J. theor. Biol. (1995) **175,** 405–421

The Cost of Threat Displays and the Stability of Deceptive Communication

ELDRIDGE S. ADAMS† AND MICHAEL MESTERTON-GIBBONS‡

†*Department of Biology*, *University of Rochester*, *Rochester*, *New York* 14627 *and*
‡*Department of Mathematics*, *Florida State University*, *Tallahassee*, *Florida* 32306, *U.S.A.*

(*Received on* 1 *February* 1993, *Accepted in revised form on* 20 *March* 1995)

Several analyses of intraspecific animal communication have suggested that threat displays must convey reliable information about the abilities of the signaller in order to be evolutionarily stable. In this paper, a game-theoretic model shows that bluffing by animals of low fighting ability can persist as a profitable tactic in a stable communication system. It is assumed that use of the threat display depends upon variation in fighting ability that is not visible to the opponent and that there is a fitness cost, or "handicap", paid by animals that threaten and subsequently lose. Analysis of the model shows that a handicap is necessary for stable communication and that the effectiveness of the threat increases with the magnitude of the handicap. However, the handicap does not ensure fully reliable communication and bluffing always forms part of the evolutionarily stable strategy (ESS). At the ESS, the very strongest and the very weakest members of the population threaten, while animals of intermediate strength do not. This is possible because, although weaker animals are liable to greater handicaps when they signal, they also gain greater benefits than strong animals using the same display. If all animals that threaten pay the handicap regardless of the outcome of the fight, then there is no ESS. These results provide a possible explanation for bluffing by the stomatopod crustacean, *Gonodactylus bredini*, a species in which animals weakened by molting successfully repulse stronger opponents by use of threat displays.

1. Introduction

Considerable controversy remains concerning the role of deception in intraspecific animal threat communication. Some authors have argued that threat displays must be "honest" in order to be evolutionarily stable; that is, they must convey reliable information about the abilities or motivations of the signaller (Zahavi, 1977, 1982, 1987; Markl, 1985; Grafen, 1990). According to the "handicap principle", reliability is enforced by signal costs, or handicaps. Animals of higher quality can better afford to pay these handicaps; therefore, animals of lower quality do not threaten or threaten with lower intensity (Zahavi, 1977, 1987; Andersson, 1982; Nur & Hasson, 1984; Grafen, 1990; Johnstone & Grafen, 1992). Under this hypothesis, communication systems in which threats do not impose handicaps are subject to bluffing, which in turn eliminates the incentive for receivers to respond to the threat. Therefore, displays with widespread bluffing are not expected in nature.

An alternative possiblity is that threat communication systems comprise a mix of reliable and deceptive signals. Deceptive threats have been documented in empirical studies (e.g. Steger & Caldwell, 1983; Adams & Caldwell, 1990) and may be expected as well on theoretical grounds (Dawkins & Krebs, 1978; Gardner & Morris, 1989; Dawkins & Guilford, 1991). Deception is possible because it is costly for receivers of displays to probe the signaller in ways that would discriminate bluffs from legitimate threats. Since the advantages of bluffing and probing are frequency dependent, deceptive threats may persist and succeed, as long as they remain at a sufficiently low frequency (Dawkins & Krebs, 1978; Wiley, 1983; Dawkins & Guilford, 1991). Indeed, proponents of the handicap principle have recognized that deception may be possible, but only if there is a limit to the frequency of

bluffing so that receivers, on average, benefit by respecting the threat display (Zahavi, 1987; Grafen, 1990).

In order to show how deceptive communication can be evolutionarily stable (Maynard Smith, 1982), it is necessary to explain what limits the frequency of deceptive threats. In some circumstances, the frequency of bluffs may be restricted by factors external to the signalling interaction. For example, in Batesian mimicry systems, the frequency of unreliable signals is set by the population densities of the models and mimics (Wiley, 1983). Similarly, if it is assumed that there are fixed proportions of two categories of signaller within a single species, then a stable mix of reliable and deceptive displays may be reached (Grafen, 1990; Johnstone & Grafen, 1993). Alternatively, the limited frequency of bluffing may be an outcome of the signalling interaction itself. Adams & Caldwell (1990) suggested that variation in costs and rewards to different animals could select for a mix of reliable and deceptive displays such that recipients are favored to respond to threats. Here we develop a model of threat communication in which the presence of a handicap and variable fighting costs leads to an ESS characterized by partial bluffing.

Previous models have shown that handicaps can ensure reliable communication if animals of lower quality pay higher costs for a given display, or if they are less able to afford those costs (Andersson, 1982; Nur & Hasson, 1984; Grafen, 1990). These models generally assume that the benefits deriving from a given signal are at least as great for animals of high quality as for animals of low quality. Thus, the net benefit for the use of the signal increases with the quality of the signaller. The result is that animals in poor condition are not favored to signal in the same way as animals of high condition, and display behavior is consequently a reliable indicator of the signaller's quality. The assumption of equal benefits may be appropriate in many signalling contexts, such as the attraction of mates. However, in the context of threat communication, it is likely that weak animals benefit more from the use of a threat display than do strong animals. This is because strong animals can win many conflicts without threatening (i.e. by direct fighting), while weak animals cannot. Furthermore, weak animals have more to gain by avoiding direct fights since they are less able to defend against injury. These points are developed further below. Since weak animals gain greater benefits for a given display, as well as paying greater average costs, the net benefit for a given advertisement may not increase monotonically with the signaller's strength. The pattern of display may therefore not provide fully reliable information about the signaller's condition.

This study was motivated by observations on bluffing by a stomatopod crustacean, *Gonodactylus bredini* (Steger & Caldwell, 1983; Caldwell, 1986; Adams & Caldwell, 1990). *G. bredini* are marine crustaceans that fight vigorously for possession of cavities in coral rubble. In addition to delivering blows with powerful raptorial appendages, many fighting animals use the meral spread threat display to deter opponents and to increase the probability of victory (Dingle & Caldwell, 1969; Adams & Caldwell, 1990). Surprisingly, newly molted residents threaten more often than intermolts, even though their soft condition renders them completely unable to fight (Steger & Caldwell, 1983; Adams & Caldwell, 1990). These bluffs significantly reduce the probability that the opponent will probe the signaller and discover its vulnerability (Adams & Caldwell, 1990). A striking feature of threat communication in *G. bredini* is that threats are especially likely by the weakest members of the population.

In this paper, we develop a game-theoretic model that examines the relationship between the costs of threat displays, threat reliability, and the stability of threat communication. In the following sections, we first clarify our use of the term "handicap" and outline our approach for including handicaps in models of threat communication. In Section 3, we describe a simple model of threat communication, and in Section 4 we describe the equilibrium solution of the game, which is characterized by partial bluffing. The full proof that this is the only ESS is presented in the Appendix. Since the principal result of the analysis may seem counterintuitive, we then discuss in Section 5 why stable bluffing is possible and why the result of this model differs from that of other models of the handicap principle. Finally, we discuss the implications of these results for understanding of threat communication and the behavior of *G. bredini*.

2. Handicaps in Threat Communication

In Zahavi's discussions of communication (1977, 1987), his examples encompass a variety of ways in which displays may handicap the signaller. For our purposes, two major categories may be distinguished. In the first, the handicap is incurred during the production of the signal. For example, the signal may be exhausting (e.g. Clutton-Brock & Albon, 1979), its development may consume large quantities of limited nutrients (e.g. antlers; Zahavi, 1987; Nur & Hasson, 1984), or its production may be enhanced by physiological changes that bear other costs (e.g.

