Anim. Behav., 1993, **46**, 759–764

Dishonesty and the handicap principle

RUFUS A. JOHNSTONE & ALAN GRAFEN
Department of Plant Sciences, South Parks Road, Oxford OX1 3RA, U.K.

(Received 9 July 1992; initial acceptance 28 August 1992; final acceptance 5 January 1993; MS. number: 4147)

Abstract. The handicap principle states that stable biological signals must be honest. Here, it is argued that they need only be honest 'on average'. If signallers employ a number of different signalling strategies at equilibrium, then the handicap principle cannot entirely rule out dishonesty. A formal demonstration of this possibility, using evolutionarily stable strategy techniques, is provided and the conditions that might lead to the evolution of multiple signalling strategies are discussed. It is concluded that the ideal of perfect honesty will almost never be met.

Zahavi (1975, 1987) has argued that stable biological signals must be honest. In recent years, moreover, his views have received increasing support. The development of mathematical models of biological signalling has provided formal verification of his 'handicap principle' (Enquist 1985; Grafen 1990; Godfray 1991; Maynard Smith 1991; Johnstone & Grafen 1992a, b). At the same time, evidence has accumulated that animal ornaments and displays sometimes convey honest information (see, for instance, Eckert & Weatherhead 1987; Møller 1990, 1992; Knapp & Kovach 1991; Wedekind 1992). In this paper, however, we show that Zahavian signals need only be honest 'on average'. Cheating can be part of a stable signalling system, provided that its incidence is low enough.

Existing models of the handicap principle appear to rule out dishonesty because they assume that all signallers employ the same signalling strategy. Under these circumstances, each level of signal will invariably be associated with signallers of a particular quality. Consequently, an interpretation strategy that sometimes yields an appropriate response to a particular level of signal must always yield an appropriate response to that level of signal. Honesty 'on average' implies honesty on each and every occasion of signalling.

In the real world, as opposed to the abstract world of the models, signallers are bound to differ in many respects. The signalling strategies they adopt are also, therefore, bound to differ. Variation in the degree of relatedness of signallers to the receiver, in the cost of signalling and in the benefits to be gained, will all influence signalling strategy. Such variation introduces a kind of 'noise' into the signalling system, similar to the effects of perceptual error (Johnstone & Grafen 1992a), in that receivers cannot determine which strategy a particular signaller employs. Each level of signal will be associated with a range of signalling types, some of whom are more prone to exaggeration than others.

If, therefore, there are several types of signaller, an interpretation strategy that sometimes yields an appropriate response to a particular level of signal cannot always yield an appropriate response to that level of signal. Even when, at equilibrium, receivers employ the best possible interpretation strategy, they will consistently overestimate the quality of some signalling types and underestimate the quality of others. Individuals of the former signalling types can be thought of as 'cheats', because they fool receivers into assigning them a dishonestly high quality.

In the next section, we demonstrate formally that cheating can be part of an evolutionarily stable signalling system. We present a version of Maynard Smith's (1991) 'Philip Sidney' game, a simple evolutionarily stable strategy (ESS) model of biological communication, in which there are two classes of signaller. These classes differ in the cost they pay for signalling, and in their degree of relatedness to the receiver. We show that an ESS can exist in which members of one class signal honestly, while members of the other class signal dishonestly. In the following section, we consider the generality of the model, and the likelihood of multiple signalling strategies evolving in nature. We argue that the conditions necessary for stable cheating will commonly be satisfied, so that signals will rarely, if ever,

0003–3472/93/100759 + 06 $08.00/0

be perfectly honest. Finally, we compare our view of cheating with other interpretations.

DECEPTION IN THE PHILIP SIDNEY GAME

The Philip Sidney game features two players, a potential donor and a beneficiary. Both players are at risk of dying. The donor controls an indivisible resource, such as a water bottle, which he may keep for himself or give to the beneficiary. If the donor keeps the water, he ensures his own survival; if he gives it to the beneficiary, then he ensures the beneficiary's survival.

