Evolutionarily Stable Communication and Pragmatics

Thomas C. Scott-Phillips

Language Evolution and Computation Research Unit
School of Psychology, Philosophy and Language Sciences
University of Edinburgh

In the past 20 or so years there has been much research interest in the evolution of cooperation in humans (Axelrod, 1995; Boyd & Richerson, 1992; Fehr & Fischbacher, 2003; Milinski et al., 2002; West et al., 2006). The foundational problem addressed by this work is how cooperation can remain evolutionarily stable when individuals have incentives to freeride; that is, to take but not contribute from the public good (Hardin, 1968). There is an analogous problem associated with the evolution of communication: how can signalling remain evolutionarily stable when individuals have incentives to be dishonest? This game-theoretic question is the defining problem of animal signalling theory (Maynard Smith & Harper, 2003; Searcy & Nowicki, 2007). The main goals of this chapter are to explore the various possible solutions to this problem and to ask which most likely applies to human communication. In addition to this it will also, using insights from pragmatics, provide some insight as to the nature of the problem and hence clarify some of the relevant issues.

It is somewhat remarkable that the question of the evolutionary stability of human communication has historically received little interest relative to the attention given to the evolution of cooperation and the burgeoning literature on the evolution of language. In the last 15–20 years both have expanded dramatically. Language evolution in particular has grown from a niche interest into a well-recognised academic discipline in its own right, with regular conferences, an ever-increasing number of papers on the topic (Google Scholar returns[1] 13,800 hits for the search *language evolution* in 1990, increasing almost monotonically to 54,400 in 2005), and special issues of relevant journals (e.g. *Lingua*, volume 117(3), 2007; *Interaction Studies*, volume 9(1), 2008). It would be reasonable to assume that solutions to the problem of evolutionarily stable communication in humans would be a central explanandum for such a discipline, but that is not the case: very few papers have made this question a central focus (exceptions include Knight, 1998; Lachmann et al., 2001; Scott-Phillips, 2008; Szmad & Szathmry, 2006). There has thus been only limited progress beyond speculative discussion, and the contrast with developments in the evolution of cooperation is striking.

This chapter begins with the observation that although there are important equivalences between the problems of cooperation and communication, it can be misleading to think about the latter exclusively in terms of the former, as that masks the fact that there are in fact two problems associated with the

[1] Searched on 22nd May, 2008.

A. Benz et al. (Eds.): Language, Games, and Evolution, LNAI 6207, pp. 117–133, 2010.

evolutionary stability of human communication, rather than the one in cooperation. These problems are: (i) how can we know that a signal means what we take it to mean?; and (ii) how can we have trust in that meaning? This division is made all the more clear once we recognise that it maps directly onto a distinction that is central to pragmatics; that between *communicative* and *informative intent*. Indeed, the relationship between these two interdependent aspects of communication shines valuable light on the matters at hand. Thus while this chapter in general seeks to explore how evolutionary considerations can inform pragmatic concerns, there are also valuable lessons that pass in the opposite direction.

How can these two problems be solved? The possible answers to this question are critically evaluated and classified. Accordingly, we see that one answer in particular fits with our instinctive ideas about how social contracts work: unreliable and dishonest communication is deterred because a reputation for such behaviour is socially maladaptive. Despite its intuitiveness, this idea has not been empirically tested, and as such represents a potentially fruitful topic for future research.

1 The Problems of Evolutionarily Stable Communication

The problem (note the singular — the inconsistency with the title of this section will become clear shortly) may be simply stated: if the signaller can gain more from an unreliable or dishonest signal than a reliable or honest one then we should expect just such signals to evolve. If the receiver's payoff to responding to such a signal is negative, as seems reasonable, then we should expect the receiver to evolve not to attend to the signal. Now that the receiver is not attending to the signal there is no possible benefit to the signaller, and so they will evolve not to produce the signal at all (if nothing else, there are likely to be metabolic and opportunity costs associated with signal production (Maynard Smith & Harper, 1995), and hence a pressure not to incur such costs if there is no consequent payoff). The system has now collapsed in much the same way as it does in Aesop's fable of the Boy Who Cried Wolf, in which the shepherds learnt not to attend to the boy's calls, since they were so frequently dishonest (Maynard Smith, 1982).

That makes the problem sound like a conceptually straight-forward one, and in many ways it is. However for human communication the matter is more complex, since there are two (analogous) problems rather than one. At one level communication is an inherently cooperative act: there must be some agreement about what signals refer to what phenomena in the world — there must be a shared agreement on the mappings between 'meaning' and form. (I put 'meaning' in scare quotes only because it is not clear what it might mean for an animal to have meanings in any recognisable sense of the term, a point that is expanded on below.) At another level that signal must be something that the audience can place their trust in, so that they are not misinformed in any way. And then, of the course, at a third level the goals to which communication is applied may

be more or less cooperative: two individuals with a shared goal will use communication for cooperative ends, but two individuals with mutually incompatible goals will use it antagonistically. Importantly, however, for it to be even used antagonistically it must already be cooperative in the first two senses. An explanatory analogy between the first and third levels is with a game of tennis (or indeed any competitive sport). To even be able to play tennis with each other we must both recognise the rules of the game and play within them; refusal to do so means that we cannot even play a meaningful game at all. In the context of communication we can call this communicative cooperation: interlocutors must agree upon the meaning of a signal. However, once we have agreed to play by the rules of tennis we will, if we are intent on winning the game, play as uncooperatively as possible, pushing the ball to the corners of the opponent's court and generally trying to force errors in their play. This is material cooperation (or rather: non-cooperation), and within communication it is entirely optional.

