper explores the possibility of creating marginal cost, without creating actual cost, in the context of signaling among relatives by focusing on the possibility that repeated interactions might influence the evaluation of signals. It is plausible that children honestly signal their need to their parents because their signaling habits can be used to condition the parent's response. Signaling thus furnishes children with a kind of "reputation," and a child with a reputation for signaling too much would eventually be ignored by the parent and denied food in a way that harmed the child. At the outset, we should be clear that the word "reputation" as we are using it does not suppose there is

secondary communication like gossip (as used in Nowak and Sigmund, 1998; Ohtsuki and Iwasa, 2006). Instead we are supposing that the parent learns how frequently the child signals and this is what we call the child's reputation. This limited kind of reputation, we argue, could replace direct cost as a mechanism for keeping signaling honest.

We show that this intuitive idea is indeed formally tenable, *even when dishonesty cannot be directly observed*. To do so we augment Maynard Smith's (1991) Sir Philip Sidney game with *reputation-based* strategies, and show that pairs of such strategies can constitute equilibria. Most importantly, these equilibria exist when the direct signal cost is too low to function as a traditional handicap. While we do not extend the analysis to other communicative games, we believe that these results should generalize to other communicative interactions that feature partial conflict of interest.

2 Handicaps in the Sir Philip Sidney game

Maynard Smith (1991) invented the Sir Philip Sidney game in order to provide a relatively tractable example of Grafen's (1990) model of the handicap principle. Lying mortally wounded on the battle field, Philip Sidney is said to have given his water to a fellow soldier with the declaration, "thy necessity is yet greater than mine." This idea – of the transfer of resources between two individuals who share a common interest – provided the basis for the game.

The formal game, shown in figure 1, involves two players; these players are typically imagined as a chick and a parent, although the model is more general. At the first node of the game some exogenous force, usually called "nature" determines whether the chick is needy (with probability p) or healthy (with probability $1-p$). At the second node the chick, conditioning on the decision by nature, either begs for food – signals to the parent – or not. Finally, in response to the signal (but not to the choice by nature) the parent either provides the chick food or keeps the food for itself. Several variations of this game have been proposes where there are more states of need, more signals, and differing amounts of transfer (Johnstone and Grafen, 1992; Bergstrom and Lachmann, 1997, 1998).

Each player's individual fitness is 1 minus the value of any penalty parameters given by the game's outcome: a chick who signals pays a signal cost $0 \leq c < 1$; a parent who gives the chick food loses fitness $0 < d < 1$; a needy chick who doesn't receive food pays a fitness cost of $0 < a < 1$ and a healthy chick $0 < b < 1$, where $a > b$. The inclusive fitness of each player is determined by their individual fitnesses plus a fraction, r of the fitness of the other individual. We will presume that, at a minimum, the parent wishes to transfer the resource to the needy chick, i.e. $d < ra$.

Holding the parameters a, b, d, and p fixed and allowing r and c to vary defines four areas of interest (Huttegger and Zollman, 2010), which are pictured in figure 2. In region 1, the cost is so high that the chick should not send the signal regardless of its state of need. Regions 2 and 3 represent the classic situ-

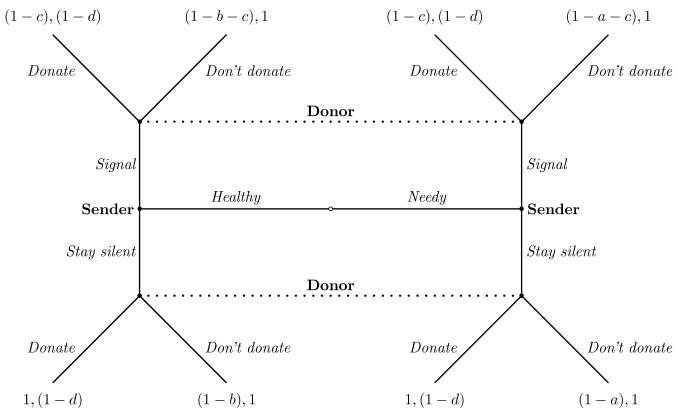

Figure 1: A game tree illustrating the Sir Philip Sidney game without inclusive fitness (from Huttegger and Zollman, 2010). The game begins at the node in the center where "nature" determines if the chick is healthy or needy. The chick conditions its behavior on this choice and decides whether to send a costly signal or stay silent. The parent can condition on the signal, but not on the state of need of the chick, and can choose wether or not to donate a resource or not. The fitness for the chick (first) and parent (second) are given at the terminal nodes.