As long as the behaviour of the two players is determined by natural selection maximizing individual fitness, the donor will keep the water for himself. Let us suppose, however, that the two players are related by a coefficient of relatedness r. Now, selection will maximize inclusive fitness, and it may pay the donor to hand over the water. Whether or not it does pay him to do so depends on the survival chances of the two players, and on the value of r.

We shall assume that the donor has a fixed survival chance without water, denoted S_D. The beneficiary, however, can be in one of two states, 'thirsty' or 'not thirsty', with probabilities of P and $(1-P)$, respectively. A beneficiary who is thirsty has a survival chance of 0 without water, while a beneficiary who is not thirsty has a survival chance of S_B without water. Under these circumstances, it may pay the donor to hand over the water to a beneficiary who is thirsty, but to keep the water when faced with a beneficiary who is not thirsty.

Suppose that the donor cannot directly perceive the state of the beneficiary. Now, it may pay the beneficiary to give a signal of his state. The donor could then make his decision as to whether he keeps or gives up the water on the basis of this signal. Maynard Smith (1991) demonstrated that honest signalling of this sort was possible. He also showed, however, that when there was a conflict of interest between signaller and receiver, honest signals had to be costly. That is, if we assume that the fitness of a beneficiary who signals is reduced by a factor of $(1-t)$, where t represents the cost of the signal, then honest signalling could only be stable for positive values of t.

Maynard Smith's signalling equilibria were perfectly honest. This reflects his assumption that beneficiaries differ only in their degree of thirst, and therefore employ the same signalling strategy. If, by contrast, there are multiple signalling strategies at equilibrium, then a partially honest signalling system might be possible. We shall show that if beneficiaries vary in the cost they pay for signalling, or in their degree of relatedness to the donor, then there can be multiple signalling strategies at equilibrium, and the signalling system need not be perfectly honest.

For the sake of simplicity, we assume that there are only two distinct classes of beneficiary, with frequencies of q and $(1-q)$. Individuals of the first class pay a cost t_1 for signalling, and are related to the donor by a coefficient of relatedness r_1. Individuals of the second class pay a cost t_2 for signalling, and are related to the donor by a coefficient of relatedness r_2. We wish to show that a partially honest signalling equilibrium is possible, or in other words that the following set of strategies can be stable: (1) class 1 beneficiaries signal only when thirsty; (2) class 2 beneficiaries always signal; (3) donors hand over water only in response to a signal. At an equilibrium of this type, class 1 beneficiaries signal honestly while class 2 beneficiaries do not.

Signalling Strategies in the New Model

If, at equilibrium, beneficiaries of the first class signal only when thirsty, then this conditional strategy must yield a higher inclusive fitness than the two possible alternatives, 'never signal' and 'always signal'. Using the payoff functions given in Table I, we can express these conditions as follows

$$1-t_1+r_1 S_D > r_1 \quad (1a)$$

$$S_B+r_1 > 1-t_1+r_1 S_D \quad (1b)$$

If (1a) is not satisfied, then the strategy 'never signal' yields higher fitness, and if (1b) is not satisfied, then the strategy 'always signal' yields higher fitness.

Similarly, if beneficiaries of the second class always signal, then this strategy must yield higher inclusive fitness than either of the alternatives, 'never signal' and 'signal only when thirsty'. Using the payoff functions given in Table I, we can see that these requirements reduce to a single condition

$$1-t_2+r_2 S_D > S_B+r_2 \quad (2)$$

If (2) is satisfied, then neither alternative strategy yields higher fitness.

Table I. Payoffs to beneficiaries of class i adopting various signalling strategies, when donors give only in response to a signal

Strategy	Payoff
Always signal	$W=(1-P)(1-t_i+r_iS_D)+P(1-t_i+r_iS_D)$
Signal when thirsty	$W=(1-P)(S_B+r_i)+P(1-t_i+r_iS_D)$
Never signal	$W=(1-P)(S_B+r_i)+Pr_i$

See text for further explanation.