Johnstone & Grafen, 1993). We call these handicaps "production costs". Production costs are paid by every animal that signals regardless of the effect of the signal upon the receiver. For example, it is costly to grow antlers whether or not the antlers deter rivals. Several models of the handicap principle assume costs of this kind (e.g. Nur & Hasson, 1984; Grafen, 1990).

A second type of handicap, also proposed by Zahavi (1977, 1987), arises when threat displays increase the vulnerability of the signaller. In this case, the handicap derives not from the energy or materials needed to produce the display, but from the increased risk of injury if the threat is not successful in deterring attack by the opponent. The production costs may be negligible. For example, threat displays in some species involve a lateral display which exposes the animal's flanks to the opponent. Under Zahavi's (1977) interpretation, a signaller adopting this posture has a greater risk of injury than if it directly faced its opponent; therefore, only animals that are truly strong can afford to take such a risk. We call this type of handicap a "vulnerability cost". In contrast to production costs, vulnerability costs are paid by the signalling animal only if the opponent attacks and is able to take advantage of the signaller's increased exposure. This kind of handicap is not paid if the threat deters the opponent or if the signaller is strong enough that it is not harmed by the increased vulnerability. Zahavi (1977, 1987) provides other examples of this kind and other authors have sometimes assumed that the handicaps for threat displays are due to increased vulnerability to injury (e.g. Enquist *et al*., 1985; Grafen, 1990). Furthermore, in empirical studies, the signal costs that are believed to underlie reliable communication are sometimes quantified by the probability of attacks towards the signaller, rather than by estimates of production costs (e.g. Popp, 1987; Waas, 1991). There is also supporting evidence that some threat displays produce vulnerability handicaps. For example, threat displays by fulmars deter attacks by some opponents, but increase the fighting costs for the signalling animal if its opponent does not withdraw (Enquist *et al*., 1985).

A distinction may also be drawn between threat displays that are graded, showing continuous variation in the degree of expression, and those that are discrete (Dawkins & Krebs, 1978). Several previous models of the handicap principle relevant to threat communication consider graded displays (Andersson, 1982; Nur & Hasson, 1984; Grafen, 1990; Johnstone & Grafen, 1992). An important result of these studies is that handicaps can lead to the evolution of graded displays in which the level of advertisement corresponds to the signaller's true abilities so that information is reliably extracted by the receiver.

In nature, many threat displays are discrete (Morris, 1957; Brown, 1975). While the display of the stomatopod, *G. bredini*, can be delivered with various intensities, cavity residents usually display with full intensity whether they are newly molted or between molts (E. Adams and R. Caldwell, personal observations). This stereotyped intensity potentially reduces the amount of information transmitted to the receiver. Nonetheless, discrete displays may carry handicaps, the magnitude of which depend upon the true quality of the signaller, and can conceivably convey reliable information to the receiver (see also Enquist, 1985).

3. A Threat Communication Game

This model concerns discrete threat displays that carry information about variation in the strength of the signaller that is not otherwise apparent from visual inspection. In stomatopods, this variation is due primarily to molt condition, but similar uncertainty could be caused by disease, hunger, or fatigue, so long as these disabilities do not produce obvious external cues. It is assumed that each animal knows its own strength, but cannot discern the true strength of an opponent except by fighting. In an escalated fight, both animals pay a cost, the true strengths are discovered, and the stronger animal wins.

We model a contest for an indivisible resource. The game is deliberately kept simple, encompassing only a single signal and response (Fig. 1). The first animal, called the signaller, begins with possession of the resource. The signaller decides whether to threaten or to remain with the resource without giving a threat display. The second animal, called the receiver, observes the behavior of the first animal, then decides whether to attack or to flee. Animals base their decisions in part upon their own strength. If the receiver attacks, a fight ensues and the stronger animal wins.

The parameters are defined as follows:

V = the gain in fitness derived from possession of the resource. This is the difference in fitness between an animal in control of a resource and an animal that must search elsewhere for other resources (Maynard Smith, 1982). An animal that flees without fighting receives a payoff of 0.

C = the cost of an escalated fight. This is the cost due to expenditure of time and energy or to injuries incurred while determining which animal has the greater strength. In an escalated contest, both the

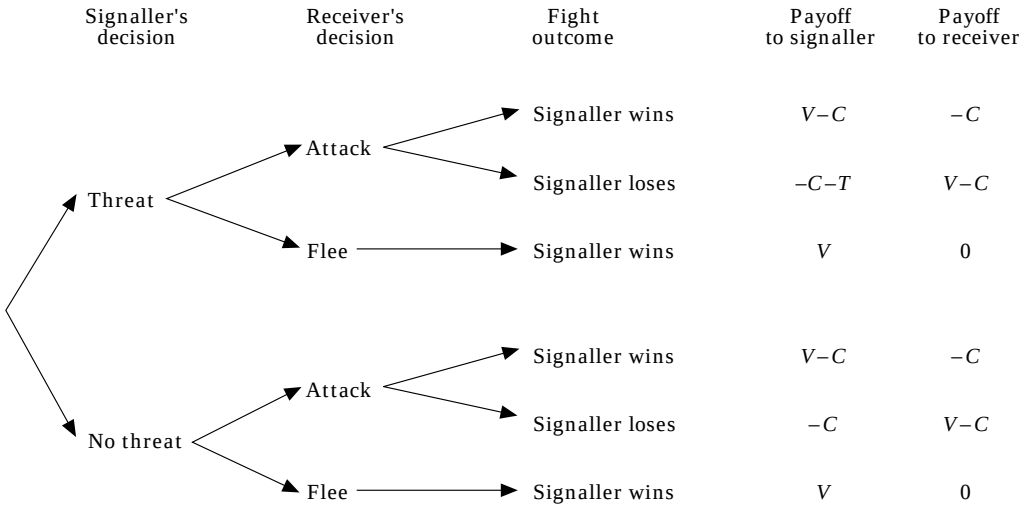

FIG. 1. Behavioral decisions and associated payoffs.

winner and the loser pay this cost, although its magnitude may differ for the two animals.

T = the handicap associated with the threat display.

The rules of the game and the assignment of payoffs are illustrated in Fig. 1. In alternative versions of the model, the handicap is represented either as a production cost or as a vulnerability cost; however, as shown below, only the vulnerability cost leads to an ESS. The costs associated with these two types of handicaps must be assigned differently. When the handicap is a production cost, then T is subtracted from the signaller's payoff whenever it threatens (the first three rows of Fig. 1). However, if the handicap is due to vulnerability, then subtraction of this cost term is contingent on the course of events in the subsequent interaction. The handicap is paid only if the opponent attacks and exploits the vulnerability of the signaller. If the receiver does not attack, no handicap is paid by the threatening animal because the production costs of the signal are negligible. Furthermore, it is assumed that the vulnerability handicap is paid only if the signaller loses. If the opponent attacks and the increased vulnerability of the signaller allows the opponent to inflict severe injury, then the handicap is paid and the signaller is more likely to lose the fight. However, if the opponent attacks but is not able to exploit the vulnerability of the signaller, then the handicap is not paid and the signaller is comparatively likely to win. For example, an animal that is truly strong and agile can recover quickly from a lateral display and fight effectively with its opponent. Thus, when the handicap is due to an increase in risk of injury, it is likely that handicaps are borne more heavily by animals that lose than by animals that win escalated fights. Other rules for the handicap are possible and the effects of altering these rules are discussed below.

