Honesty through repeated interactions

Patricia Rich Kevin J. S. Zollman*

Abstract

In the study of signaling, it is well known that the cost of deception is an essential element for stable honest signaling in nature. In this paper, we show how costs for deception can arise endogenously from repeated interactions between individuals. Utilizing the Sir Philip Sidney game as an illustrative case, we show that repeated interactions can sustain honesty with no observable signal costs, even when deception cannot be directly observed. We provide a number of potential experimental tests for this theory which 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 two or more individuals. This occurs even when a first analysis suggests that deception would be fitness enhancing for one of the parties. In order for honest communication to be stable to invasion by dishonesty, there must be some countervailing force which reduces the fitness of deception.

It was originally suggested that the only way to make deception unprofitable would be for the communicating individual to spend a high cost to send the signal – to take on a "handicap" (Zahavi, 1975; Zahavi and Zahavi, 1997). It has since been shown that ubiquitous cost is not necessary to sustain honesty (Hurd, 1995; Számadó, 1999; Lachmann et al., 2001; Számadó, 2011b). Instead, the cost of deception is critical. Honesty can be free, so long as lying is costly. For example, Hurd (1997) showed that reliable communication of fighting ability is possible with very low observed cost so long as the penalty imposed on a weak individual who imitates a strong one is sufficiently high to deter deception – a plausible assumption in animal contests.

The cost of deception – sometimes called "marginal cost" – might not be observed in systems in equilibrium, 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 marginal cost provided in the literature (for examples, 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 marginal cost difficult.

This paper explores the possibility of creating out-of-equilibrium cost, without creating observable costs, in the context of signaling among relatives. We do so 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 will eventually be ignored by the parent and denied food in a way that harms 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 suppose 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.* This restriction distinguishes our model from the few existing models of signaling reputation (Silk et al., 2000; Catteeuw et al., 2014), where dishonesty must be directly observed. In this paper, 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, these results should generalize to other communicative interactions that feature partial conflict of interest.

In section 2 we review the Sir Philip Sidney game and present the various equilibria which exist for different parameter settings in this game. We then modify the game in section 3 and present the central results of the paper. The paper concludes with a discussion of the idealizations and potential empirical tests of the model in section 4.

2 Handicaps in the Sir Philip Sidney game

The handicap principle was initially formulated by the Zahavis (1975; 1997) to explain the presence of honesty in situations where there is an incentive for deception. 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, if signaling conveys relevant information, receivers do best to make use of the accurate information carried by signals. Mathematical 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 the Zahavis' claim that the handicap principle is uniquely able to account for reliable signaling in nature.

The Zahavis' 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. Maynard Smith's (1991) Sir Philip Sidney game provides a relatively tractable example of Grafen's (1990) model of the handicap principle. The game, shown in figure 1, involves two players. These players are typically imagined as a chick and a parent, although the model can be interpreted more generally. At the first node of the game some exogenous force (usually called "nature") determines whether the chick is in need of food or not in need of food. Following Bergstrom and Lachmann (1997) we will refer to these states as "needy" – which occurs with probability p – and "healthy" – which occurs 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 proposed 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 chick who doesn't receive food pays a fitness cost of $0 < a < 1$ if it is needy and $0 < b < 1$ if it is healthy (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 presume that, at a minimum, the parent wishes to