ation for the Sir Philip Sidney game and the handicap principle more generally. Here, when $rd < b$, there is a situation of parent–offspring conflict; the parent only wishes to transfer the resource to the needy chick, but both the needy and healthy chick would like to acquire the resource from the parent. In such a case, without signal cost the healthy chick has an incentive to imitate the needy chick in order to secure the resource.

In region 2, the cost of the signal is sufficiently high, however, that the healthy chick is unwilling to pay the cost necessary to successfully imitate the needy chick. As a result, in region 2, honest communication is stable because of the presence of a significant signal cost. Alternatively, in region 3 the cost is too low, and as result totally honest communication is impossible. Here the only pure strategy equilibrium is where neither chick sends the signal and the parent transfers the resource or not depending on the underlying probability that the chick is needy. (Recent work has illustrated another polymorphic equilibrium exists in this region where *partially* honest communication can be sustained Huttegger and Zollman 2010; Wagner 2013; Zollman et al. 2013.)

Finally, in region 4, the game is a game of pure common interest – both the parent and the chick prefer the parent transfer the resource when the chick is needy, and neither prefer the parent transfer the resource when the chick is healthy. In such a situation, totally honest communication is possible even with no signal cost.

Region 2 is central to discussions of the Sir Philip Sidney game. Here one

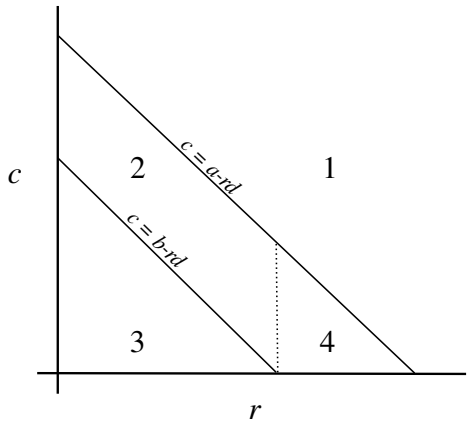

Figure 2: Regions of interest in the Sir Philip Sidney game holding a, b, and d fixed.

predicts that when honest communication exists, one should observe significant signal cost. Empirical confirmation of this prediction has been rare. While there is clear evidence that signaling from children to parents communicates information about the state of need of the child, there is little evidence of significant cost to begging, especially in birds (for an overview see Searcy and Nowicki, 2005). In addition to these empirical concerns, a number of models have shown that evolving to one of these signaling equilibria might be very difficult (Bergstrom and Lachmann, 1997; Rodríguez-Gironés et al., 1998; Hamblin and Hurd, 2009; Huttegger and Zollman, 2010; Zollman et al., 2013).

When the cost of the signal is too low to sustain signaling, a healthy chick who fails to signal is engaging in an altruistic act (Zollman, 2013). That is, the chick is lowering its own fitness in order to enhance the fitness of the parent. The study of biological altruism is voluminous, but one central finding is the possibility that repeated interactions may furnish a potential explanation for altruistic behavior (Trivers, 1971). (Inclusive fitness cannot solve this problem because inclusive fitness is already accounted for in the model, and the act remains altruistic.)

Traditional models of repeated interactions presume *perfect monitoring* where a partner can determine whether or not a conspecific has behaved altruistically. This assumption is been used in the models of cost-free honest signaling (Silk et al., 2000; Catteeuw et al., 2014). In the case of the Sir Philip Sidney game, this would translate to the parent becoming aware, after the interaction, whether the chick was needy or healthy when it begged. This is biologically implausible, and as a result, models of the repeated prisoner's dilemma can not be uncritically applied to this case. In the following section, we demonstrate another method by which reputations might stabilize honest communication in the Sir Philip Sidney game even with signals that have no cost.