In any case, the costs of escalated fights depend upon the strength of the fighting animals. Each population of fighting animals has a distribution of strengths. Let s represent a given animal's relative strength scaled between 0 and 1; e.g. if $s=0.85$, the animal can win escalated contests against 85% of opponents. A convenient way to represent variation in fighting costs is to assume that C is a linear function of s, such that the strongest animal pays A and the weakest animal pays a greater cost of $A+B$. Thus, $C=A+B(1-s)$. If $B=0$, then the costs of fighting are the same for all animals, but with $B>0$, weaker animals pay greater costs. To develop an intuitive description of the game, we assume first that $B=0$, which simplifies the mathematical expressions. However, it is shown in the appendix that there is an ESS only when $B>0$.

4. Bluffing in an Equilibrium Population

Consider first two possible solutions to this game (Fig. 2). A more complete set of strategies is described and analyzed in the Appendix. Here we wish to clarify the meaning of reliability and deception by reference to the most important possible solutions. In the first case [Fig. 2(a)], communication is "honest", since only strong animals threaten (those with $s>J$). From the receiver's point of view, the threat carries reliable information about the strength of the signaller in that animals that threaten are always stronger than animals that do not threaten. If the signaller threatens, then receivers with $s>L$ attack; others flee [Fig. 2(c)]. If the signaller does not threaten, then receivers with $s>K$ attack [Fig. 2(d)]. Since $L>K$, threat displays reduce

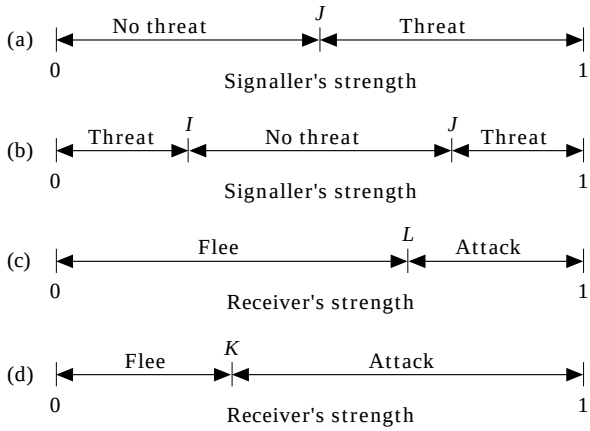

FIG. 2. Two conceivable threat communication systems. (a) No bluffing. Only signallers with strengths greater than the critical value J give threat displays. (b) Bluffing. The weakest signallers, those with strength less than the critical value I, also produce threat displays. (c) In either case, if the signaller threatens, receivers with strength greater than L attack. (d) If the signaller does not threaten, receivers with strength greater than K attack.

the probability of attack by receivers. This can be considered a reliable threat communication system, since animals that threaten are all stronger than animals that do not threaten, although some signallers drive off opponents against which they could not win an escalated contest (in the region $J<s<L$). We show below that such reliable communication is never an ESS for this game.

A threat communication system characterized by partial bluffing is shown in Fig. 2(b). In this case, threats are given by the weakest signallers ($s<I$) as well as by the strongest ($s>J$). Since the receivers cannot determine the true s of the signaller without escalation, these two categories of signals are equally effective in driving off opponents. Threat displays by the weak signallers can be called "bluffs", since none of these animals could win escalated fights against the opponents that are deterred by the threat display (those with $K<s<L$). Threat displays by the strongest animals are called "reliable threats", since any of these animals could win escalated fights against the opponents that are deterred by the display. This solution incorporates a remarkable feature of threat communication in stomatopods—the weakest animals often threaten; indeed, they do so in circumstances in which many stronger animals do not threaten (Adams & Caldwell, 1990). We will show that there is always an ESS of this form; that is, bluffing is favored for the weakest members of the population, yet the threat display does not lose its effectiveness.

I, J, K and L (Fig. 2) represent threshold values of s at which animals switch tactics. If the population is at equilibrium, then the payoffs to alternative tactics should be equal at these threshold values.

At the equilibrium value of I, the expected payoff to a signaller that threatens is the same as the expected payoff to a signaller that does not threaten. Assuming initially that fighting costs do not vary with s, the average payoff to an animal with $s=I$ that threatens is

$$L(V)+(1-L)(-C-T).$$

A proportion, L, of the receivers will flee, in which case the signaller receives a payoff of V, and the remainder of receivers (probability $=1-L$) will attack and win, with the signaller receiving a payoff of $-C-T$ (see Fig. 1). Similarly, the average payoff to a signaller with $s=I$ that does not threaten is

$$K(V)+(1-K)(-C).$$

These expressions are nearly identical, but in the second case the animal does not pay the handicap, T, when it loses. Putting these together, at the equilibrium value of I:

$$L(V)+(1-L)(-C-T) = K(V)+(1-K)(-C). \quad (1)$$

Following similar reasoning, the equations for fitness at J, K, and L, respectively are:

$$L(V)+(J-L)(V-C)+(1-J)(-C-T) = KV+(J-K)(V-C)+(1-J)(-C) \quad (2)$$

$$0=[(K-I)/(J-I)](V-C)+[(J-K)/(J-I)](-C) \quad (3)$$

$$0=[I/(I+1-J)](V-C)+[(1-J)/(I+1-J)](-C). \quad (4)$$

The left halves of eqns (3) and (4) are set to zero, the payoff for fleeing, since the payoff for fleeing is equal to the payoff for attack at the threshold values of K and L. Expressions (1–4) can be solved to give I, J, K and L, in terms of V, C and T. This yields:

$$I=C^2/[V(V+C+T)] \quad (5)$$

$$J=(V^2+C^2+TV)/[V(V+C+T)] \quad (6)$$

$$K=C/V \quad (7)$$

$$L=(CV+C^2+TV)/[V(V+C+T)]. \quad (8)$$

The requirement that $1>J>L>K>I>0$, as depicted in Fig. 2, is met provided that V, C and T are positive and that $V>C$. The Appendix considers other possible orderings of I, J, K, and L; however, none produces an ESS.

So long as weaker animals suffer greater costs of fighting (i.e. $B>0$), then there is a unique ESS (see Appendix). At the ESS, $1>J>L>K>I>0$ as depicted in Fig. 2. Figure 3 illustrates how the values of the thresholds I, J, K and L depend upon the cost of escalated fights, varied in this case by increasing A, which increases fighting costs for all animals. As fighting costs rise, the fraction of weak animals bluffing (those with $s<I$) increases steadily while the fraction of strong animals threatening (those with $s>J$) increases, then declines. A broader range of receivers flee as costs rise and the proportion deterred by the threat, given by $L-K$, diminishes. Qualitatively similar results are obtained for other values of V, T and B.

The effects of variation in the magnitude of the handicap are shown in Figure 4. As the handicap becomes more severe, fewer animals deliver either bluffs or reliable threats, and the range of receivers that are deterred by the threat, given by $L-K$, increases. The highest values of T shown in Fig. 4 are probably biologically unrealistic, but it is helpful to note that the ESS persists regardless of the value of T.

Thus far, the model has assumed that the handicap is paid only by animals that threaten and lose contests (Fig. 1). If instead it is assumed that animals that threaten always pay the handicap, regardless of the outcome of the fight, then it can be shown that there is no ESS either with or without bluffing (see Appendix).