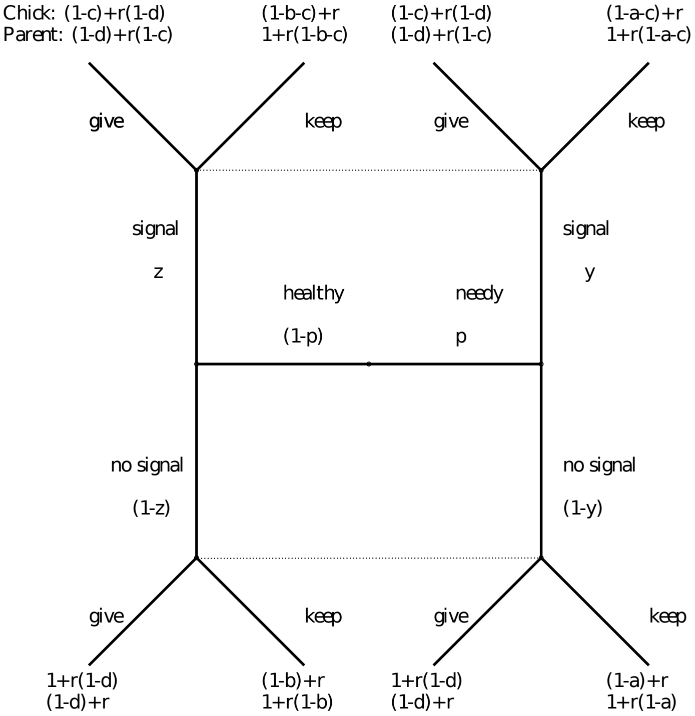


Figure 1: A game tree illustrating the Sir Philip Sidney game with inclusive fitness. "Nature" determines whether the chick is healthy or needy (at the center node). The chick conditions its behavior on this choice and decides whether to send a costly signal or stay silent. The parent conditions its behavior on the signal, but not on the state of need of the chick. The parent chooses whether or not to donate a resource. Inclusive fitness for each individual is derived from adding r multiplied by the other's individual fitness.

transfer the resource to the needy chick, i.e. $d < ra$.

Depending on the value of the parameters, this game features three types of equilibria, illustrated in figure 2. The equilibrium labeled “Signaling” is the state where the neediness of the chick is perfectly communicated to the parent. In addition to the signaling equilibrium, one of the “pooling” equilibria always are present. These are equilibria where no information is communicated and the parent responds by either always transferring or never transferring the resource (which parental response is best depends on the underlying probability that the child is needy). The final equilibrium, the hybrid equilibrium, is not critical to our discussion here (for a discussion of the hybrid equilibrium, please see Huttegger and Zollman, 2010; Wagner, 2013; Zollman et al., 2013). We will return to this equilibrium in section 4.

Holding the parameters a, b, d, and p fixed and allowing r and c to vary defines four regions of interest (Bergstrom and Lachmann, 1997; Huttegger and Zollman, 2010), which are pictured in figure 3. In all four areas, at least one of the pooling equilibria exits and is stable in a weak sense (Huttegger and Zollman, 2010). In region 1, the cost is so high that the chick should not send the signal regardless of its state of need – only pooling equilibria exist in this region. Regions 2 and 3 represent the classic situation for the Sir Philip Sidney game and the handicap principle more generally. Here, when $rb < d$ and $rd < b$, there is 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 imitate the needy chick successfully. 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 a result totally honest communication is impossible. (This is where the hybrid equilibrium exists.)

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. These results are summarized in table 1.

Region 2 is central to discussions of the Sir Philip Sidney game. Here one 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, several models have shown that evolving to one of these signaling equilibria might be 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