Table II. Payoffs to donors adopting various response strategies, when signallers of class 1 signal only if thirsty and signallers of class 2 always signal

Strategy	Payoff
Always give	$W=q[(1-P)(S_D+r_1)+P[S_D+r_1(1-t_1)]]+(1-q)[S_D+r_2(1-t_2)]$
Give in response to signal	$W=q[(1-P)(1+r_1S_B)+P[S_D+r_1(1-t_1)]]+(1-q)[S_D+r_2(1-t_2)]$
Never give	$W=q[(1-P)(1+r_1S_B)+P]+(1-q)[(1-P)(1+r_2(1-t_2)S_B)+P]$

See text for further explanation.

Now, if there is a conflict of interest between donor and beneficiary (see Maynard Smith 1991), then (1) and (2) can be jointly satisfied only if $t_1>0$, and if either $t_2<t_1$ or $r_2<r_1$, or both. In other words, a partially honest signalling system can be stable only if signalling is costly for some individuals, so that they are constrained to honesty. At the same time, other individuals must pay lower costs, or be less closely related to the recipient, or both, so that they are freed from this constraint.

Response Strategies in the New Model

If, at equilibrium, donors hand over the water only in response to a signal, then this conditional strategy must yield a higher inclusive fitness than the two possible alternatives, 'always give' and 'never give'. Using the payoff functions given in Table II, we can express these conditions as follows

$$1+r_1S_B>S_D+r_1 \qquad (3a)$$

$$qA+(1-q)B>0 \qquad (3b)$$

where $A=P[S_D-1+r_1(1-t_1)]$ and $B=S_D-1+r_2(1-t_2)[(S_B(1-P)-1]$. If (3a) is not satisfied, then the strategy 'always give' yields higher fitness, and if (3b) is not satisfied then the strategy 'never give' yields higher fitness.

Now, (3) can be jointly satisfied at the same time as (1) and (2). If, for example, $r_1=0{\cdot}5$ and $r_2=0{\cdot}2$, then the values $S_D=S_B=0{\cdot}8$, $t_1=0{\cdot}4$, $t_2=0{\cdot}3$, $P=0{\cdot}6$ and $q=0{\cdot}9$ satisfy the conditions. We may therefore conclude that a partially honest signalling system can be stable, although we can see from (3b) that this is possible only if the proportion of honest signallers exceeds a critical level. The minimum frequency of honest signallers required for stability depends on the cost to the donor of responding to a 'cheat'. If the cost is high, then there must be a high proportion of honest signallers for the system to remain stable. If, on the other hand, the cost to a donor of being 'cheated' is low, then the system can tolerate a low level of honesty.

EVOLUTION OF MULTIPLE SIGNALLING STRATEGIES

The handicap principle claims that stable signalling systems must be honest. Our arguments show that they need not be perfectly honest. Instead, stable dishonesty is possible, provided that signallers employ a number of different signalling strategies. In this section, we consider when and why multiple signalling strategies should evolve.

Previous models of the handicap principle have featured a single signalling strategy because all signallers of a given quality were assumed to be identical. Under these conditions, the best possible

strategy for one signaller will be the best possible strategy for all signallers. By contrast, in our model, signallers differed in their degree of relatedness to the receiver, and in the cost they paid for signalling. It was these differences that gave rise to multiple signalling strategies at equilibrium. We must therefore determine whether similar differences will occur in real signalling systems, and whether they will have similar consequences.

Differences between signallers in their degree of relatedness to the recipient are likely to be widespread in nature. However, for these differences to result in dishonesty, signallers must be able to distinguish levels of kinship more accurately than receivers can. In the model, we assumed that beneficiaries could employ different signalling strategies according to their degree of relatedness to the donor, while donors could employ only a single response strategy. Implicitly, therefore, we assumed that signallers could distinguish the two levels of relatedness r_1 and r_2, while donors could not. If, on the other hand, receivers can determine kinship as well as signallers, then at equilibrium, there will be a different pair of signal and response strategies for each level of relatedness. Signalling costs will vary according to the degree of relatedness between signaller and receiver, but there will be no opportunity for dishonesty (see Johnstone & Grafen 1992b).