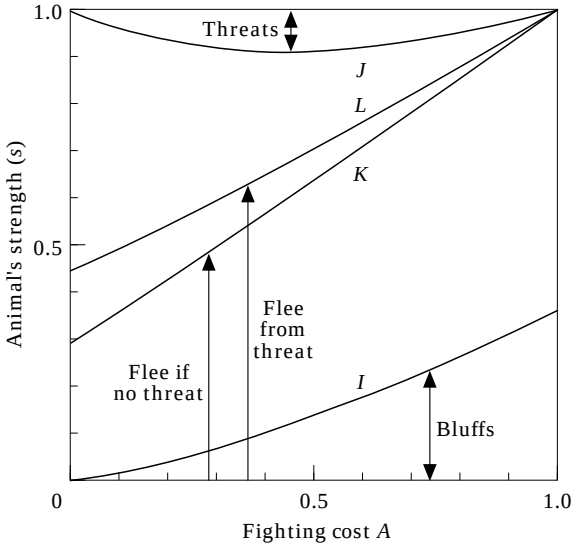

FIG. 3. The effect of varying A, the component of fighting costs that does not vary with strength, upon the ESS thresholds. (See Fig. 2 for an explanation of the four thresholds.) In this example, $V=1$, $B=0.4$, and $T=0.4$.

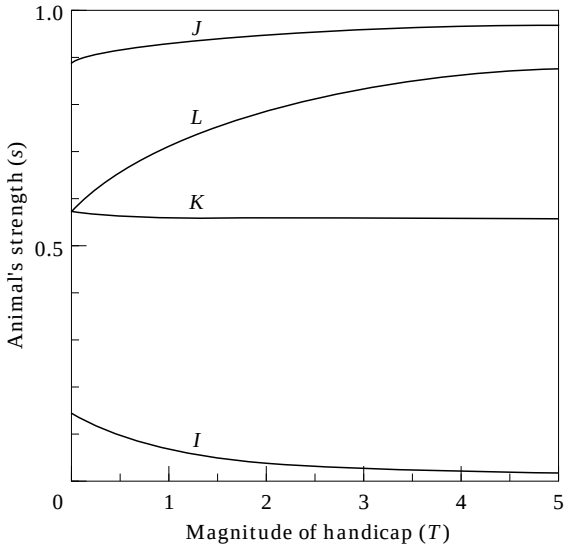

FIG. 4. The effect of varying T, the magnitude of the handicap, upon the ESS thresholds. (See Fig. 2 for an explanation of the four thresholds.) In this example, $V=1$, $A=0.4$, and $B=0.4$.

5. Why Bluffing is Favored

As described above, analysis of the model shows that threats are favored for both the strongest and the weakest members of the population, but not for animals of intermediate strength. Because this outcome may seem counterintuitive, we discuss here how bluffing by the weakest animals is possible in a stable system of threat communication.

Consider an animal that must decide whether or not to deliver a threat display. Although the payoff cannot be predicted with certainty in a particular case, because the strength of the opponent is unknown, the average payoff can be calculated for an animal of any given strength (s). Figure 5(d) shows the relative payoff for signallers that threaten as a function of s; that is, the difference between the expected fitness if the animal threatens and the expected fitness if the animal does not threaten. The resident should threaten if the net relative payoff is positive; otherwise, it should refrain from threatening. Notice that the graph of the net relative payoff crosses the abscissa twice, dividing the spectrum of animal strengths into three regions: weak animals ($s\leqslant I$), which are favored to threaten; moderate animals ($I\leqslant s\leqslant J$), which should not threaten; strong animals ($s\geqslant J$), which should threaten.

To understand this in further detail, various effects of the threat display can be examined separately. The costs associated with the threat display are due to the vulnerability handicap. Since the handicap is paid only if the signaller threatens and loses, the expected handicap can be calculated as the product of the

probability of losing and T, the magnitude of the handicap [Fig. 5(a)]. These costs are especially strong for weak animals, which are more likely to lose contests. This aspect of our model is in keeping with Zahavi's arguments and with some previous models of handicaps: the average handicap is greater for weaker animals, thus strong animals may be favored to signal when weaker animals are not.

However, the benefits associated with threats also vary with the signaller's strength. Threats produce two kinds of benefits. First, by threatening, the resident may win some fights that it would otherwise lose. This benefit is enjoyed chiefly by weak animals. The magnitude of this benefit is the product of V, the value of the resource, and the increased likelihood of winning due to the threat display. Only receivers with strengths between K and L are deterred by threats. Since strong signallers (those with strengths greater than L) would win all of these contests even if they did not threaten, they do not increase their probability of victory by displaying. Animals weaker than K benefit the most from threats, since they would not otherwise win any contests with the opponents that flee from threats. Between K and L, the benefit declines linearly with resident strength [Fig. 5(b)].

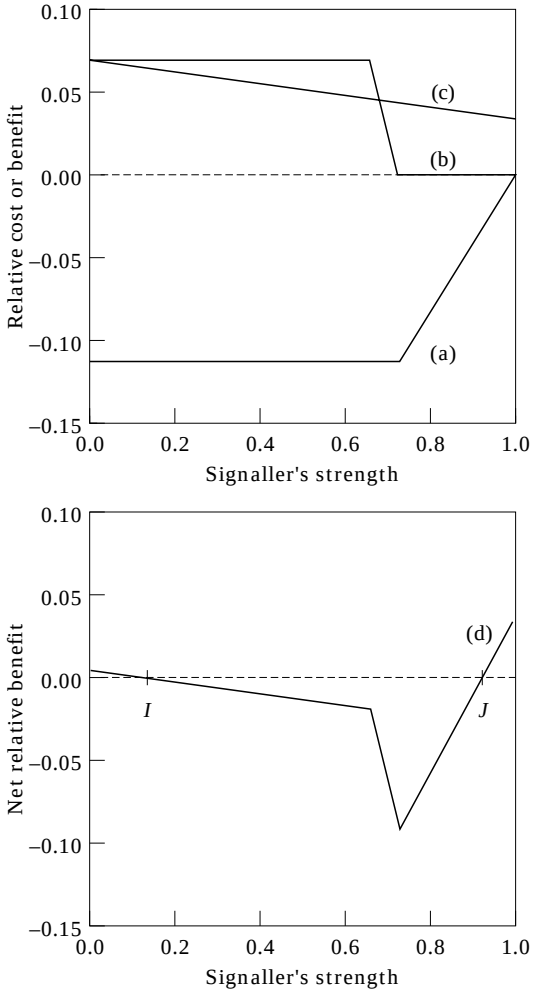

FIG. 5. The costs and benefits of threatening, relative to not threatening, as a function of the signaller's strength. In this example, $V=1$, $A=0.5$, $B=0.5$, and $T=0.5$. (a) The average handicap paid. (b) The relative benefit due to the increased likelihood of victory. (c) The relative benefit due to decreased fighting costs. (d) The sum of all relative costs and benefits.

The second benefit of threatening is that it is cheaper to win a contest by the use of a display than by escalated fighting. This also benefits weak animals more than strong animals, since weak animals suffer greater fighting costs in escalated fights [Fig. 5(c)]. By driving off receivers with strengths between K and L, threat displays lower fighting costs for a fraction of encounters. The expected benefit is given by the product of this fraction, $L-K$, and the cost of engaging in an escalated fight, $C=A+B(1-s)$.

When the cost and benefit curves are added together, it can be seen that threats are favored for strong animals and for weak animals [Fig. 5(d)]. At the strong end of the spectrum, the expected handicap changes faster with resident's strength than does the expected benefit; this favors threats by the stronger animals. At the weak end of the spectrum, the benefits change faster with strength than the costs. This favors threats by the weakest animals.