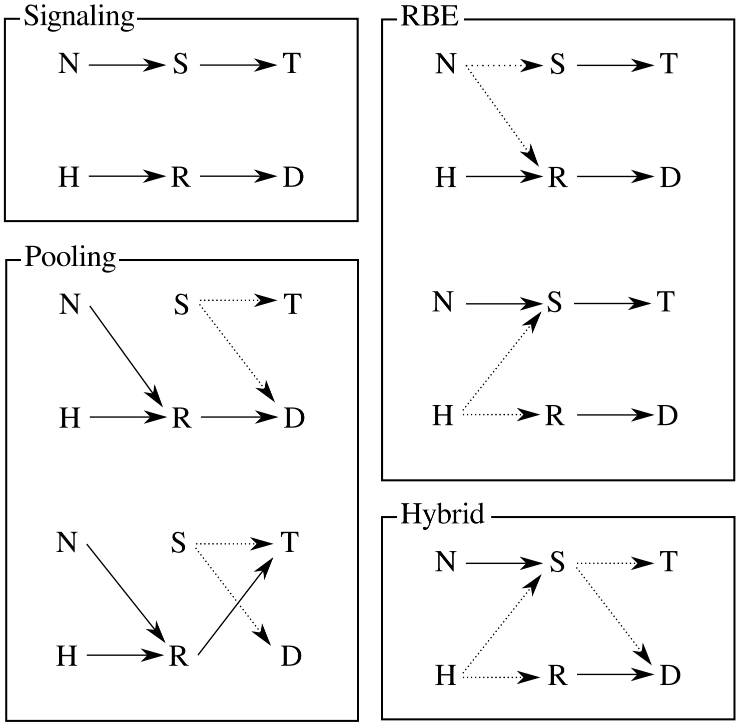


Figure 2: An illustration of the various equilibria of the Sir Philip Sidney game with and without reputation-based strategies. Each diagram represents the strategies for the chick and parent. The nodes H and N represent the state of need of the chick; healthy and needy respectively. The nodes S and R represent the signal sent by the chick, S is "send a signal" and R is "refrain." The arrows from state to signal represent the strategy of the chick. Solid lines represent plays with probability 1 or monomorphic populations while dotted lines represent mixed strategies or polymorphic populations. The nodes labeled T and D are the actions available to the parent, T is "transfer" and D is "don't transfer." The lines from signal to action are the strategies for the parent.

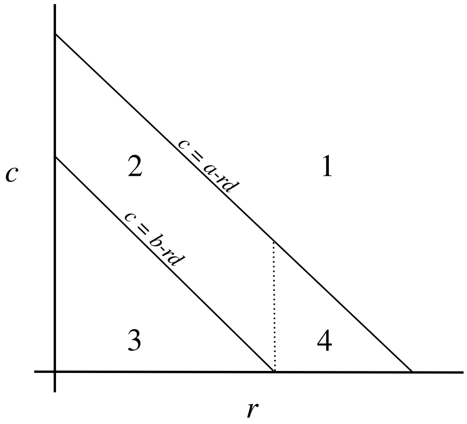


Figure 3: Regions of interest in the Sir Philip Sidney game with a, b, and d held fixed.

	Regions			
Equilibrium	1	2	3	4
Pooling	‡	‡	‡	‡
Cost free honest signaling				∗
Handicap honest signaling		†		‡
Hybrid equilibrium			∗	
Reputation-Based Equilibria		X	X	X

Table 1: Summary of the existence of three equilibria in the different regions of the parameter space illustrated in figure 3. The presence of a symbol indicates that an equilibrium can exist, under some conditions, in that region of the parameter space. The symbol marks the reference for conditions of the equilibria's existence in that region, † for (Maynard Smith, 1991), ‡ for (Bergstrom and Lachmann, 1997), ∗ for (Huttegger and Zollman, 2010), and X for section 3 of the current paper.

who fails to signal engages 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 explainaltruistic 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 a conspecific has behaved altruistically; this assumption has been used in the models of cost-free honest signaling (Silk et al., 2000; Catteeuw et al., 2014). In the Sir Philip Sidney game, this translates 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 cannot be uncritically applied to this case. In the following section, we introduce 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 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 negative reputation. 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.

One approach to modeling this interaction is with the theory of repeated games, where a single stage game is repeated many times (Mailath and Samuelson, 2006). While powerful, this strategy is extremely complex. We will instead opt to idealize the repeated interaction structure by assuming that the parent can effectively learn one component of the chick's behavior, namely the unconditional probability with which the chick signals. We assume that this information is learned by the parent without error (an assumption we discuss in section 4) and that the parent can condition its behavior on this information.

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 use because it represents an idealized proxy for information that the parent can observe – frequency of signaling – and not what is unobservable – the state of need of the chick. A chick acquires a reputation based on its signaling probability in the two states of need, i.e. based on the probability of signaling when needy (denoted by y) and when healthy (denoted by z). A chick's reputation, then, is $F := py + (1-p)z$. We use $B_{y,z}$ to 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 D_x where the chick is given food in response to a signal if and only if $F \leq x$. In other words, the strategy D_x sets a begging quota 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 $(B_{y,z}, D_x)$ with $py+(1-p)z = F \leq x$ that constitutes a Nash equilibrium. (By allowing the parent to condition its strategy on F we assume that the parent has a very accurate assessment of one part of the child's strategy – the unconditional probability of signaling – while remaining ignorant of other parts. This is an idealization we explore in more detail in the discussion.)