This distinction between the first and third levels has been previously outlined (Hurford, 2007). My suggestion is that we also recognise another type of cooperation involved in communication, nestled between these two: the honest use of signals. Interestingly there is, for the pragmatician, an obvious term for this type of cooperation: informative cooperation. The reason it is obvious is that it recognises the distinction, central to pragmatics, between an individual's informative intent and their communicative intent. To outline: the former refers to the speaker's intention to inform the listener of something, and the latter to the speaker's intention that the listener recognise that they have an informative intention (Grice, 1975; Sperber & Wilson, 1995). Pragmatics thus recognises that when a speaker produces an utterance they do not just intend that the listener understand whatever it is they are talking about, but also that they intend that the listener understand that the utterance is an act of communication designed to achieve an informative intention. We can thus distinguish between the communicative layer, which is about the fact that there is a coherent communicative act, and which requires a reliable mapping between meaning and form; and the informative layer, which is about the fact that the content of the utterance is a reliable guide to the world, and which requires honesty on the part of the speaker.

Correspondingly, we have two types of cooperation necessary for communication: communicative cooperation and informative cooperation. We also have a third, entirely optional type: material cooperation. For example, when I lie to my colleague I am reliable but dishonest; but when she argues with me and in doing so prevents me from doing my work she may be both reliable and honest but is materially uncooperative. As necessary conditions, the first two layers demand evolutionary explanation. Indeed, in many respects the evolutionary stability of cooperative enterprises is the defining problem of social evolutionary theory (Axelrod & Hamilton, 1981; Frank, 1998; Maynard Smith, 1982; West et al., 2007). There are two problems to be addressed, then: one regarding how signaller and receiver can agree upon a shared 'meaning' for a given signal (communicative cooperation); and another about whether the signaller uses that meaning in an honest way (informative cooperation). To distinguish between the two

problems, and to disambiguate between two terms that have previously been used synonymously, I suggest that the former problem be termed the problem of signal reliability, and the latter the problem of signal honesty. The difference is depicted in figure 1. These two problems are formally equivalent; that is, they have the same logical structure. As a result the possible solutions are identical too. Of course, this does not mean that the two problems need actually have the same solution — it is perfectly possible that the problem of reliability will be solved differently to the problem of honesty in any particular case.

Before we ask about the possible solutions to these problems, I want to comment briefly on why these distinctions have not previously been recognised by animal signalling theorists. One key difference between humans and other animals is that humans exercise what has been termed *epistemic vigilance* (Sperber & Wilson, 2008): once we comprehend utterances we can evaluate whether or not we consider them true. This distinction between comprehension and acceptance does not, in general, seem to apply to other animals; once informed, they act (but see below). Importantly, the distinction maps directly onto the previously identified distinction between communicative and informative cooperation. Communicative cooperation is a matter of whether or not signals are reliable (that is, whether individuals share the same signal-form mappings), and once it is achieved then signals become comprehensible. Similarly, informative cooperation is a matter of whether or not signals are honest, and once that is achieved then receivers can accept them as true, and are thus worth attending to. When receivers do not, or rather cannot, exercise epistemic vigilance then these two problems collapse into one. There are, of course, some occasions in which non-humans do seem to exercise epistemic vigilance to at least some degree (Cheney & Seyfarth, 1990). In such cases, we have two problems to solve rather than the usual one studied by animal signaling theorists.

What term should we use to refer to the situation when honesty and reliability collapse into a single problem? Both *reliability* and *honesty* seem to depend upon a coherent notion of the meaning of signals: reliability is a problem about a disjoint between the meaning-form mappings held by different individuals, while the notion of honesty seems to presuppose that a signal has a propositional meaning whose truth-value can be assessed. However it is at best unclear whether it is coherent to talk about animal signals having meanings in the same way that human utterances do. Despite this, I will not suggest an alternative term, for at two reasons. First, a suitable alternative is not forthcoming; and second, the two terms are in such widespread use in the animal signalling literature, with little if any apparent confusion, that redefinition seems both unwise and unlikely to succeed. On the contrary, the use of anthropomorphic gloss is a common strategy in behavioural ecology and social evolution (Grafen, 1999). Of the two, honesty seems the more preferable, if only because it seems to be the more common. This is perhaps because it is the more theoretically interesting: it is hard to see what payoffs could be attained through unreliable communication (if one cannot be understood then why should one signal at all?), but the potential payoffs to dishonesty are clear.