3 Reputation through repeated interaction

It is intuitively quite plausible that, instead of a direct cost, signaling can be indirectly costly for an individual because their pattern of signaling over time furnishes the individual with a reputation that the receiver uses to determine how to behave towards the individual in question. This possibility is most viable when the signalers and receivers interact repeatedly or have many opportunities to observe others' behavior, as with parents and young or group-living species. As an example, chicks may be harmed by dishonesty if it gives them a reputation that causes the parent to ignore the chick's begging and forgo feeding it in the future.

Define a *reputation* for a chick as the probability F that the chick signals on an arbitrary round of the Sir Philip Sidney game. This is the appropriate quantity to utilize because it takes into account the information the parent can observe (frequency of signaling) and not what is unobservable (the health of the chick). A chick acquires a reputation based on its signaling frequency in each of the two situations it may find itself in, i.e. based on the frequency of signaling when needy (denoted by y) and when healthy (denoted by z). A chick's reputation, then, is $F := py + (1-p)z$. Then let Byz denote the chick's strategy which is defined by probabilities y and z.

A parent who has observed F through previous interactions chooses a strategy Dx where x is a probability and the chick is given food in response to a signal if and only if $F \leq x$. In other words, the strategy Dx sets an upper limit to the chick's signaling frequency above which the chick will not be fed but below which the signal will be trusted.

Let a reputation-based equilibrium (RBE) be any reputation-based strategy pair (Byz, Dx) with $py+(1-p)z = F \leq x$ that constitutes a Nash equilibrium. Generally, the case where $y = 1$ and $z = 0$ (honest signaling) and the case where $y = z = 0$ (never signaling) will not be referred to as RBE.

First we note that RBE cannot exist when the cost is too high to sustain a signaling equilibrium, i.e. when $c > a - rd$ (region 1 of figure 2). Once the cost of signaling becomes so high that even the needy chick is unwilling to signal, no RBE exist. This is consistent with Maynard Smith's version of the Sir Philip Sidney game.

Henceforth we will assume that the cost to the signal is below this threshold, that we occupy regions 2, 3, or 4 of figure 2. For all these regions we can partially characterize the chick's best response to a given parental strategy.

Suppose that the parent adopts a strategy Dx. The best response for the chick is to choose y and z in order to maximize this equation: $p[y[(1-c)+r(1-d)]+(1-y)[(1-a)+r]]+(1-p)[z[(1-c)+r(1-d)]+(1-z)[(1-b)+r]]$, subject to the constraint that $F \leq x$. (The chick's signaling frequency F cannot exceed x as then the chick will sometimes paying the cost of signaling but never receive food.) Now, by the assumption that $a \geq c + rd$, the payoff for the y cases is at least as high as the payoff for the $1 - y$ cases, and so the chick prefers for y to be as high as possible relative to $1 - y$. The payoffs for the y cases and the z cases are equal. However, the payoff for the $1 - y$ case is strictly less than the

payoff for the $1 - z$ case as $a > b$.

This means that for any instantiation of the Sir Philip Sidney game in regions 2, 3, or 4, whenever a parent adopts a strategy Dx, the chick does best by setting y to satisfy this equation,

$$y = \begin{cases} 1 & \text{if } x \geq p \\ \frac{x}{p} & \text{otherwise} \end{cases} \tag{1}$$

Should the parent adopt a strategy x such that $x \leq p$, this constraint fully characterizes the chick's best response. The chick does best by setting $y = x/p$ and $z = 0$. However, when $x > p$, this equation does not fully characterize the best response for the chick, because it does not determine the value of z. To fully characterize the best response of the chick, we must first consider regions 2 and 4 separately from region 3.

Recall that in regions 2 and 4 honest signaling is an equilibrium. It is an equilibrium in region 2, because the cost is sufficiently high to prevent the healthy chick from profitably imitating the needy one, and sufficiently low to allow the needy chick to profitably signal. In region 4, cost is unnecessary because the chick and parent are related to a sufficiently high degree that the healthy chick does not wish to secure the resource.

As shown in (Maynard Smith, 1991), when $c > b - rd$, the healthy chick does better by refraining from signaling than by paying the cost and securing the resource. As a result, a positive z is strictly worse than $z = 0$. So in regions 2 and 4, the best response for a chick to strategy Dx is:

$$y = \begin{cases} 1 & \text{if } x \geq p \\ \frac{x}{p} & \text{otherwise} \end{cases}$$

and,

$$z = 0$$

This presumed that the parent is adopting a conditional transfer strategy Dx. Alternatively, if the parent is choosing an unconditional strategy, either to always transfer or never transfer (regardless of signal), the chick always does best by never signaling, since there is no reason to pay the cost c. (In the case where $c = 0$ the chick may either signal or not.)