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 3). 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 assume that the cost of the signal is below this threshold, that we occupy regions 2, 3, or 4 of figure 3. 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 D_x. The best response for the chick is to choose y and z in order to maximize this equation:

$$p\big(y(1-c+r-rd)+(1-y)(1-a+r)\big)+ \\ (1-p)\big(z(1-c+r-rd)+(1-z)(1-b+r)\big) \quad (1)$$

subject to the constraint that $F \leq x$. (The chick's signaling probability F cannot exceed x as then the chick will sometimes pay the cost of signaling but never receive food.) Now, by the assumption that $a > c+rd$, the payoff for the y cases is higher than 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 D_x, the chick does best by setting y to satisfy this equation,

$$y = \begin{cases} 1 & \text{if } x \geq p \\ \frac{x}{p} & \text{otherwise} \end{cases} \quad (2)$$

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. (We henceforth ignore parameter settings that lie on the boundary between two regions.)

3.1 Regions 2 and 4

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 signal profitably. 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 when it refrains from signaling than when it pays the cost to secure 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 D_x is:

$$y = \begin{cases} 1 & \text{if } x \geq p \\ \frac{x}{p} & \text{otherwise} \end{cases} \tag{3}$$

and $z = 0$.

This presumes that the parent adopts a conditional transfer strategy D_x. Alternatively, if the parent chooses 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 chick's best responses to a parent strategy, D_x in regions 2 and 4. Figure 4(a-c) illustrates the fitnesses of various strategies for different parental responses.

In order to fully characterize the equilibrium behavior, we must now consider the parent's response to a chick who adopts a strategy $B_{y,z}$ which yields a signaling probability of F.

Suppose the chick adopts a strategy $B_{y,z}$. First note that D_x and $D_{x'}$ are behaviorally equivalent for the parent when $x, x' \geq F$ (both transfer the resource when the signal is received). Similarly when $x, x' < F$, D_x and $D_{x'}$ are behaviorally equivalent to the strategy *never transfer*. As a result we only must compare three strategies: D_x, where $x \geq F$, *always transfer*, and *never transfer*.

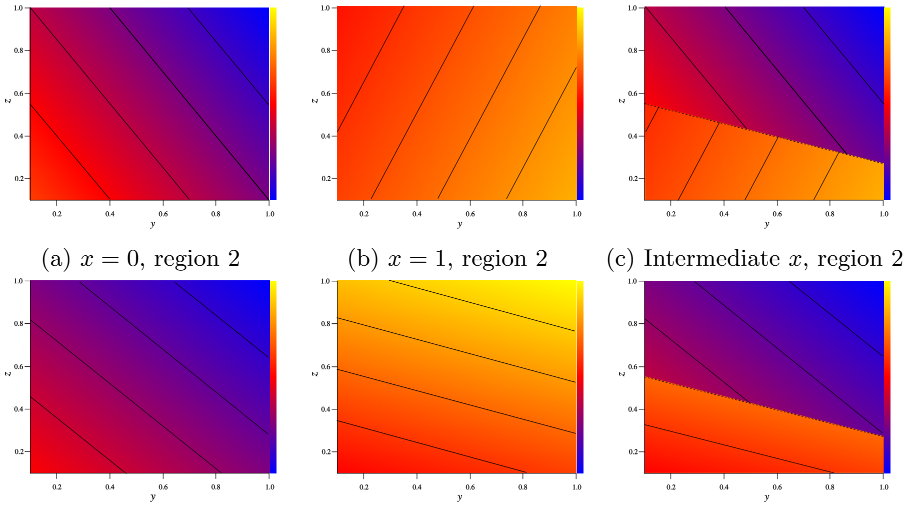
Suppose the chick pursues a strategy $B_{y,z}$ such that $z = 0$. Because $d < ra$, the parent wishes to transfer the resource to the needy chick, but because $d > rb$ the parent does not wish to transfer to the healthy chick. The strategy D_x (where $x \geq py$) will transfer the resource to the needy chick with probability y and will never transfer to the healthy chick. As a result, it is superior to the strategy *never transfer* whenever $y > 0$.