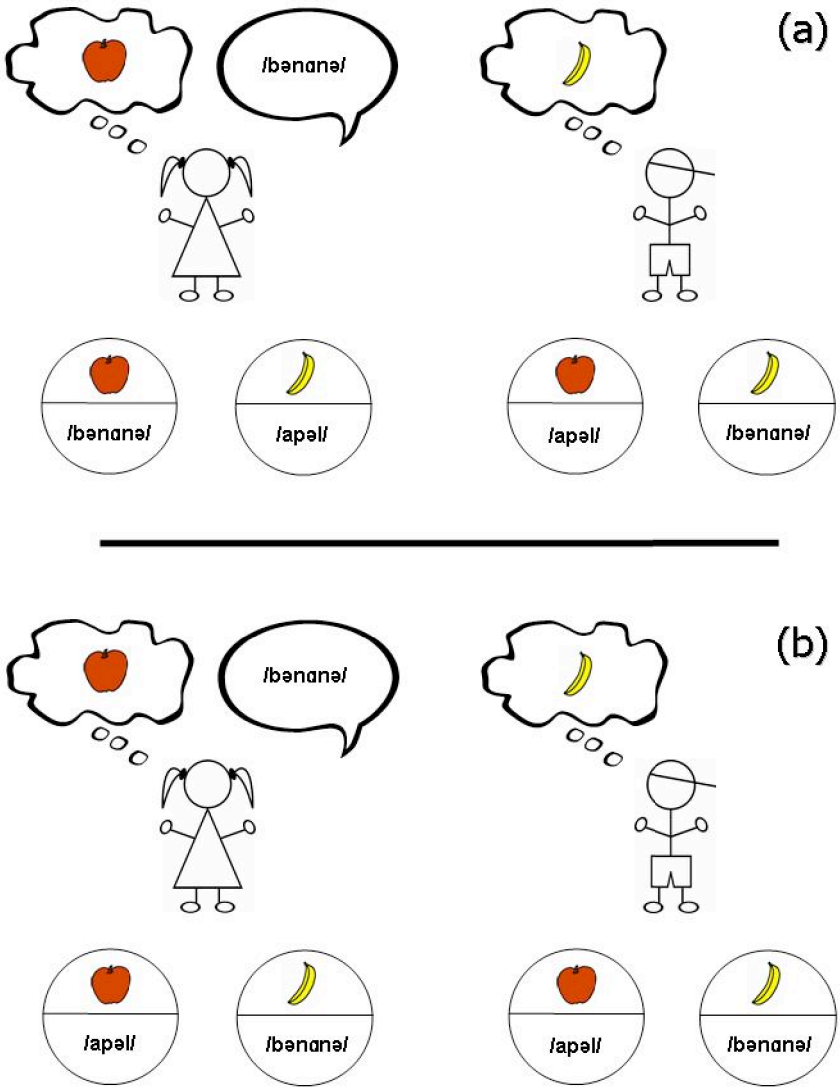

Fig. 1. The twin problems of (a) reliability; and (b) honesty. In both cases the girl has said "banana" having thought of an apple, and this fails to correspond to the boys mapping of the sound (which is as per the convention in English). However the reasons for this failure are different in each case. In (a) the girl has a different (in fact, the precise opposite) mapping from sounds to meaning than the boy, and this makes her unreliable. In (b) she has the same mappings as the boy but has chosen to communicate a different meaning than the one she has thought of, and this makes her dishonest.

Table 1. The different types of cooperation involved in communication

type of cooperation	gloss	corresponding evolutionary problem
communicative	Do interlocutors have the same meaning-form mappings as each other?	reliability
informative	Does the signal carry information that is worth the receivers attention?	honesty
material	Is communication being used to achieve mutually beneficial goals?	none

This section has introduced and discussed a number of other terms, and so it seems appropriate to summarise them and their relationships to each other. Table 1 does this. The nature of the problems of evolutionarily stable communication should now be clear, and we can thus ask about possible solutions.

2 Solutions to the Problems of Evolutionarily Stable Communication

We turn now to possible solutions. *Inclusive fitness theory* (Grafen, 2006; Hamilton, 1964), or *kin selection* (Maynard Smith, 1964), is the most significant contribution to evolutionary theory since Darwin. Haldane's quip that he wouldn't give up his life for one of his brothers, but that he would for both of them, or eight of his cousins, nicely captures the basic idea: that since many of my genes are shared with my relatives, it is in my own genetic interests to help them. This insight is captured by Hamilton's simple rule, that altruistic behaviours will favoured if the cost incurred by the actor is outweighed by the benefit to the recipient times the degree of relatedness between the two individuals: $br > c$. If this inequality is satisfied then dishonest or unreliable behaviour should be unexpected. Accordingly, there are many instances of kin-selected communication in nature, most obviously among eusocial insects. A related point is that spatial organisation is important: the individuals with which an organism communicates are not chosen at random, but instead tend to be those that are nearby. The degree to which populations disperse themselves is measured in terms of *viscosity*, a notion that is closely tied to that of kin selection: if viscosity is high, meaning that there is limited dispersal, then over time individuals in the same area tend to be related to one another, and can hence ensure stable communication due to Hamilton's rule.

The most famous explanation of how non-kin can maintain stable communication is the *handicap principle*. Although introduced to animal signalling theory in the 1970s (Zahavi, 1975), and almost simultaneously to economics (Spence, 1973), the idea goes back much further, at least to 19$^{\text{th}}$-century sociological discussions of the conspicuous consumption of the leisure class (Veblen, 1899). Indeed, such expenditure serves as a nice illustration of the basic idea: the purchase of expensive, conspicuous objects (Ferraris, Tiffany jewellery, etc.) advertises to onlookers that the purchaser can afford to make such purchases, and

therefore must be well-off; the cost of the objects is a handicap that only the most affluent can afford. As a further example, while writing this article I came across the following passage in a newspaper article about stock market trading in the City of London: "Certain clients even expected such behaviour [drinking and drug taking, often to excess, during working hours] from their brokers, viewing their antics as proof that they were so good at their job, they were given free rein to behave as they pleased" ("Tricks of the traders", 2008). Similarly, large tails make peacocks less dexterous and slower than they would otherwise be. Only the highest quality peacocks can afford such a handicap, and hence the peacock tail acts as a reliable indicator of quality. Consequently the peacock tail has become the exemplar *par excellence* of the handicap principle. The idea was originally met within evolutionary biology with some skepticism, with a number of models and arguments produced that purported to show that it was unlikely to work (e.g. Maynard Smith, 1976), but that changed once a formal proof of its stability was published (Grafen, 1990). There was no such similar skepticism within economics; on the contrary, its proponent was awarded the Nobel prize in part for his articulation of the idea.

Although initially paradoxical, once grasped the logic of the handicap principle is often recognised as an ingenious solution to the problem of evolutionary stability. Perhaps for this reason, it has sometimes been assumed (e.g. Knight, 1998) that it is the only process by which we might stabilise communication, and hence that if communication is to remain stable then signals *must* incur costs over-and-above those necessary to actually produce the signal in the first place. This distinction between the costs that are necessary to produce the signal and any additional costs that are paid as a handicap is captured by the terms *efficacy costs* and *strategic costs* respectively (Maynard Smith & Harper, 1995). The handicap principle is in essence a statement that communication can be stabilised by the payment of strategic costs. However, although they are sufficient, it is not the case that strategic costs are necessary for stability. On the contrary, such a claim is false both theoretically and empirically: several alternative processes have been identified by animal signalling theorists (Maynard Smith & Harper, 2003), and there are many instances of signalling in nature in which no strategic costs are paid: the status badges of male passerines (Rohwer, 1975; Whitfield, 1987); the display of an already fertilised ovipositor by female fruit flies (Maynard Smith, 1956); the hunting calls of killer whales (Guinet, 1992); and, of course, human language are just a few of the many examples (for more, and more details on the listed examples, see Maynard Smith & Harper, 2003).