We do not suggest that this double-crossing of payoff curves will characterize all models of threat communication. But neither is this result dependent upon a particular contrived example. The illustration in Fig. 5 adopts particular values for V, A, B and T, but a similar result is obtained in this game for any parameter values over the allowable range. Furthermore, this result persists with some minor changes in the rules of the game. For example, we have assumed in this paper that T and V do not vary with strength. However, if the magnitude of the handicap or the value of victory is greater for weaker animals, ESS solutions can still arise with bluffing by weak animals (unpublished results).

6. Discussion

The principal conclusion resulting from this analysis is that threat communication may be stable despite successful bluffing by the weakest members of the population. Indeed, such bluffing always forms part of the ESS in this game. Threats are delivered by the weakest and the strongest members of the population, but not by animals of intermediate strength. The result is a mix of reliable and deceptive displays, which are

both effective in eliciting withdrawals by opponents of intermediate strength.

When the population plays according to the evolutionarily stable strategy, some receivers of threat displays are deceived. Nevertheless, many threats are reliable and receivers would only lower their fitness by ignoring threat displays. This is true because a sufficiently large fraction of threats are given by strong animals that are able to inflict injury and are likely to win escalated fights. Responding to threats benefits the receivers on average, but lowers their fitness in some specific cases.

This result supports Zahavi's suggestion that a handicap is necessary for stable threat communication (Zahavi, 1977, 1982), but shows that handicaps do not ensure fully reliable communication in all signalling systems. Bluffers pay high costs when exposed by probing and weak signallers are especially likely to pay the handicap, but this does not weed out all deceptive threats. Despite the existence of bluffing, the cost of the threat is related to its effectiveness; that is, to the proportion of opponents that are deterred by the threat display, calculated as $L-K$ (Fig. 4). When there is no handicap (when T is 0), the threat has no effect on the receiver's behavior. In support of this prediction, studies of threat communication in birds show that threat displays are more effective when they cause a greater increase in the signaller's vulnerability to attack (Enquist *et al*., 1985; Popp, 1987; Waas, 1991). The model also shows that bluffs are not necessarily rare or even less common than legitimate threats. If the costs of escalated fights are large with respect to the value of the resource, a small number of legitimate threats can protect a large number of bluffers from discovery (Fig. 3).

The outcome of the game theoretic analysis is sensitive to particular assumptions about the signalling interaction and the assignment of handicap costs. The rules adopted are very simple (Fig. 1), representing only one of several possible ways to incorporate vulnerability handicaps. Not all of these will produce stable bluffing; indeed, the plausibility of complete "honesty" has been shown both for discrete threat displays (Enquist, 1985) and for displays that vary in intensity (Grafen, 1990). This variation in the conclusions of alternative game theoretical models implies that there will be diversity in the reliability of natural signalling systems. Honest signalling cannot be expected universally (see also Bond, 1989; Gardner & Morris, 1989; Dawkins & Guilford, 1991; Johnstone & Grafen, 1993).

Nonetheless, the features in our model that favor bluffing by weak animals are likely to arise in other models of threat communication, as long as the animals are assumed to have a spectrum of strengths. Previous models of the handicap principle have emphasized that, for the handicap principle to operate, animals of lower quality must suffer greater average costs from a given signal (Andersson, 1982; Grafen, 1990), or be less able to afford those costs (Nur & Hasson, 1984). This requirement is upheld in our model as well [Fig. 5(a)]. However, in addition to greater costs, weak animals also realize greater benefits from threat displays, as argued above. The changes in signal costs and benefits with animal strength do not precisely compensate for one another; instead, when they are added together, the resulting profile of predicted behaviors may have unexpected properties, including bluffing by weak animals [e.g. Fig. 5(d)]. By contrast, when the function of a signal is to attract mates, rather than to deter competitors, there is no obvious reason why the average benefits of a given signal should be greater for animals of lower quality. As a result, models of communication in other contexts usually assume that the benefits of a given level of advertisement are equal or greater for animals of higher quality (Nur & Hasson, 1984; Grafen, 1990). The result is a fully reliable display.

Many assumptions of this model are confirmed by data on threat communication in *G. bredini*. In agreement with the model, hidden variation in fighting ability has strong effects on the probable outcome of fights (Steger & Caldwell, 1983; Adams & Caldwell, 1990). Animals adjust their fighting tactics according to changes in their own physical condition (Adams & Caldwell, 1990). Receivers cannot distinguish bluffs from legitimate threats, except by risky probing (Adams & Caldwell, 1990). Moreover, the threat display of *G. bredini* does not appear to be a production handicap. The meral spread threat display is a presentation of weaponry by an extension of the raptorial appendages (Dingle & Caldwell, 1969), which is unlikely to consume significant amounts of energy. However, adopting this posture reduces surprise and lowers the speed at which a strike can be delivered (R. L. Caldwell, personal communication), which may increase the odds that an opponent can deliver the first blow. The game-theoretic analysis provides a possible explanation for succesful threats by newly molted animals.

In the interests of simplicity, the model makes numerous restrictive assumptions. For example, it is assumed that the resource is of equal value to all animals. In *G. bredini*, it is more likely that weak animals suffer greater penalties than strong animals from loss of the cavity due to their vulnerability to predators and their inability to supplant intermolt stomatopods elsewhere on the reef. This variation

in resource value should increase the likelihood of bluffing by newly molted animals. The newly molted resident's best chance for survival is to drive off its opponents without allowing probing. Thus, weak animals may be better able to afford the cost of threats than strong animals for which the cavity is not as valuable.

Another recent model (Gardner & Morris, 1989) also examines discrete signals with specific reference to bluffing in *G. bredini*. In Gardner & Morris' model, a resident may either flee or defend, but all residents that defend cavities use the threat display. The use of the threat is thus completely coincident with the decision to fight. From the receiver's point of view, all animals with which they might fight deliver threats; thus, no information is provided by the threat and no decision by the recipient varies with the signallers' display behavior. This is better considered to be a model of fighting decisions, rather than of signalling *per se*. Their results show that weak animals are less likely to fight than strong animals for some parameter values and that the probability of fighting by small animals may change cyclically. No such cyclical fluctuations have been observed in *G. bredini*, but it is valuable to note that the dynamics of fighting decisions can be analyzed even when there is no point equilibrium.

This research was supported by a Postdoctoral Fellowship from the Smithsonian Institution and by a Fellowship in Science and Engineering from the David and Lucile Packard Foundation to ESA. We thank R. L. Caldwell for numerous discussions on communication and deception in stomatopods and N. Knowlton for helpful comments on an early version of the manuscript.

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, M. (1982). Sexual selection, natural selection and quality advertisement. *Biol. J. Linn. Soc.* **17,** 375–393.

Bond, A. B. (1989). Toward a resolution of the paradox of aggressive displays: I. Optimal deceit in the communication of fighting ability. *Ethology* **81,** 29–46.

Brown, J. L. (1975). *The Evolution of Behavior*. New York: W. W. Norton.

Caldwell, R. L. (1986). The deceptive use of reputation by stomatopods. In: *Deception: Perspectives on Human and Nonhuman Deceit* (Mitchel, R. W. & Thompson, N. S., eds) pp. 129–145. Albany: State University of New York Press.

Clutton-Brock, T. H. & Albon, S. D. (1979). The roaring of red deer and the evolution of honest advertisement. *Behaviour* **69,** 145–170.

Dawkins, M. S. & Guilford, T. (1991). The corruption of honest signalling. *Anim. Behav.* **41,** 865–873.

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

Dingle, H. & Caldwell, R. L. (1969). The aggressive and territorial behaviour of the mantis shrimp *Gonodactylus bredini* Manning (Crustacea: Stomatopoda). *Behaviour* **33,** 115–136.