D_x yields at least as high a payoff as the strategy *always transfer* when,

$$\begin{aligned} &p(1-y)(1+r-ra) + (1-p)(1+r-rb) \\ &\qquad \geq p(1-y)(1-d+r) + (1-p)(1-d+r). \end{aligned} \tag{4}$$

Since it is required that $a > d/r$, this is satisfied when,

$$y \geq 1 + \frac{(1-p)(br-d)}{p(ar-d)}. \tag{5}$$



(a) $x = 0$, region 2

(b) $x = 1$, region 2

(c) Intermediate x, region 2

(d) $x = 0$, region 3

(e) $x = 1$, region 3

(f) Intermediate x, region 3

Figure 4: Fitness contour plots with isoclines (solid lines) for the chick in the y, z-space . Plotted values are for $a = 0.9, b = 0.5, c = 0.3, d = 0.3, p = 0.25,$. (a-c) are for region 2, where $c = 0.4$. (d-f) are for region 3, where $c = 0.3$ (a and d) illustrate the fitness of the chick when the parent never transfers ($x = 0$). (b and e) illustrates the fitness of the chick when the parent transfers whenever a signal is sent, regardless of the chick's strategy ($x = 1$). (c and f) illustrates the fitness of the chick when the parent adopts an intermediate value of x. The discontinuity, which occurs along the dotted line, is caused by the sudden transition in parental response when the chick's signal probability F crosses the threshold x.

When this inequality is satisfied the strategy D_x where $x \geq F = py$ maximizes fitness for the parent. Otherwise, the strategy *always transfer* yields a higher fitness to the strategy D_x.

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. Pooling equilibria – where the chick never signals and the parent always transfers – also remain.

The reputation game has introduced 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.

3.2 Region 3

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 in which some experiments appear to, somewhat paradoxically, place actual interactions between parents and offspring (Searcy and Nowicki, 2005). Our most significant results are found here.

Suppose the parent adopts a strategy D_x where $x = p$, which will only tolerate the chick signaling with probability p. p is the probability with which the chick is needy, and by equation 2 we see the optimal strategy for the chick is $y = 1$ and $z = 0$, signaling only when it is needy. The parent is now transferring the resource if and only if the chick is needy, which given our earlier assumptions is optimal for the parent. Therefore, this strategy pair 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. Adding 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 is 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 probability allowed by the parent.

This allows us to fully characterize the chick's strategy in region 3. If the parent adopts a strategy D_x, the chick will choose $B_{y,z}$ such that:

$$y = \begin{cases} 1 & \text{if } x \geq p \\ \frac{x}{p} & \text{otherwise} \end{cases} \tag{6}$$

and,

$$z = \begin{cases} \frac{x-p}{1-p} & \text{if } x \geq p \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Figure 4(d-f) presents fitness isoclines for the chick to various parental strategies in region 3.

Suppose that the parent adopts a strategy D_x and the chick adopts a strategy $B_{y,z}$ that satisfies those constraints. Does the parent's strategy represent an optimal choice? We only must compare the strategy D_x where $x = F$ to the strategies *always transfer* and *never transfer*. Given the chick's strategy, the parent's payoff of adopting D_x is

$$p\big(y(1-d+r-rc)+(1-y)(1+r-ra)\big) \\ +(1-p)\big(z(1-d+r-rc)+(1-z)(1+r-rb)\big). \tag{8}$$

The payoff for the strategy *never transfer* is:

$$p\big(y(1+r-ra-rc)+(1-y)(1+r-ra)\big) \\ +(1-p)\big(z(1+r-rb-rc)+(1-z)(1+r-rb)\big). \tag{9}$$

Because $br < d$, when the chick is healthy (the $(1-p)$ terms) the parent prefers *never transfer.* However, when the chick is needy, the parent prefers D_x to *never transfer* when $(1-d)+r(1-c) > 1+r(1-a-c)$ which is true because $ar > d$. As a result, the parent prefers D_x to *never transfer* when $py(ar-d)+(1-p)z(rb-d) \geq 0$. In words, this constraint requires that the chick must not signal too frequently when healthy. When the healthy chick signals too frequently, and 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

$$p\big(y(1-d+r-rc)+(1-y)(1-d+r)\big) \\ +(1-p)\big(z(1-d+r-rc)+(1-z)(1-d+r)\big). \tag{10}$$