One further point should be emphasised, as it will be of critical importance later: it is the signal itself that must incur the strategic costs. The handicap principle is unstable if the costs are transferred onto some other associated behaviour. That is, there must be a causal relationship between signal form and the cost incurred. For example, there is no strategic cost associated with the size of a male passerine's badge of status (Rohwer, 1975; Whitfield, 1987), the size and colouration of which correlates with the bird's resource holding potential (a composite measure of all factors that influence fighting ability (Parker, 1974)).

However low-status birds that have large badges will incur the costs of getting into fights they cannot win (Rohwer & Rohwer, 1978). To call such a scenario a handicap seems to render the notion of a handicap far too general, a point that will be expanded upon below, where the formal difference between this scenario and the peacock's tail is made clear. Despite this, the notion of a handicap has been used in this more general sense. For example, the two previous suggestions about how the handicap principle might be relevant to the evolution of language, discussed in the next section, are precisely examples of where the costs are not associated with the signal itself but instead with its social consequences.

What are the alternatives to the handicap principle? *Indices* are causal associations between signal meaning and signal form. This link precludes even the possibility of unreliability or dishonesty. An example is the roar of red deer, where formant dispersion is reliably (negatively) correlated with the deer's size (Reby & McComb, 2003) as an inevitable consequence of the acoustics of the deer's vocal apparatus. The deer's larynx descends upon vocalisation, and the comparative evidence suggests that this is the result of a selection pressure to exaggerate one's size (Fitch & Reby, 2001). However that process seems to have gone as far as it can without compromising other aspects of the deer's anatomy (*ibid.*). As a result it is actually impossible for the deer's roar not to carry reliable information about its size and hence its social dominance; deer can lower their larynx no further, and hence the formant dispersion of their vocalisations is unfakeable. Other examples of indices include male jumping spiders, who expose the ventral surface of their abdomen as an indicator of their current condition (Taylor et al., 2000) and snapping shrimps, who advertise their claws to each other as a way to avoid physical conflict (Versluis et al., 2000) (again, for more examples and more details on these examples, see Maynard Smith & Harper, 2003). To state the idea of an index in formal game-theoretic terms, signals can be free of strategic costs and evolutionarily stable so long as the efficacy cost of the signal is a function of the trait in question (e.g. a function of size, in the red deer example) (Hurd, 1995; Lachmann et al., 2001; Szmad & Szathmry, 2006).

There is some potential for confusion here. Indices and handicaps are supposed to be mutually exclusive, yet on the one hand handicaps are stable only if the strategic costs associated with the handicap are tied to signal form, but on the other hand indices are defined by a causal relationship between signal meaning and signal form. So we have a chain of associations from strategic costs to meaning (handicaps), and from meaning to form (indices). Strategic costs are thus associated with form, and the difference between a handicap and an index becomes unclear. To retain the distinction we must be more precise in our terminology: a handicap is indexical of signal *cost*, while what we would normally call an index is indexical of signal *meaning*.

Moving on, are there any non-indexical solutions to the problem of evolutionarily stable communication? That is, what explanations are available when a signal is free of strategic costs and is not indexical of meaning? Several possibilities have been identified, but there is an open question about how best to categorise them. One classification (Maynard Smith & Harper, 2003) suggests a three-way

division between coordination games, repeated interactions, and punishment. *Coordination games* are those in which in which some common interest overrides any conflicting motivations the participants might have (Silk et al., 2000). The classic example is the 'War of the Sexes': the husband wants to go to the pub for the evening and the wife wants to go to the theatre, but they share an overriding common interest that whatever they do they want to do it together. A real-world example is courtship in fruit flies, who mate only once in their lifetime. If a male attempts to court a female after this mating she will display her ovipositor to him and thus advertise that his efforts are futile. He then ceases courtship immediately (Maynard Smith, 1956). In this way the female's signal saves both of them wasting time. Formally such games can be settled only if there an asymmetry in the relationship such that one player or the other backs down, and if this asymmetry is known to both players (Maynard Smith, 1982). In a *repeated interaction* the longer-term payoffs of honesty may outweigh the short-term payoff of dishonesty (Silk et al., 2000), and hence the problem should not arise. Repeated interactions are more likely in viscous populations, a point highlighted in the literature on cooperation but not much considered with respect to communication (but see Grim et al., 2006; Skyrms, 1996). Indeed, repeated interactions are a candidate explanation for both communication and cooperation (it is, after all, the basic logic behind reciprocity, in which individuals trade what would otherwise be altruistic and hence evolutionarily unstable behaviours (Trivers, 1971)). Finally there is *punishment*, in which one individual actively punishes another for unreliable/dishonest signalling (Clutton-Brock & Parker, 1995). This will of course act as an incentive against such behaviour, but this only really moves the problem on to a different locus, since we must now ask why punishing behaviour will evolve if it is itself costly. Indeed, it seems to be a prime candidate to fall foul of the tragedy of the commons, since all individuals get an equal share of the payoff (stable communication) but can let others pay the costs of punishing that are necessary to obtain it.