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

Enquist M., Plane, E. & Röed, J. (1985). Aggressive communication in fulmars (*Fulmarus glacialis*). *Anim. Behav.* **33,** 1007–1020.

Gardner, R. & Morris, M. R. (1989). The evolution of bluffing in animal contests: an ESS approach. *J. theor. Biol.* **137,** 235–243.

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

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

Johnstone, R. A. & Grafen, A. (1993). Dishonesty and the handicap principle. *Anim. Behav.* **46,** 759–764.

Markl, H. (1985). Manipulation, modulation, information, cognition: some of the riddles of communication. In: *Experimental Behavioral Ecology and Sociobiology* (Hölldobler, B. & Lindauer, M., eds) pp. 163–194. Sunderland, MA: Sinauer.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Mesterton-Gibbons, M. (1992). *An Introduction to Game-Theoretic Modelling*. Redwood City, California: Addison-Wesley.

Morris, D. (1957). Typical intensity and its relation to the problem of ritualization. *Behaviour* **11,** 1–12.

Nur, N. & Hasson, O. (1984). Phenotypic plasticity and the handicap principle. *J. theor. Biol.* **110,** 275–297.

Popp, J. W. (1987). Risk and effectiveness in the use of agonistic displays by American goldfinches. *Behaviour* **103,** 141–156.

Steger, R. & Caldwell, R. L. (1983). Intraspecific deception by bluffing: a defense strategy of newly molted stomatopods (Arthropoda: Crustacea). *Science, N.Y.* **221,** 558–560.

Waas, J. R. (1991). The risks and benefits of signalling aggressive motivation: a study of cave-dwelling little blue penguins. *Behav. Ecol. Sociobiol.* **29,** 139–146.

Wiley, R. H. (1983). The evolution of communication: information and manipulation. In: *Communication* (Halliday, T. R. & Slater, P. J. B., eds) pp. 156–189. New York: W. H. Freeman.

Zahavi, A. (1977). Reliability in communication systems and the evolution of altruism. In: *Evolutionary ecology* (Stonehouse, B. & Perrins. C. M., eds) pp. 253–259. London: Macmillan.

Zahavi, A. (1982). The pattern of vocal signals and the information they convey. *Behaviour* **80,** 1–8.

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

APPENDIX

Mathematical Model

In this Appendix we develop a formal mathematical model of the game presented heuristically in the main body of the paper. More detailed discussion of the relevant game-theoretic concepts can be found in Mesterton-Gibbons (1992), whose notation we adopt.

First we define the strategy set, S, to be the set of all four-dimensional vectors whose components are numbers between 0 and 1. We interpret S as follows. Let the vector $\mathbf{u}=(u_1, u_2, u_3, u_4)$ denote the strategy of a representative individual or protagonist, called Player 1 for convenience, and let the random variable

TABLE A1
Payoffs to resident protagonist and intruding opponent

Relative magnitudes of X and Y	Protagonist's payoff	Opponent's payoff
$X<u_1$ or $X>u_2$ and $Y>v_4$	$\rho(X, Y)$	$\sigma(Y, X)$
$X<u_1$ or $X>u_2$ and $Y<v_4$	V	0
$u_1<X<u_2$ and $Y>v_3$	$\sigma(X, Y)$	$\sigma(Y, X)$
$u_1<X<u_2$ and $Y<v_3$	V	0

X denote Player 1's fighting strength. We assume that X is continuously distributed between 0 and 1. Then, in the role of resident, Player 1 threatens if either $X<u_1$ or $X>u_2$, but does not threaten if $u_1<X<u_2$; because X is continuously distributed, the event that $X=u_1$ or $X=u_2$ occurs with zero probability, and so we can safely ignore it. In the role of intruder, on the other hand, Player 1 attacks when $X>u_4$ if its opponent threatens but when $X>u_3$ if its opponent does not threaten; correspondingly, Player 1 flees when $X<u_4$ or when $X<u_3$ according to whether its opponent threatens or not (and $X=u_3$ or $X=u_4$ again occurs with zero probability).

Similarly, let $\mathbf{v}=(v_1, v_2, v_3, v_4)$ denote the strategy of Player 1's opponent, called Player 2 for convenience, and let the random variable Y denote Player 2's fighting strength, which again is continuously distributed between 0 and 1. The interpretation of $\mathbf{v}$ is analogous to that of $\mathbf{u}$: for example, in the role of resident, Player 2 threatens if either $Y<v_1$ or $Y>v_2$.

Using notation that temporarily suppresses dependence on $\mathbf{u}$ and $\mathbf{v}$ let $F(X, Y)$ denote the payoff to the $\mathbf{u}$-strategist (Player 1) against the $\mathbf{v}$-strategist (Player 2), let $F_k(X, Y)$ denote the payoff to $\mathbf{u}$ against $\mathbf{v}$ in role k, and let p_k be the probability of occupying role k. Then, if r stands for resident and i for intruder, we have $p_r+p_i=1$ and

$$F(X, Y)=p_rF_r(X, Y)+p_iF_i(X, Y). \qquad (A.1)$$

Note that F, F_r and F_i are random variables because X and Y are random variables. Accordingly, let fighting strength X or Y be continuously distributed between 0 and 1 with density g, let E denote expected value over the joint distribution of X and Y, and define $f_k(\mathbf{u}, \mathbf{v})=E[F_k(X, Y)]$. Then for $k=$ r and $k=$ i we have

$$f_k(\mathbf{u}, \mathbf{v})=\int_0^1\int_0^1 F_k(x, y)g(x)g(y)\,dx\,dy \qquad (A.2)$$

and, from (A.1), the reward or expected payoff to Player 1 against Player 2 is

$$f(\mathbf{u}, \mathbf{v})=E[F(X, Y)]=p_rf_r(\mathbf{u}, \mathbf{v})+p_if_i(\mathbf{u}, \mathbf{v}). \qquad (A.3)$$

We interpret $f(\mathbf{u}, \mathbf{v})$ as the reward to a $\mathbf{u}$-strategist in a population of $\mathbf{v}$-strategists.

It is convenient at this juncture to define functions ρ and σ of X and Y as follows:

$$\rho(X, Y)=\begin{matrix} V-C & \text{if} & X>Y \\ -C-T & \text{if} & X<Y \end{matrix} \qquad (A.4)$$

$$\sigma(X, Y)=\begin{matrix} V-C & \text{if} & X>Y \\ -C & \text{if} & X<Y. \end{matrix} \qquad (A.5)$$

From Fig. 1, ρ and σ are the respective payoffs to a threatening or non-threatening resident protagonist against an attacking opponent. Then the payoffs from any contest to either contestant are defined by Tables A1 and A2; in each table, the first two rows correspond to threatening behavior by the resident, whereas the last two rows correspond to non-threatening behavior. Note that, by symmetry, the middle columns of both tables together imply the final columns. The middle columns of Tables A1 and A2 define, respectively, F_r and F_i. Thus if $u_1\leqslant u_2$