Differences between signallers in the costs they pay for signalling and in the benefits they stand to gain are likely to be of greater importance. Existing models of costly signalling have generally assumed that signallers vary either in their ability to pay signalling costs, or in the amount they stand to gain. In the former case, exemplified by displays of male quality during mating, honest signalling is stable because high-quality individuals pay lower costs for a given level of advertising (see, for instance, Grafen 1990; Johnstone & Grafen 1992a). In the latter case, exemplified by chicks begging for food, signallers differ in their degree of need, and honest signalling is stable because needy individuals stand to gain more from a given level of advertising (see, for instance, Godfray 1991; Maynard Smith 1991; Johnstone & Grafen 1992b). But while receivers may often be interested in ability to pay or ability to gain alone, real signallers are bound to differ in both respects.

Where signallers vary both in their ability to pay the cost of signalling and in the amount they stand to gain, their level of signalling will reflect both factors. This introduces an element of receiver uncertainty. A receiver interested in need must cope with multiple signalling strategies reflecting differences in ability to pay. A given level of signal may indicate a needy signaller who cannot easily bear the costs of signalling, or a less needy individual who can easily bear them. Conversely, a receiver interested in ability to pay must cope with multiple signalling strategies reflecting differences in need. A given level of signal may indicate a high-quality signaller with less to gain, or a low-quality signaller with more to gain. Our model described above considered the former situation. Donors were interested in whether or not the beneficiary was thirsty, or in other words in his level of need. Stable dishonesty was possible because beneficiaries differed in the cost they paid for signalling, or in other words in their 'quality'.

Even ignoring differences in need, moreover, signals of quality can be partially dishonest at equilibrium if some individuals find advertising cheaper for a given level of quality. Under these circumstances, the 'real' cost of the signal may differ from the 'apparent' cost that the receiver assumes is being paid. Thus, the cost of an energetic vocal display such as birdsong depends on an enormous number of different factors, the time of day, the energy reserves of the singer, the chance of attracting predators (which will differ from one time or place to another), the signaller's need to perform other actions, etc. The quality in which the receiver is interested is only one among these many factors (see Richner 1992). It may be the major determinant of a signaller's ability to bear the cost of signalling, but it cannot be the only determinant. Different signallers, or even a single signaller at different times, will therefore employ different signalling strategies. Deception is the inevitable result.

In all the above cases, the extent to which cheats benefit depends critically on their frequency. If they are rare, then the chances are that any particular signal has come from an honest individual, so receivers will attribute a high quality to the signaller. If cheats are common, then the chances are that any particular signal has come from a dishonest individual, so receivers will attribute a lower quality to the signaller. Given the diversity of factors that influence signalling strategy, signalling systems are almost bound to feature a range of common strategies, all slightly different, and some of these will benefit slightly from the effects of receiver uncertainty. In general parlance, however, the term 'cheat' is reserved for a distinct type

of signaller that is uncommon enough to gain a substantial advantage of this sort.

DIFFERENT KINDS OF CHEATING

The term 'cheat' has been used in a number of different ways. In one common usage, for example, it refers to any strategy that might perturb a proposed signalling equilibrium. Thus, Maynard Smith & Harper (1988) analysed the fate of a 'cheat' with a dishonestly large badge that might disturb the signalling equilibrium in their 'Badges of Dominance' model, and other potential equilibrium-breakers have since been considered (Owens & Hartley 1991). Clearly, by this definition, cheating cannot be part of a stable signalling system. It is not, however, a definition that allows the term to be applied to real signalling systems.

We have used the word 'cheat' to refer to any signaller that is consistently misinterpreted to its own advantage. This is the sense in which the word is most commonly used in empirical studies. Thus, the stomatopod crustaceans described by Adams & Caldwell (1990), which continue to perform their normal threat displays in the period immediately after moulting, are deceptive because receivers treat their signals as indicating aggression despite the fact that newly moulted individuals cannot fight effectively.