On the surface this three-way distinction seems a reasonable one, but in fact it is not at all clear how we should distinguish between the three. If, for example, an individual abuses a system of mutually-beneficial repeated interactions and their partner then refuses to re-enter into the relationship, should we classify that refusal as the collapse of an arrangement of repeated interactions or as punishment? At a functional level the two behaviours are identical: they are an adaptive reaction to a partner that has abused a previously established relationship. Any criteria we use to distinguish between them must therefore be to do with mechanisms rather than functions. Such distinctions are desirable, but are not suggested here, since it would require a full review of possible mechanisms, a project that is outside the purview of the present article. At the same time, we should recognise when behaviours are functionally equivalent. For that reason I suggest a high-order classification of *deterrents*. In general, deterrents refer to the situation where reliable/honest communication is cost-free, but where dishonesty is costly. It can be shown that not only is such an arrangement stable, but that where it occurs costly signals will be selected against (Gintis et al., 2001; Lachmann et al., 2001).

That is, with deterrents the costs are paid by those who deviate from the ESS. This differs from the handicap principle in the following important sense: there, costs are paid as part of the ESS. This is the reason why the handicap principle should not be extended to scenarios in which the costs are paid socially rather than in production; the two are formally quite different. In one, handicaps, costs are incurred by honest signallers; while in the other, deterrents, costs are incurred by dishonest signallers. This is a fundamental difference that is not captured by present terminology. In fact, it has become standard to use the term handicap to refer to both scenarios. The suggestion here is that they be distinguished.

We thus have a three way classification of the basic functional outcomes by which communication between non-relatives may be stabilised (Scott-Phillips, 2008). Each of these, and particularly deterrents, could be subdivided using mechanistic criteria, but that matter is not discussed here:

- *indices*, in which signal form is tied to signal meaning;
- *handicaps*, in which signal cost is tied to signal form, and hence acts as a guarantee that is incurred by reliable/honest signallers;
- *deterrents*, in which costs are incurred by signallers who deviate from reliability/honesty.

These possibilities are mutually exclusive and are provisionally taken to be exhaustive — additional suggestions are not forthcoming. The present trichotomy covers scenarios in which unreliability/dishonesty is either precluded (indices), expensive (handicaps) or unwise (deterrents). The possibility of further alternatives is not discounted, but it is hard to see what form they might take. We now turn to the question of which one most likely applies to human communication, and natural language in particular.

3 Stable Communication in Humans

Three of the solutions discussed above can be discounted an explanation of why linguistic communication is evolutionarily stable. First, kin selection has been proposed as a partial explanation of the stability of human communication (Fitch, 2004), but there is, of course, an obvious flaw — that we freely speak to non-relatives. This is the reason why kin selection can only be a partial explanation of honesty and reliability; the suggestion is only that contemporary language evolved "primarily in a context of kin selection" (Fitch, 2004, p.275). Exactly what is entailed by this needs to be more fully developed before the idea can be properly evaluated. Second, linguistic form is famously unrelated to meaning (Saussure, 1959): *dog*, *chien* and *Hund* all refer to the same set of canine animals, despite no similarities in form. (Onomatopoeia is a rare exception.) Third, handicaps may be excluded because, as discussed above, the notion of a handicap should be restricted to those scenarios in which there are strategic costs associated with signal production, yet human utterances do not seem to carry such costs. Despite this "crippling problem" (Miller, 2000, p.348), researchers have still looked to the handicap principle as an explanation of stability in language.

On example is the suggestion that the sort of ritualised performance witnessed in many societies acts as a costly signal of one's commitment to the group, and hence performers are trusted as in-group members (Knight, 1998; Power, 1998). However there is nothing in this model to stop an individual paying the costs to enter the in-group and then once accepted behaving dishonestly or even unreliably. This is because the costs of the performance are not causally tied to the individual's subsequent utterances. A second example is the hypothesis that politeness phenomena act as a handicap (van Rooij, 2003), in that they reduce the speaker's social standing relative to the listener, place them in the listener's debt or otherwise incur socially relevant costs. For example, the utterance "I don't suppose there'd be any possibility of you..." can be read as an announcement that the speaker is prepared to incur some social cost in order to ensure that the desire which follows the ellipsis is satisfied. Let us accept, for the sake of argument, that this argument is correct. This does not make politeness a handicap, since the costs incurred are not paid *as part of* the signal. If politeness does place us in social debt then this would be an example of a self-imposed *deterrent* rather than a handicap: it imposes a social obligation on us to return the favour in some way, and we do not renege on this because the threat of social exclusion deters us from doing so. The difference between deterrents and handicaps is implicitly acknowledged by this paper, which discusses how the costs can be due to either by the signal (handicaps) or the receiver(s) (deterrents). However the term handicap is then used to refer to both scenarios — and as such offers a good example of how such usage has become standard.

By deduction, then, we are drawn towards deterrents as a solution to the two evolutionary problems of reliability and honesty. There is an intuitiveness to both ideas: unreliable communication, for example if one says "dog" to refer to feline pets, is deterred because it means that one will not be understood, and hence cannot achieve one's communicative goals; and dishonest communication will result in a loss of trust and the consequent social costs. In fact, deterrents are what we should logically expect to find in humans. In general, when indices are not available, and when the expected gains from dishonesty or unreliability outweigh the costs, then costly signals must be employed to ensure stability (Lachmann et al., 2001). Deterrents will be used only if verification is both possible and cheap — which is precisely what epistemic vigilance gives us. A similar finding is that signalling can be stable if unreliability is costly (Gintis et al., 2001), and it should also be noted that deterrents allow signals to take an arbitrary form (Lachmann et al., 2001). The fact that utterances are cheap yet arbitrary is too often taken to be paradoxical: "resistance to deception has *always* selected against conventional [arbitrary –TSP] signals — with the one puzzling exception of humans" (Knight, 1998, p.72, italics added). This is, as the examples discussed above show, simply not true. Instead, once we remove the requirement that costs be causally associated with signal form, as we do if we place the onus of payment on the *dis*honest individual, then the signal is free to take whatever form the signaller wishes. This paves the way for an explosion of symbol use.