When the chick signals (with the py and $(1-p)z$ terms) the payoff for *always transfer* is equivalent to D_x. D_x performs, when the chick is healthy and does not signal (the $(1-z)$ term), better. When the chick is needy and does not signal (the $(1-y)$ term) *always transfer* performs better. Overall D_x performs better when $p(1-y)(d-ar)+(1-p)(1-z)(d-br) \geq 0$. This constraint 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.

These two conditions can be written as,

$$\frac{y}{z} \geq \frac{(1-p)(d-br)}{p(ar-d)} \geq \frac{1-y}{1-z} \tag{11}$$

When the chick adopts a strategy $B_{y,z}$ that satisfies equation 11, then a parent strategy $x = F$ is optimal. If either inequality is violated the parent prefers either to transfer to the chick regardless of the signal, or to withhold the resource regardless of signal. 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 that the parent adopts such a strategy, the chick too is optimizing by choosing $B_{y,z}$. 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 reputation-based equilibria (RBE). 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.

We have shown that RBE enable totally honest signaling in equilibrium without *any* appreciable signal costs. This occurs when the parent can condition its behavior on the probability of signaling by its offspring. The parent can thus impose a cost that would only be observed when the system was out of equilibrium – when the parent or chick is manipulated to give the impression of too-frequent signaling.

Our model assumes that a parent can observe something which is correlated with the probability with which its offspring is signaling. Without this ability, parents could not adopt a reputation-based strategy and RBE would be impossible. To do so requires individual recognition of offspring in species with brood sizes greater than one. Assuming recognition among offspring is plausible; individual recognition has been found in several species including recognition via begging calls (Lefevre et al., 1998; Insley et al., 2003; Draganoiu et al., 2006). But mere recognition is not sufficient; parents must be able to condition their behavior on the past begging frequencies of the offspring. This may not always be possible. For example, Kilner et al. (1999) found that the common cuckoo (*Cuculus canorus*) manipulates reed warbler (*Acrocephalus scirpaceus*) parents by signaling more frequently than reed warbler chicks. But in other cases of signaling, animals can differentiate between signalers and can respond based on past signaling behavior. 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 not all species who signal have the ability to recognize individuals or to condition on signaler's past behavior, we do not present this model as the explanation for honesty in *all* signaling interactions. We view it as one of many potential mechanisms whereby honesty is maintained.

Furthermore, our model utilizes a particular idealization, that parents have direct access to a consequence of the chick's strategy – they can observe the probability of signaling (F) directly. This, of course, will never occur; it is an idealization of our model. Instead, a parent will "estimate" F from its sample of past behavior. For instance, the parent might observed how often the chick begs in the presence of the parent. A more realistic model would need to account for errors in the parent's estimate of F caused by parent's sampling of the chick's actual strategy. The accuracy of the parent's perception of signaling probability will depend on the length of the parent's memory, on its ability to accurately recognize its offspring, and on its ability to correctly identify if a chick is signaling or not. This could be achieved in this model by adding a noise function that alters how the parent perceives actual choice of F by the chick. In such a case, a chick might do best by adopting a strategy which signals slightly less often than allowed by the parent in order to ensure that errors would not cause the chick to be perceived as violating the begging quota. While the details of such a model will be more complicated, RBE would likely exist in these more realistic models. We leave development of such a model for future work.

Along the same lines, we assume that the probability that the chick is needy is stationary – that is p is fixed and unchanging. In reality this parameter may change with environmental condition and age of the chick. For example in species where the young can forage for themselves, the probability of need may decrease as the child ages (Smiseth et al., 2003). In response to this, parents might need to adjust their thresholds over time or in different environments. Of course, such adjustments would depend on the parent's ability to recognize these other variables. Not only does the environment and age of the child affect the value of p, but it is influenced by the parent's past behavior. Offspring who are underfed will be needy more often than those who have been fed recently. In this paper we have followed the norm in the signaling literature of ignoring the complex interaction between the environment, development, and past parent behavior on the current state of need. Future work on this model, and others in the signaling literature, should explore the effect of this idealization on our understanding of the evolution of signaling.