This characterizes the best response of the chick to the parent's strategy. Now we must consider the parent's response to a chick who adopts a strategy Byz which yields a signaling frequency of F.

Suppose the chick adopts a strategy Byz. First note that Dx_1 and Dx_2 are behaviorally equivalent for the parent when $x_1, x_2 \geq py$ (both transfer the resource when the signal is received). Similarly when $x_1, x_2 < py$ Dx_1 and Dx_2 are behaviorally equivalent to the strategy *never transfer*. As a result we only must compare three strategies: Dx_1, where $x_1 \geq py$, *always transfer*, and *never transfer*.

Suppose the chick pursues a strategy Byz such that $z = 0$. Because $d < ra$, the parent wishes to transfer the resource to the needy chick, but because $d > rb$

the parent does not wish to transfer to the healthy chick. The strategy Dx_1 (where $x_1 \geq py$) will transfer the resource to the needy chick with probability y and will never transfer to the healthy chick. As a result, it is superior to the strategy *never transfer* whenever $y > 0$.

Dx_1 yields at least as high a payoff as the strategy *always transfer* when $p(1-y)[1+r(1-a)]+(1-p)[1+r(1-b)] \geq p(1-y)[(1-d)+r]+(1-p)[(1-d)+r]$. In the $p(1-y)$ case, since it is required that $a \geq \frac{d}{r}$, Dx yields as high a payoff as *always transfer* only when $a = \frac{d}{r}$. Considering only the $(1-p)$ case, as $b \leq \frac{d}{r}$, *always transfer* is never better than Dx_1. These cases show that *always transfer* yields a strictly higher payoff than Dx when $p(1-y)[ar-d]+(1-p)[br-d] > 0$. Or, alternatively,

$$y > 1 + \frac{(1-p)(br-d)}{p(ar-d)}$$

When this equation is satisfied, the parent does best to ignore the signal and always transfer the resource. If, on the other hand this equation is not satisfied the parent prefers to choose a strategy Dx such that $x \geq F = py$.

This now allows us to characterize the equilibrium properties of the reputation based Sir Philips Sidney game in regions 2 and 4. In these regions, the traditional signaling equilibrium continues to exist. The chick will signal only if it is needy and the parent will transfer only if the chick signals. The pooling equilibria, where the chick never signals and the parent always transfers also remain.

The reputation game has introduced a number of new equilibria, where the needy chick only occasionally signals, and the parent transfers the resource only upon receiving the signal. In these equilibria the chick is signaling as frequently as the parent would tolerate and so cannot signal more frequently.

Let us now turn to region 3. This is the region where, in the traditional Sir Philip Sidney game, signaling is not an equilibrium because the healthy chick would like to secure the resource from the parent – although the parent does not want to transfer to the healthy chick – and the cost of signaling is too low to prevent the healthy chick from profitably imitating the needy chick. It is also the region that some recent experiments appear to, somewhat paradoxically, place actual interactions between parents and children. This where our most significant results are found.

Recall that equation 1 applies in this case as well. This already establishes an important fact that distinguishes our analysis of region 3 from the traditional analysis of the Sir Philip Sidney game. If the parent adopts a strategy Dp, which will only tolerate the chick signaling with frequency p – the exact frequency with which the chick is needy – then the chick does best by setting $y = 1$ and $z = 0$, signaling only when it is needy. It is trivial to show that this constitutes an equilibrium – an equilibrium where the chick is honestly signaling its need despite a signal cost that would be judged by the traditional analysis as too low to sustain an honest signaling equilibrium. The addition of reputation has made cheap – indeed free – honest signaling possible.

To continue our analysis we must consider what is a best response by the chick to an arbitrary parent strategy. Because $c < b - rd$ the healthy chick would be willing to signal in order to secure the resource. Because $c > 0$, however, the healthy chick does not want to signal if it will not secure the resource. So as a result, the healthy chick will signal with exactly the frequency allowed by the parent.