What keeps humans from dishonesty and unreliability? There is an obvious candidate: *reputation*. For this to work it must be possible for individuals to modify their future behaviour in the light of other's behaviour. This is a task that the human memory performs with ease, often subconsciously (see Pentland, 2008) but we should nevertheless recognise it as a crucial prerequisite. Emotions like anger ensure that we do not repeatedly trust those that have cheated us (Ekman, 1992; Tooby & Cosmides, 1990), and the empirical literature contains many illustrations of our sensitivity to untrustworthy behaviour. For example, we are more likely to recall the identities of cheaters than cooperators (Chiappe et al., 2004; Mealey et al., 1996; Oda, 1997). We are well attuned to the detection of unfakeable physical cues of dishonest behaviour, for example a lack of eye contact and a large number of unfilled pauses in speech (Anolli & Ciceri, 1997; Scherer et al., 1985), and these appear to be cross-cultural (Bond Jr. et al., 1990). In fact such cues may even be seen not only when we are deceptive but also in our everyday appearance: when presented with a number of faces and asked to recall them later, experimental participants are more likely to recall the identities of individuals who later defected in a game of prisoner's dilemma, even when they do not have access to this information (Yamagishi et al., 2003).

We are also very sensitive to our own reputational status within the social group in general, and are keen to maintain our standing: cooperation can be maintained in various economic games once reputational effects are added, but not otherwise (Milinski et al., 2002; Piazza & Bering, 2008; Wedekind & Milinski, 2000). This is true even if we experience only subtle cues of a potential loss of reputation, such as stylised eyespots on a computer (Haley & Fessler, 2005). Such effects have also been found in more ecologically-valid conditions: an honesty box for tea, coffee and milk in a University common room received greater contributions when the small picture above it was a pair of human eyes rather than a flower (Bateson et al., 2007). This attentiveness to one's own reputation and to cues that it may be affected by current behaviour should not be a surprise, since a loss of reputation will mean exclusion from the local group, a heavy penalty for a social species like ourselves. Indeed, the emerging consensus from the burgeoning literature on the evolution of cooperation is that reputational effects are crucial to stability (Fehr, 2004); without such effects scenarios like the tragedy of the commons are far more likely to arise (Milinski et al., 2002). A similar story seems to hold in primate societies (Gouzoules & Gouzoules, 2002). Note also that this effect is likely to snowball once language in some form or another is off the ground, since individuals then become able to exchange information about the honesty and reliability of others (Enquist & Leimar, 1993). This may explain why so much of our conversational time is dedicated to gossip (Dunbar, 1997).

An important implication of the hypothesis that unreliability and dishonesty are deterred by the threat of poor reputations is that the second-order problem of how deterrents are implemented does not arise. Nobody is asked to bear the brunt of the costs of punishing others, because social exclusion is not itself costly to enforce. On the contrary, it is the most adaptive response to individuals with a reputation for unreliability or dishonesty.

4 Concluding Remarks

In this chapter I have sought to review the various ways in which communication can be evolutionarily stable, and ask which most likely applies to linguistic communication. Language is a more complex case than most if not all animal signals, since it sets two problems rather than one. The first is *reliability*: we must agree upon the meaning of signal. The second is *honesty*: why should signallers be honest if dishonesty pays? These two terms are often used synonymously, but the case of language makes it clear that they are separate problems. They correspond to two different layers of communication, analogous to the well-recognised distinction within pragmatics between communicative and informative intent. A third layer, to do with whether or not communication is used to achieve mutually beneficial goals, is also identified. This material cooperation is, of course, not necessary for stable communication: we can antagonise and argue with our interlocutors, but still maintain stability.

One way in which the evolutionary problems of communicative and informative cooperation can be solved is for there to be a causal relationship between meaning and form. This ensures that the signal cannot be faked, and is termed an *index*. Alternatively, the signals may be costly, and if there is a causal relationship between that cost and the signal's meaning then we have *handicaps*. Finally, there may be some costs associated with dishonesty or unreliability that outweigh the potential benefits. These costs act as *deterrents*. Note that these deterrents are often a consequence of the environmental make-up rather than pro-active punishment, since such enforcement would only replace the first-order problems of dishonesty and reliability with an analogous second-order problem, under the reasonable assumption that this enforcement is itself costly. In humans, deterrents seem the most likely solution to both problems — we are deterred from unreliable and dishonest communication because that would give us a bad reputation, with obvious evolutionary consequences. There is scope for empirical investigation of this proposal. One way in which this could be done would be to use the economic games that have been profitably used to study the effects of reputation in the evolution of cooperation in humans (e.g. Axelrod, 1995; Milinski et al., 2002), but with the independent variable as honesty in a communication game rather than cooperation in a prisoner's dilemma or some other cooperative game. Investigation of whether and how humans might differ from other primates in this regard would also be useful.

One matter that has not been discussed is the informational value of utterances. Conversation is sometimes thought of as an exchange of information, which is kept stable through reciprocity (e.g. Ulbaek, 1998). This would imply that we keep track of who we have given information to, punish those who do not provide information in return, and compete to listen to others. These predictions do not seem to be correct; on the contrary, we compete to speak rather than to listen (Dessalles, 1998, 2007). In general, speaking appears to be a selfish rather than an altruistic act (Scott-Phillips, 2007). One reason for this is to gain a better reputation: the scientist who presents good work at a conference will go up in his colleagues' esteem, for example. As such, then, this story also involves

reputation, in this case the attainment of good reputation. This is the other side of the bad reputation that will follow if we speak unreliably or dishonestly.