Cheating, in the sense described above, has been the subject of much discussion, and our account may seem rather different from previous explanations. All theories of cheating, however, ours included, are based on receiver constraints. Without some restriction on the strategies that receivers can adopt, dishonesty would be impossible, because receivers would not allow themselves to be fooled. Natural selection would eliminate those response strategies that led to the misinterpretation of a signal. Different explanations of cheating merely emphasize different kinds of constraint.

Prior to the acceptance of the handicap principle, it seemed that constraints on receiver evolution were such that dishonesty might be widespread. Dawkins & Krebs (1978; Krebs & Dawkins 1984), for instance, emphasized the possibility of time lags in receiver evolution. Signaller and receiver, or 'manipulator' and 'mind-reader', are engaged in a coevolutionary arms race. If receivers are constrained in their rate of evolution, they may fall behind in the race, allowing signallers to exploit their outdated response strategies dishonestly. From this perspective it seems that dishonesty should be just as comomn as honesty, there being no obvious reason why receivers should win the arms race more often than signallers. Indeed, receivers may be constrained more frequently by the need to satisfy conflicting selective pressures. In mate choice, for example, female preferences may be partly determined by neural mechanisms that have evolved for other purposes (Ryan 1990; Ryan & Rand 1990; Ryan & Keddy-Hector 1992).

Once the handicap principle was accepted, however, it began to seem that honesty must be widespread. Zahavi (1975, 1987) pointed out that receivers could ensure honesty by paying attention to costly signals that only high quality signallers could afford to produce. This would cut short the coevolutionary arms race between mind-reader and manipulator, and lead to a stable, honest signalling system. Unless receivers were severely constrained by conflicting selective pressures in the response strategies they could adopt, it seemed that deceit could be only a temporary phenomenon.

We have shown, however, that dishonesty can be part of a stable signalling system. Our arguments are once again based on receiver constraints, but constraints of a different sort. The handicap principle may ensure a degree of honesty, because only high quality signallers can afford to produce costly signals, but unfortunately the link between signal and quality is bound to be less than perfect. Signallers will differ in a variety of ways other than in quality, differences that influence their signalling strategy, and this variation will lead to receiver uncertainty. If receivers were able to assess all the relevant factors when interpreting a signal, they would be able to determine the signalling strategy in use, and dishonesty would be impossible (see the discussion of relatedness above). It is because receivers are constrained in the degree of information they can acquire that deceit becomes possible. The recipient of a signal, unaware of exactly which signalling strategy is being employed, can only make a best guess as to its meaning.

Over the course of evolutionary time, signalling systems may evolve to become less prone to cheating. Where receivers respond to a number of different but concurrent displays (Johnstone & Grafen 1992a), for instance, as with the tail plumage, 'jump display' and display court of the lekking widowbird, *Euplectes jacksoni* (Andersson 1991,

1992), natural selection may favour a shift in attention towards the display least susceptible to cheating. Our arguments, however, show that even the most honest signal cannot be perfectly honest. Moreover, it is not clear that the less cheatable of two signalling systems must inevitably be more stable. Once receivers have evolved a favourable response to some display, it may be difficult or impossible to switch to a different signal, even if it would result, in the long run, in more honest communication. The question of whether or not signalling systems will evolve to a global 'optimum' remains, at the moment, unanswered.

To conclude, living organisms are bombarded by a host of signals, from members of their own and of other species. They are engaged in a continuing evolutionary struggle to extract meaning from these signals, to avoid being manipulated and deceived. The handicap principle tells us that there is at least a partial solution to their difficulties. By concentrating on costly signals, they can ensure that they are but rarely fooled. Yet, the apparent cost of a signal is not a perfect guide to the true quality of the signaller. Constrained in the extent of their information, receivers cannot hope to interpret all signals appropriately. Even, therefore, in a stable signalling system, there will be cheats.

REFERENCES

Adams, E. S. & Caldwell, R. L. 1990. Deceptive communication in asymmetric fights of the stomatopod crustacean *Gonodactylus bredini*. *Anim. Behav.*, **39**, 706–716.