Finally, this model, like many others in the signaling literature, assumes a two-player interaction – one parent and one chick. The harm incurred by the parent from transferring the resource comes from a reduction in the number of future offspring in later breeding periods. Costs like this might be endogenized by introducing several offspring who are competing (i.e. Godfray, 1995). It remains an open question whether reputation-based equilibria would exist in this context. We, however, see no reason why they would be excluded.

RBE are consistent with several observed features in biological systems. Low signal costs are consistent with, but not required by RBE. Similarly, occasional signaling by both needy and healthy individuals are consistent with RBE. The RBE theory does have empirical predictions, however. First, if any healthy individuals signal, all needy individuals should as well. The RBE theory would be contradicted if healthy individuals signaled more frequently than needy individuals. Second, signalers which are experimentally manipulated to signal too frequently should eventually be ignored by the parent entirely.

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 is necessary to test how individual parents respond to extravagant amounts of signaling. As is the case with all research of this kind, particular assumptions must be made about the time scales on which the parent will condition its behavior. Such assumptions would be motivated by cognitive sophistication of the particular species under consideration, and would thus vary on a case-by-case basis.

RBE bear a superficial similarity to the so-called "hybrid equilibria" (pictured in 2 and discussed in Huttegger and Zollman, 2010; Wagner, 2013; Zollman et al., 2013). Both the hybrid equilibrium and some RBE feature partially honest communication and can 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 RBE the parent always transfers the resource when the signal is observed.

Beyond parent–offspring interactions, we believe that these equilibria will be present in different models of signaling. Therefore, RBE might explain honesty in other types of interaction. What is required is that the two parties interact repeatedly and are capable of recognizing each other (see Tibbetts and Dale 2007 for a discussion of the evidence for individual recognition in many different contexts).

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 outside of equilibrium – that is critical (Hurd, 1995; Számadó, 1999; Lachmann et al., 2001). In these RBE the cost is imposed when the parent punishes 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 withheld resources, 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 equilibrium and thus require different empirical tests.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. EF 1038456. We are indebted to several anonymous referees for helpful suggestions on earlier drafts of this manuscript.

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, apr). 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, sep). 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.

Draganoiu, T. I., L. Nagle, R. Musseau, and M. Kreutzer (2006). In a songbird, the black redstart, parents use acoustic cues to discriminate between their different fledglings. *Animal Behaviour 71*(5), 1039–1046.

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

Godfray, H. C. J. (1995). Signaling of need between parents and young: parent-offspring conflict and sibling rivalry. *American Naturalist 146*, 1–24.

Gouzoules, H., S. Gouzoules, and K. Miller (1996). Skeptical Responding in Rhesus Monkeys (¡i¿Macaca mulatta¡/i¿). *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, dec). 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.

Insley, S. J., R. Paredes, and I. L. Jones (2003). Sex differences in razorbill Alca torda parent–offspring vocal recognition. *J Exp Biol 206*(1), 25–31.

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. Számadó, 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.

Lefevre, K., R. Montgomerie, and A. J. Gaston (1998). Parent – offspring recognition in thick-billed murres (Aves : Alcidae). *Anikmal Behavior 55*, 925–938.

Mailath, G. J. and L. Samuelson (2006). *Repeated Games and Reputations.* Oxford University Press.

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

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, apr). 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, apr). 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, nov). 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, feb). Cheap talk when interests conflict. *Animal behaviour 59*(2), 423–432.

Smiseth, P. T., C. T. Darwell, and A. J. Moore (2003). Partial begging: an empirical model for the early evolution of offspring signalling. *Proceedings. Biological sciences / The Royal Society 270*(1526), 1773–1777.

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

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

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

Tibbetts, E. a. and J. Dale (2007, oct). 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, apr). 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, jan). Between cheap and costly signals: the evolution of partially honest communication. *Proceedings of the Royal Society B: Biological Sciences 280*(1750), 20121878.