It hardly bears stating that the honesty and reliability of utterances are central to pragmatics. Without reliability communication cannot take place at all, and honesty is so crucial that Grice saw fit to make it one of his four maxims. He also desired a naturalistic basis for his ideas: "I would like to think of the standard type of conversational practice not merely as something that all or most do *in fact* follow but as something that it is *reasonable* for us to follow, that we *should not* abandon" (1975, p.48, italics in original). This chapter has sought to explore how animal signalling theory can be applied to language so as to provide an important part of that foundation: evolutionary stability.

References

Anolli, L., Ciceri, R.: The voice of deception: Vocal strategies of naive and able liars. Journal of Nonverbal Behaviour 21, 259–284 (1997)

Axelrod, R.: The Evolution of Cooperation. Basic Books, New York (1995)

Axelrod, R., Hamilton, W.D.: The evolution of cooperation. Science 211, 1390–1396 (1981)

Bateson, M., Nettle, D., Roberts, G.: Cues of being watched enhance cooperation in a real-world setting. Biology Letters, 412–414 (2007)

Bond Jr., C.F., Omar, A., Mahmoud, A., Bonser, R.N.: Lie detection across cultures. Journal of Nonverbal Behavior 14(3), 189–204 (1990)

Boyd, R., Richerson, P.J.: Punishment allows the evolution of cooperation (or anything else) in sizable groups. Ethology and Sociobiology 13, 171–195 (1992)

Cheney, D.L., Seyfarth, R.M.: How Monkeys see the World. University of Chicago Press, Chicago (1990)

Chiappe, D., Brown, D., Dow, B., Koontz, J., Rodreiguez, M., McCulloch, K.: Cheaters are looked at longer and remembered better than cooperators in social exchange situations. Evolutionary Psychology 2, 108–120 (2004)

Clutton-Brock, T.H., Parker, G.A.: Punishment in animal societies. Nature 373, 209–216 (1995)

Dessalles, J.-L.: Altruism, status and the origin of relevance. In: Hurford, J.R., Studdert-Kennedy, M., Knight, C. (eds.) Approaches to the Evolution of Language, pp. 130–147. Cambridge University Press, Cambridge (1998)

Dessalles, J.-L.: Why We Talk: The Evolutionary Origins of Language. Oxford University Press, Oxford (2007)

Dunbar, R.I.M.: Grooming, Gossip, and the Evolution of Language. Faber, London (1997)

Ekman, P.: An argument for basic emotions. Cognition and Emotion 6, 169–200 (1992)

Enquist, M., Leimar, O.: The evolution of cooperation in mobile organisms. Animal Behaviour 45(4), 747–757 (1993)

Fehr, E.: Don't lose your reputation. Nature 432, 449–450 (2004)

Fehr, E., Fischbacher, U.: The nature of human altruism. Nature 425, 785–791 (2003)

Fitch, W.T.: Evolving honest communication systems: Kin selection and mother tongues. In: Oller, D.K., Griebel, U. (eds.) The Evolution of Communication Systems: A Comparative Approach, pp. 275–296. MIT Press, Cambridge (2004)

Fitch, W.T., Reby, D.: The descended larynx is not uniquely human. In: Proceedings of the Royal Society of London, series B, vol. 268, pp. 1669–1675 (2001)

Frank, S.A.: Foundations of Social Evolution. Princeton University Press, Princeton (1998)
Gintis, H., Alden Smith, E., Bowles, S.: Costly signaling and cooperation. Journal of Theoretical Biology 213, 103–119 (2001)
Gouzoules, H., Gouzoules, S.: Primate communication: By nature honest, or by experience wise? International Journal of Primatology 23(4), 821–847 (2002)
Grafen, A.: Biological signals as handicaps. Journal of Theoretical Biology 144, 517–546 (1990)
Grafen, A.: Optimisation of inclusive fitness. Journal of Evolutionary Biology 238, 541–563 (2006)
Grice, H.P.: Logic and conversation. In: Cole, P., Morgan, J. (eds.) Syntax and Semantics III: Speech Acts, pp. 41–58. Academic Press, New York (1975)
Grim, P., Wardach, S., Beltrani, V.: Location, location, location: The importance of spatialization in modeling cooperation and communication. Interaction Studies 7(1), 43–78 (2006)
Guinet, C.: Predation behaviour of Killer Whales (Orcinus orca) around Crozet islands. Canadian Journal of Zoology 70, 1656–1667 (1992)
Haley, K.J., Fessler, D.M.T.: Nobody's watching? Subtle cues affect generousity in an anonymous economic game. Evolution and Human Behavior 26(3), 245–256 (2005)
Hamilton, W.D.: The genetical evolution of social behaviour. Journal of Theoretical Biology 7, 1–52 (1964)
Hardin, G.: The tragedy of the commons. Science 162, 1243–1248 (1968)
Hurd, P.L.: Communication in discrete action-response games. Journal of Theoretical Biology 174, 217–222 (1995)
Hurford, J.R.: Origins of Meaning. Oxford University Press, Oxford (2007)
Knight, C.: Ritual/speech coevolution: a solution to the problem of deception. In: Hurford, J.R., Studdert-Kennedy, M., Knight, C. (eds.) Approaches to the Evolution of Language, pp. 68–91. Cambridge University Press, Cambridge (1998)
Lachmann, M., Szmad, S., Bergstrom, C.T.: Cost and conflict in animal signals and human language. Proceedings of the National Academy of Sciences 98(23), 13189–13194 (2001)
Maynard Smith, J.: Fertility, mating behaviour and sexual selection in Drosophila subobscura. Journal of Genetics 54, 261–279 (1956)
Maynard Smith, J.: Group selection and kin selection. Nature 201, 1145–1147 (1964)
Maynard Smith, J.: Sexual selection and the handicap principle. Journal of Theoretical Biology 57, 239–242 (1976)
Maynard Smith, J.: Evolution and the Theory of Games. Cambridge University Press, Cambridge (1982)
Maynard Smith, J., Harper, D.G.C.: Animal signals: Models and terminology. Journal of Theoretical Biology 177, 305–311 (1995)
Maynard Smith, J., Harper, D.G.C.: Animal Signals. Oxford University Press, Oxford (2003)
Mealey, L., Daood, C., Krage, M.: Enhanced memory for faces of cheaters. Ethology and Sociobiology 17, 119–128 (1996)
Milinski, M., Semmann, D., Krambeck, H.-J.: Reputation helps solve the 'tragedy of the commons'. Nature 415, 424–426 (2002)
Miller, G.F.: The Mating Mind. BCA, London (2000)
Oda, R.: Biased face recognition in the prisoner's dilemma games. Evolution and Human Behavior 18, 309–315 (1997)
Parker, G.A.: Assessment strategy and the evolution of animal conflicts. Journal of Theoretical Biology 47, 223–243 (1974)