Andersson, S. 1991. Bowers on the savanna: display courts and male choice in a lekking widowbird. *Behav. Ecol.*, **2**, 210–218.

Andersson, S. 1992. Female preference for long tails in lekking Jackson's widowbirds: experimental evidence. *Anim. Behav.*, **43**, 379–388.

Dawkins, R. & Krebs, J. R. 1978. Animal signals: information or manipulation. In: *Behavioural Ecology: An Evolutionary Approach* (Ed. by J. R. Krebs & N. B. Davies), pp. 282–309. Oxford: Blackwell Scientific Publications.

Eckert, C. J. & Weatherhead, P. J. 1987. Ideal dominance distributions: a test using red-winged blackbirds *Agelaius phoeniceus*. *Behav. Ecol. Sociobiol.*, **20**, 143–152.

Enquist, M. 1985. Communication during aggressive interactions with particular reference to variation in choice of behaviour. *Anim. Behav.*, **33**, 1152–1161.

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

Grafen A. 1990. Biological signals as handicaps. *J. theor. Biol.*, **144**, 517–546.

Johnstone, R. A. & Grafen, A. 1992a. Error-prone signalling. *Proc. R. Soc. Lond. Ser. B.*, **248**, 229–233.

Johnstone, R. A. & Grafen, A. 1992b. The continuous Sir Philip Sidney Game: a simple model of biological signalling. *J. theor. Biol.*, **156**, 215–234.

Knapp, R. A. & Kovach, J. T. 1991. Courtship as an honest indicator of male parental quality in the bicolor damselfish, *Stegastes partitus*. *Behav. Ecol.*, **2**, 295–300.

Krebs, J. R. & Dawkins, R. 1984. Animal signals: mind-reading and manipulation. In: *Behavioural Ecology: An Evolutionary Approach* 2nd edn (Ed. by J. R. Krebs & N. B. Davies), pp. 380–402. Oxford: Blackwell Scientific Publications.

Maynard Smith, J. 1991. Honest signalling: the Philip Sidney game. *Anim. Behav.*, **42**, 1034–1035.

Maynard Smith, J. & Harper, D. G. C. 1988. The evolution of aggression: can selection generate variability? *Phil. Trans. R. Soc. Lond. Ser. B.*, **319**, 557–570.

Møller, A. P. 1990. Fluctuating asymmetry in male sexual ornaments may reliably reveal male quality. *Anim. Behav.*, **40**, 1185–1187.

Møller, A. P. 1992. Patterns of fluctuating asymmetry in weapons: evidence for reliable signalling of male quality in beetle horns and bird spurs. *Proc. R. Soc. Lond. Ser. B.*, **248**, 199–206.

Owens, I. P. F. & Hartley, I. R. 1991. 'Trojan Sparrows': evolutionary consequences of dishonest invasion for the badges-of-status model. *Am. Nat.*, **138**, 1187–1205.

Richner, H. 1992. Assessment of expected performance and Zahavi's notion of signal. *Anim. Behav.*, **45**, 399–401.

Ryan, M. J. 1990. Sexual selection, sensory systems, and sensory exploitation. *Oxford Surv. Evol. Biol.*, **7**, 156–195.

Ryan, M. J. & Keddy-Hector, A. 1992. Directional patterns of female mate choice and the role of sensory biases. *Am. Nat.*, **139**, S4–S35.

Ryan, M. J. & Rand, A. S. 1990. The sensory basis of sexual selection for complex calls in the tungara frog, *Physalaemus pustulosus* (sexual selection for sensory exploitation). *Evolution*, **44**, 305–314.

Wedekind, C. 1992. Detailed information about parasites revealed by sexual ornamentation. *Proc. R. Soc. Lond. Ser. B.*, **247**, 169–174.

Zahavi, A. 1975. Mate selection: a selection for a handicap. *J. theor. Biol.*, **53**, 205–214.

Zahavi, A. 1987. The theory of signal selection and some of its implications. In: *International Symposium of Biological Evolution* (Ed. by V. P. Delfino), pp. 305–327. Bari: Adriatica Editrice.