This allows us to fully characterize the chick's strategy in region 3. If the parent adopts a strategy Dx, the chick will choose Byz such that:

$$y = \begin{cases} 1 & \text{if } x \geq p \\ \frac{x}{p} & \text{otherwise} \end{cases}$$

and,

$$z = \begin{cases} \frac{x-p}{1-p} & \text{if } x \geq p \\ 0 & \text{otherwise} \end{cases}$$

Suppose that the parent adopts a strategy Dx and the chick adopts a strategy Byz that satisfies those constraints. Does the parent's strategy represent an optimal choice? We only must compare the strategy Dx where $x = F$ to the strategies *always transfer* and *never transfer*. Given the chick's strategy, the parent's payoff of adopting Dx is $p[y[1 + r(1 - a - c)] + (1 - y)[1 + r(1 - a)]] + (1-p)[z[1+r(1-b-c)]+(1-z)[1+r(1-b)]]$. The payoff for the strategy *never transfer* is: $p[y[(1-d)+r(1-c)]+(1-y)[1+r(1-a)]]+(1-p)[z[(1-d)+r(1-c)]+(1-z)[1+r(1-b)]]$. Because $br < d$, when the chick is healthy (the $(1-p)$ terms) the parent weakly prefers *never transfer*. However, when the chick is needy, the parent prefers Dx to *never transfer* when $(1-d)+r(1-c) \geq 1+r(1-a-c)$ which is true because $ar > d$. As a result, the parent prefers Dx to *never transfer* when $py(ar - d) + (1 - p)z(rb - d) \geq 0$ or, equivalently, when

$$\frac{y}{z} \geq \frac{(1-p)(d-br)}{p(ar-d)} \tag{2}$$

In words, this constraint requires that the chick must not signal too frequently when healthy. When the healthy chick signals too frequently, and the the healthy chick is sufficiently common, the parent would prefer to withhold the resource even though this would negatively impact the needy chick.

Now we will consider the strategy *always transfer*. The payoff for *always transfer* is $p[y[(1-d)+r(1-c)]+(1-y)[(1-d)+r]]+(1-p)[z[(1-d)+r(1-c)]+(1-z)[(1-d)+r]]$. When the chick signals (with probability py and $(1-p)z$ the payoff for *always transfer* is equivalent to Dx. When the chick is healthy and does not signal (the $(1 - z)$ term), Dx performs better. When the chick is needy and does not signal (the $(1 - y)$ term) *always transfer* performs better. Overall Dx performs better when $p(1 - y)[d - ar] + (1 - p)(1 - z)[d - br] \geq 0$ or, equivalently, when

$$\frac{(1-p)(d-br)}{p(ar-d)} \geq \frac{1-y}{1-z} \tag{3}$$

In contrast to the previous constraint, this one requires that the chick signals often enough (i.e. y must be sufficiently high). If the chick doesn't signal enough, and the chick is needy sufficiently often, then the parent would prefer to always transfer to ensure that the needy chick always receives the resource.

When the chick adopts a strategy Byz that satisfies equations 2 and 3, then a parent strategy $x = F$ is optimal. That is, the parent optimizes by transferring to a chick who signals exactly as often as this one does, but refuses to transfer to a chick that signals more frequently. Given the parent adopts such a strategy, the chick too is optimizing by choosing Byz. Therefore these two strategies are an equilibrium of the game

4 Discussion

We have shown that in the appropriately-modified Sir Philip Sidney game, there exist a number of reputation based equilibria. In some of these the state of the chick is imperfectly communicated – either the needy chick occasionally fails to signal or the healthy chick occasionally signals – but the population will nonetheless be in equilibrium.

Some of these equilibria bear a superficial similarity to the so-called "hybrid equilibria" (Huttegger and Zollman, 2010; Wagner, 2013; Zollman et al., 2013), because they feature partially honesty communication and involve occasional signaling by the healthy type. However, these equilibria are distinguished by the parent's behavior – in the hybrid equilibria the parent occasionally ignores a signal and withholds the resource. In our RBE the parent always transfers the resource when the signal is observed.