Pentland, A.: Honest Signals: How they Shape our World. MIT Press, Cambridge (2008)
Piazza, J., Bering, J.M.: Concerns about reputation via gossip promote generous allocations in an economic game. Evolution and Human Behavior 29, 172–178 (2008)
Power, C.: Old wives' tales: The gossip hypothesis and the reliability of cheap signals. In: Hurford, J.R., Studdert-Kennedy, M., Knight, C. (eds.) Approaches to the Evolution of Language, pp. 111–129. Cambridge University Press, Cambridge (1998)
Reby, D., McComb, K.: Anatomical constraints generate honesty: Acoustic cues to age and weight in the roars of Red Deer stags. Animal Behaviour 65, 317–329 (2003)
Rohwer, S.: The social significance of avian winter plumage variability. Evolution 29, 593–610 (1975)
Rohwer, S., Rohwer, F.C.: Status signalling in Harris' Sparrows: experimental deceptions achieved. Animal Behaviour 26, 1012–1022 (1978)
de Saussure, F.: Course in General Linguistics. McGraw-Hill, New York (1959)
Scherer, K.R., Feldstein, S., Bond, R.N., Rosenthal, R.: Vocal cues to deception: A comparative channel approach. Journal of Psycholinguistic Research 14, 409–425 (1985)
Scott-Phillips, T.C.: The social evolution of language, and the language of social evolution. Evolutionary Psychology 5(4), 740–753 (2007)
Scott-Phillips, T.C.: On the correct application of animal signalling theory to human communication. In: Smith, A.D.M., Smith, K., Ferreri Cancho, R. (eds.) The Evolution of Language: Proceedings of the 7th International Conference on the Evolution of Language, pp. 275–282. World Scientific, Singapore (2008)
Searcy, W.A., Nowicki, S.: The Evolution of Animal Communication. Princeton University Press, Princeton (2007)
Silk, J.B., Kaldor, E., Boyd, R.: Cheap talk when interests conflict. Animal Behaviour 59, 423–432 (2000)
Skyrms, B.: The Evolution of the Social Contract. Cambridge University Press, Cambridge (1996)
Spence, M.: Job market signalling. Quarterly Journal of Economics 87, 355–374 (1973)
Sperber, D., Wilson, D.: Relevance: Communication and Cognition, 2nd edn. Blackwell, Oxford (1995)
Sperber, D., Wilson, D.: Epistemic Vigilance. Paper presented at the Workshop on pragmatics and social cognition, UCL, London (2008)
Szmad, S., Szathmry, E.: Selective scenarios for the emergence of natural language. Trends in Ecology and Evolution 21(10), 555–561 (2006)
Taylor, P.W., Hasson, O., Clark, D.L.: Body postures and patterns as amplifiers of physical condition. Proceedings of the Royal Society of London, series B 267, 917–922 (2000)
Tooby, J., Cosmides, L.: The past explains the present: Emotional adaptations and the structure of ancestral environments. Ethology and Sociobiology 11, 375–424 (1990)
Tricks of the traders, 3rd edn. Guardian Magazine (2008)
Trivers, R.L.: The evolution of reciprocal altruism. Quarterly Review of Biology 46, 35–57 (1971)
Ulbaek, I.: The origin of language and cognition. In: Hurford, J.R., Studdert-Kennedy, M., Knight, C. (eds.) Approaches to the Evolution of Language, pp. 30–43. Cambridge University Press, Cambridge (1998)
van Rooij, R.: Being Polite is a Handicap: Towards a Game Theoretical Analysis of Polite Linguistic Behaviour. Paper presented at the 9th conference on the theoretical aspects of rationality and knowledge (2003)

Veblen, T.: The Theory of the Leisure Class. MacMillian, London (1899)
Versluis, M., Schmitz, B., von der Heydt, A., Lohse, D.: How Snapping Shrimps snap: Through cavitating bubbles. Science 289, 2114–2117 (2000)
Wedekind, C., Milinski, M.: Cooperation through image scoring in humans. Science 288, 850–852 (2000)
West, S.A., Gardner, A., Shuker, D.M., Reynolds, T., Burton-Chellow, M., Sykes, E.M., et al.: Coooperation and the scale and competition in humans. Current Biology 16, 1103–1106 (2006)
West, S.A., Griffin, A.S., Gardner, A.: Social semantics: Altruism, cooperation, mutualism and strong reciprocity. Journal of Evolutionary Biology 20, 415–432 (2007)
Whitfield, D.P.: Plumage variability, status signalling and individual recognition in avian flocks. Trends in Ecology and Evolution 2, 13–18 (1987)
Yamagishi, T., Tanida, S., Mashima, R., Shimoma, S., Kanazawa, S.: You can judge a book by its cover: Evidence that cheaters may look different from co-operators. Evolution and Human Behavior 24, 290–301 (2003)
Zahavi, A.: Mate selection: A selection for a handicap. Journal of Theoretical Biology 53, 205–214 (1975)