TABLE A2
Payoffs to intruding protagonist and resident opponent

Relative magnitudes of X and Y	Protagonist's payoff	Opponent's payoff
$Y<v_1$ or $Y>v_2$ and $X>u_4$	$\sigma(X, Y)$	$\rho(Y, X)$
$Y<v_1$ or $Y>v_2$ and $X<u_4$	0	V
$v_1<Y<v_2$ and $X>u_3$	$\sigma(X, Y)$	$\sigma(Y, X)$
$v_1<Y<v_2$ and $X<u_3$	0	V

we have

$$f_r(\mathbf{u}, \mathbf{v}) = \int_0^{u_1}\int_{v_4}^1 \rho(x, y)g(x)g(y)\,dx\,dy + \int_{u_2}^1\int_{v_4}^1 \rho(x, y)g(x)g(y)\,dx\,dy + \int_0^{u_1}\int_0^{v_4} Vg(x)g(y)\,dx\,dy + \int_{u_2}^1\int_0^{v_4} Vg(x)g(y)\,dx\,dy + \int_{u_1}^{u_2}\int_{v_3}^1 \sigma(x, y)g(x)g(y)\,dx\,dy + \int_{u_1}^{u_2}\int_0^{v_3} Vg(x)g(y)\,dx\,dy \quad (A.6)$$

and

$$f_i(\mathbf{u}, \mathbf{v}) = \int_{u_4}^1\int_0^{v_1} \sigma(x, y)g(x)g(y)\,dx\,dy + \int_{u_4}^1\int_{v_2}^1 \sigma(x, y)g(x)g(y)\,dx\,dy + \int_{u_3}^1\int_{v_1}^{v_2} \sigma(x, y)g(x)g(y)\,dx\,dy, \quad (A.7)$$

where in each case the first integral sign corresponds to integration variable x, and the second to variable y. If $u_1 > u_2$, however, then the **u**-strategist always threatens, and in place of (A.6) we have

$$f_r(\mathbf{u}, \mathbf{v}) = \int_0^1\int_{v_4}^1 \rho(x, y)g(x)g(y)\,dx\,dy + \int_0^1\int_0^{v_4} Vg(x)g(y)\,dx\,dy = W, \quad (A.8)$$

say, where W is independent of **u**. Thus any strategy satisfying $u_1 > u_2$ is equivalent mathematically to any strategy satisfying $u_1 = u_2$, and so from now on we restrict the strategy set to

$$S = \{\mathbf{u} = (u_1, u_2, u_3, u_4) | 0 \leqslant u_1 \leqslant u_2 \leqslant 1,\ 0 \leqslant u_3 \leqslant 1,\ 0 \leqslant u_4 \leqslant 1\}. \quad (A.9)$$

It is another question entirely whether the strategies thus excluded are all equivalent biologically: an animal that always threatens because its "high" threshold (for reliable communication) is normal and its "low" threshold (for deceitful communication) abnormally high may be said to behave very differently from an animal that always threatens because its low threshold is normal and its high threshold abnormally low, but our game does not distinguish between them. Nevertheless, the question becomes irrelevant because $u_1 < u_2$ at the only ESS.

Now, $\mathbf{u}^* \in S$ [defined by (A.9)] is an evolutionarily stable strategy, or ESS, if

$$f(\mathbf{u}^*, \mathbf{u}^*) \geqslant f(\mathbf{v}, \mathbf{u}^*) \quad \text{for all } \mathbf{v} \in \mathrm{S} \quad (A.10a)$$

and

$$f(\mathbf{u}^*, \mathbf{u}^*) > f(\mathbf{v}, \mathbf{u}^*) \quad (A.10b)$$

or

$$f(\mathbf{u}^*, \mathbf{v}) > f(\mathbf{v}, \mathbf{v}) \quad (A.10c)$$

for all $\mathbf{v} \in S$ such that $\mathbf{v} \neq \mathbf{u}^*$, where "$\in$" means "belonging to". If only (A.10a) were satisfied, then $\mathbf{u}^*$ could not be selected against, but it could be invaded by random drift. If (A.10b) is satisfied for all $\mathbf{v} \neq \mathbf{u}^*$, then $\mathbf{u}^*$ is a strong ESS; otherwise $\mathbf{u}^*$ is a weak ESS. To obtain an ESS of our game we first calculate the "rational reaction set" R, defined by

$$R = \{(\mathbf{u}, \mathbf{v}) | \quad \mathbf{u} \in S, \quad \mathbf{v} \in S, \quad f(\mathbf{u}, \mathbf{v}) = \max f(\bar{\mathbf{u}}, \mathbf{v})\}, \quad (A.11)$$

which contains all feasible strategy combinations $(\mathbf{u}, \mathbf{v})$ such that **u** is a $\bar{\mathbf{u}}$-strategist's best reply to **v** in the sense of maximizing f. From (A.10), $\mathbf{u}^*$ can be an ESS only if $(\mathbf{u}^*, \mathbf{u}^*) \in R$. If also $\mathbf{u}^*$ is uniquely the best reply to itself—that is, if $(\mathbf{u}^*, \mathbf{u}^*) \in R$ but $(\mathbf{u}, \mathbf{u}^*) \notin R$ for $\mathbf{u} \neq \mathbf{u}^*$—then $\mathbf{u}^*$ is a strong ESS. If, on the other hand, $(\mathbf{u}^*, \mathbf{u}^*) \in R$ but there exist alternative best replies to $\mathbf{u}^*$, then $\mathbf{u}^*$ is at best a weak ESS (and often, as here, is not an ESS at all). For a more detailed discussion of the concept of rational reaction set see, for example, Mesterton-Gibbons (1992).

To calculate R we must maximize f defined by (A.3) with respect to u. We first observe from (A.6)–(A.7) that $f_r(\mathbf{u}, \mathbf{v})$ is independent of u_3 and u_4, whereas $f_i(\mathbf{u}, \mathbf{v})$ is independent of u_1 and u_2. Using notation that suppresses dependence on **v**, we may write

$$f_i(\mathbf{u}, \mathbf{v}) = f_3(u_3) + f_4(u_4), \quad (A.12)$$

where

$$f_3(u_3)=\int_{u_3}^{1}\int_{v_1}^{v_2}\sigma(x,y)g(x)g(y)\,\mathrm{d}x\,\mathrm{d}y \tag{A.13}$$

and f_4 is the sum of the first two integrals in (A.7). Note that $v_1\leqslant v_2$ in the above expression because $\mathbf{v}\in S$ [defined by (A.9)]. Furthermore, defining $\tau=\rho-\sigma$, i.e.

$$\tau(X,Y)=\begin{matrix}0 & \text{if} & X>Y\\ -T & \text{if} & X<Y\end{matrix} \tag{A.14}$$

(and again using notation that suppresses dependence on $\mathbf{v}$), we may write

$$f_r(\mathbf{u},\mathbf{v})=f_1(u_1)+f_2(u_2), \tag{A.15}$$

where

$$f_1(u_1)=\int_0^{u_1}\int_{v_3}^{v_4}(V-\sigma(x,y))g(x)g(y)\,\mathrm{d}x\,\mathrm{d}y + \int_0^{u_1}\int_{v_4}^{1}\tau(x,y)g(x)g(y)\,\mathrm{d}x\,\mathrm{d}y \tag{A.16}$$

and

$$\begin{aligned}f_2(u_2)=&\int_0^1\int_0^{v_3}Vg(x)g(y)\,\mathrm{d}x\,\mathrm{d}y\\ &+\int_{u_2}^1\int_{v_3}^{v_4}(V-\sigma(x,y))g(x)g(y)\,\mathrm{d}x\,\mathrm{d}y\\ &+\int_{u_2}^1\int_{v_4}^{1}\tau(x,y)g(x)g(y)\,\mathrm{d}x\,\mathrm{d}y\\ &+\int_0^1\int_{v_3}^{1}\sigma(x,y)g(x)g(y)\,\mathrm{d}x\,\mathrm{d}y\\ =&\,W-f_1(u_2),\end{aligned} \tag{A.17}$$

where W is defined by (A.8). Thus maximization with respect to u_3 may be performed separately from that with respect to u_4, and both independently of that with respect to u_1 or u_2. This separability of the reward function makes the game tractable analytically.