Most surprisingly, we have shown that totally honest signaling can be maintained in equilibrium without *any* appreciable signal costs. This occurs when the parent can condition her behavior on the frequency of signaling by the chick. This imposes a kind of cost that would only be observed in experiments where the parent or chick is manipulated to give the impression of too-frequent signaling.

It is unlikely that this model will provide an unequivocal explanation for honesty in all signaling interactions. For example, Kilner et al. (1999) found that the common cuckoo (*Cuculus canorus*) was able to manipulate reed warbler (*Acrocephalus scirpaceus*) parents by signaling more frequently than reed warbler chicks. However, a number of experiments on alarm calls have shown frequency conditioned behavior in rodents (Hare, 1998; Hare and Atkins, 2001; Blumstein et al., 2004) – but not all rodents (Schibler and Manser, 2007) – and primates (Cheney and Seyfarth, 1988; Gouzoules et al., 1996). Because most experiments on nestling begging look only at begging within the range normally seen in the wild, they neither provide evidence for or against this model. Further research would be necessary to test how individual parents respond to extravagant amounts of signaling.

Beyond parent–child interactions, we believe that these equilibria will be present in many different models of signaling. Therefore, RBE might provide

an explanation for honesty in other types of interaction. What is required is that the two parties interact repeatedly and be capable of recognizing each other (see Tibbetts and Dale 2007 for a discussion of the evidence for individual recognition).

These equilibria provide a concrete illustration of the observation that signal costs need not be present in equilibrium, but rather it is *marginal* cost – the cost of moving away from equilibrium – that is critical (Hurd, 1995; Számadó, 1999; Lachmann et al., 2001). In these RBE the cost is imposed by the parent punishing too-frequent signaling by withholding the resource. In equilibrium, this cost is never observed and as a result it will appear that honest communication is taking place without signal cost.

At a superficial level this is consistent with the handicap principle. Costs, in the form of withholding the resource, exist and they stabilize signaling. Searcy and Nowicki (2005) argue that there are fundamental differences between reputation costs and the costs typically posited by Zahavi. Critically, the traditional versions of the handicap principle posit the existence of *observable* costs to the signal which should be found both in and out of equilibrium. The costs imposed in RBE, on the other hand, will not be observed in systems that are in equilibria and thus require different empirical tests.

Acknowledgments

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

Bergstrom, C. T. and M. Lachmann (1997). Signalling among relatives. I. Is costly signalling too costly? *Philosophical Transactions of the royal Society of London B 352*, 609–617.

Bergstrom, C. T. and M. Lachmann (1998, April). Signaling among relatives. III. Talk is cheap. *Proceedings of the National Academy of Sciences of the United States of America 95*(9), 5100–5.

Blumstein, D. T., L. Verneyre, and J. C. Daniel (2004, September). Reliability and the adaptive utility of discrimination among alarm callers. *Proceedings. Biological sciences / The Royal Society 271*(1550), 1851–7.

Catteeuw, D., T. A. Han, and B. Manderick (2014). Evolution of honest signaling by social punishment. *Proceedings of the 2014 conference on Genetic and evolutionary computation - GECCO '14*, 153–160.

Cheney, D. L. and R. M. Seyfarth (1988). Assessment of meaning and the detection of unreliable signals by vervet monkeys. *Animal Behaviour 36*, 477–486.

Godfray, H. C. J. (1991). Signalling of need by offspring to their parents. *Nature 352*, 328–330.

Gouzoules, H., S. Gouzoules, and K. Miller (1996). Skeptical Responding in Rhesus Monkeys ({\it Macaca mulatta}). *International Journal of Primatology 17*(4), 549–568.

Grafen, A. (1990). Biological Signals as Handicaps. *Journal of Theoretical Biology 144*, 517–546.

Hamblin, S. and P. L. Hurd (2009). When will evolution lead to deceptive signaling in the Sir Philip Sidney game? *Theoretical population biology 75*(2-3), 176–82.

Hare, J. and B. Atkins (2001, December). The squirrel that cried wolf: reliability detection by juvenile Richardson's ground squirrels (Spermophilus richardsonii). *Behavioral Ecology and Sociobiology 51*(1), 108–112.

Hare, J. F. (1998). Juvenile Richardson's ground squirrels, {\it Spermophilus richardsonii}, discriminate among individual alarm callers. *Animal Behaviour 55*, 451–460.

Hurd, P. L. (1995, May). Communication in Discrete Action-Response Games. *Journal of Theoretical Biology 174*(2), 217–222.

Hurd, P. L. (1997). Is Signalling of Fighting Ability Costlier for Weaker Individuals? *Journal of Theoretical Biology 184*, 83–88.

Huttegger, S. M. and K. J. S. Zollman (2010). Dynamic stability and basins of attraction in the Sir Philip Sidney game. *Proceedings of the Royal Society of London B 277*, 1915–1922.

Johnstone, R. A. (1999). Signaling of need, sibling competition, and the cost of honesty. *Proceedings of the National Academy of Sciences of the USA 96*, 12644–12649.

Johnstone, R. A. and A. Grafen (1992). The continuous Sir Philip Sidney Game: A simple model of biological signaling. *Journal of Theoretical Biology 156*, 215–236.

Kilner, R. M., D. G. Noble, and N. B. Davies (1999). Signals of need in parent – offspring communication and their exploitation by the common cuckoo. *Nature 397*, 667–672.

Lachmann, M. and C. T. Bergstrom (1998). Signalling among Relatives II: Beyond the Tower of Babel. *Theoretical Population Biology 54*, 146–160.

Lachmann, M., S. Számádó, and C. T. Bergstrom (2001). Cost and conflict in animal signals and human language. *Proceedings of the National Academy of Sciences 98*(23), 13189–13194.

Maynard Smith, J. (1991). Honest Signaling, The Philip Sidney Game. *Animal Behavior 42*, 1034–1035.

Maynard Smith, J. and D. Harper (2003). *Animal signals*. Oxford: Oxford University Press.

Nowak, M. A. and K. Sigmund (1998). Evolution of indirect reciprocity by image scoring. *Nature 393*(June), 573–577.

Ohtsuki, H. and Y. Iwasa (2006, April). The leading eight: social norms that can maintain cooperation by indirect reciprocity. *Journal of theoretical biology 239*(4), 435–44.

Rodríguez-Gironés, M. a., M. Enquist, and P. a. Cotton (1998, April). Instability of signaling resolution models of parent-offspring conflict. *Proceedings of the National Academy of Sciences of the United States of America 95*(8), 4453–7.

Schibler, F. and M. B. Manser (2007, November). The irrelevance of individual discrimination in meerkat alarm calls. *Animal Behaviour 74*(5), 1259–1268.

Searcy, W. A. and S. Nowicki (2005). *The Evolution of Animal Communication*. Princeton: Princeton University Press.

Silk, J., E. Kaldor, and R. Boyd (2000, February). Cheap talk when interests conflict. *Animal behaviour 59*(2), 423–432.

Számadó, S. (1999). The validity of the handicap principle in discrete action–response games. *Journal of theoretical biology 198*(4), 593–602.

Számadó, S. (2008, November). How threat displays work: species-specific fighting techniques, weaponry and proximity risk. *Animal Behaviour 76*(5), 1455–1463.

Számadó, S. (2011a, August). Long-term commitment promotes honest status signalling. *Animal Behaviour 82*(2), 295–302.

Számadó, S. (2011b, January). The cost of honesty and the fallacy of the handicap principle. *Animal Behaviour 81*(1), 3–10.

Tibbetts, E. a. and J. Dale (2007, October). Individual recognition: it is good to be different. *Trends in ecology & evolution 22*(10), 529–37.

Trivers, R. L. (1971). The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology 46*(1), 35–57.

Wagner, E. (2013, April). The Dynamics of Costly Signaling. *Games 4*(2), 163–181.

Zahavi, A. (1975). Mate Selection – A selection for a Handicap. *Journal of theoretical biology 53*, 205–214.

Zahavi, A. and A. Zahavi (1997). *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. New York: Oxford University Press.

Zollman, K. J. S. (2013). Finding Alternatives to Handicap Theory. *Biological Theory 8*(2), 127–132.

Zollman, K. J. S., C. T. Bergstrom, and S. M. Huttegger (2013, January). Between cheap and costly signals: the evolution of partially honest communication. *Proceedings of the Royal Society B: Biological Sciences 280*(1750), 20121878.