Some general features of R now follow directly from the inequalities $V>0$, $C>0$ and $T>0$. From (A.5) and (A.13), if $u_3<v_1<v_2$ then $f_3'(u_3)>0$, where the prime denotes differentiation, and so the maximum of f_3 over $0\leqslant u_3\leqslant 1$ must occur where $v_1\leqslant u_3\leqslant 1$ for any realistic probability density function g, i.e. any g such that $g(\xi)>0$ for $0\leqslant\xi\leqslant 1$ (except perhaps at isolated points where $g=0$). Again, for any realistic g, (A.5), (A.14) and (A.16) imply that if $v_4\leqslant v_3$ then $f'_1(u_1)<0$ for all $0\leqslant u_1\leqslant 1$ unless $v_4=v_3=1$. Thus if $v_4\leqslant v_3$ then the maximum of f_1 must occur at $u_1=0$; unless $v_4=v_3=1$, in which case f_1 is independent of u_1. Correspondingly, (A.17) yields $f_2'(u_2)>0$ for all $0\leqslant u_2\leqslant 1$ unless $v_4=v_3=1$, and so the maximum of f_2 must occur at $u_2=1$; unless $v_4=v_3=1$, in which case f_2 is independent of u_2.

Nevertheless, we cannot completely calculate R without specifying V, C, T and g in (A.4)–(A.8). Accordingly, we assume that V and T are constant, but that C decreases linearly with fighting strength according to

$$C=A+B(1-X), \tag{A.18}$$

with A, B constant and $0<A<V$, $B>0$. (Of course, (A.18) is defined from the point of view of the protagonist: the corresponding cost for its opponent is $C=A+B(1-Y)$.) Thus $V>C$ for the strongest animal, and $V>C$ for every animal in the limit as $B\to 0$, but in general there may be (weaker) animals for which $C>V$. It is conveient at this juncture to define three dimensionless parameters, as follows:

$$a=\frac{A}{V},\quad b=\frac{B}{V},\quad t=\frac{T}{V}. \tag{A.19}$$

Note that $0<a<1$, $b>0$ and $t>0$, by assumption; nevertheless, it will be instructive at a later point to consider the limits $a\to 0$, $b\to 0$ and $a\to 1$. We also assume that fighting strength is uniformly distributed between 0 and 1, i.e.

$$g(\xi)=1,\quad 0\leqslant\xi\leqslant 1. \tag{A.20}$$

We now proceed with the maximization in (A.11). From (A.13), (A.18) and (A.20) we readily find that

$$f_3(u_3)=\tfrac{1}{2}V(v_2-v_1)(1-u_3)\{2(1-a) -b(1-u_3)\}-\tfrac{1}{2}V(v_2-u_3)^2 \tag{A.21}$$

if $v_1\leqslant u_3\leqslant v_2$, whereas the last (squared) term in (A.21) must be omitted to obtain the correct expression for f_3 when $v_2\leqslant u_3\leqslant 1$. It is then straightforward to show that the maximum of f_3 on $v_1\leqslant u_3\leqslant 1$ (and hence also on $0\leqslant u_3\leqslant 1$) occurs at $u_3=$

$$\theta_3 \quad \text{if} \quad b(1-v_2)\leqslant 1-a \tag{A.22a}$$

$$1-(1-a)/b \quad \text{if} \quad b(1-v_2)>1-a, \tag{A.22b}$$

where

$$\theta_3=\frac{v_1+(a+b)(v_2-v_1)}{1+b(v_2-v_1)}. \tag{A.23}$$

TABLE A3
Maximum of f_i defined by (A.7) on $0 \leqslant u_3, u_4 \leqslant 1$

Relative magnitudes of $v_1 \leqslant v_2$	Where maximum occurs u_3	u_4	Conditions on (u_3, u_4)
$\delta < bv_1,\ b(1-v_2) \leqslant 1-a$	θ_3	(A.28a)	$v_1 \leqslant u_3 \leqslant v_2,\ u_4 < v_1$
$bv_1 \leqslant \delta \leqslant bv_2,\ b(1-v_2) \leqslant 1-a$	θ_3	θ_4	$v_1 \leqslant u_3,\ u_4 \leqslant v_2$
$\delta > bv_2,\ b(1-v_2) \leqslant 1-a$	θ_3	(A.28c)	$v_1 \leqslant u_3 \leqslant v_2,\ u_4 > v_2$
$\delta < bv_1,\ b(1-v_2) > 1-a$	$1-(1-a)/b$	(A.28a)	$u_3 > v_2,\ u_4 < v_1$
$bv_1 \leqslant \delta \leqslant bv_2,\ b(1-v_2) > 1-a$	$1-(1-a)/b$	θ_4	$u_3 > v_2,\ v_1 \leqslant u_4 \leqslant v_2$
$\delta > bv_2,\ b(1-v_2) > 1-a$	$1-(1-a)/b$	(A.28c)	$u_3 > v_2,\ u_4 > v_2$
$v_1 = 0,\ v_2 = 1$	$(a+b)/(1+b)$	u_4	u_4 arbitrary

Note that $v_1 \leqslant u_3 \leqslant v_2$ in (A.22a), whereas $v_2 < u_3 \leqslant 1$ in (A.22b). A similar calculation shows that

$$f_4(u_4) = \tfrac{1}{2}V(1-u_4)\{2v_1 - (2a + b(1-u_4)) \times (v_1 - v_2 + 1)\} + \tfrac{1}{2}V(1-v_2)^2 - \tfrac{1}{2}V(v_1-u_4)^2 \quad \text{(A.24)}$$

if $0 \leqslant u_4 \leqslant v_1$; whereas the last (negative squared) term must be omitted from (A.24) to obtain the correct expression for f_4 when $v_1 \leqslant u_4 \leqslant v_2$, and

$$f_4(u_4) = \tfrac{1}{2}V(1-u_4)\{2v_1 - 2v_2 + u_4 + 1 - (v_1 - v_2 + 1)(2a + b(1-u_4))\} \quad \text{(A.25)}$$

when $v_2 \leqslant u_4 \leqslant 1$. Then if we define

$$\theta_4 = \frac{(a+b)(v_1 - v_2 + 1) - v_1}{b(v_1 - v_2 + 1)} \quad \text{(A.26)}$$

and

$$\delta = \frac{(a+b)(v_1 - v_2 + 1) - v_1}{v_1 - v_2 + 1}, \quad \text{(A.27)}$$

the maximum of f_4 on $0 \leqslant u_4 \leqslant 1$ can be shown to occur where $u_4 =$

$$\frac{(a+b)(v_1 - v_2 + 1)}{1 + b(v_1 - v_2 + 1)} \quad \text{if} \quad \delta < bv_1 \quad \text{(A.28a)}$$

$$\theta_4 \quad \text{if} \quad bv_1 \leqslant \delta \leqslant bv_2 \quad \text{(A.28b)}$$

$$\frac{(a+b)(v_1 - v_2 + 1) - v_1 + v_2}{1 + b(v_1 - v_2 + 1)} \quad \text{if} \quad \delta > bv_2, \quad \text{(A.28c)}$$

provided $v_1 - v_2 + 1 \neq 0$. Note that $0 \leqslant u_4 < v_1$ in (A.28a) and $v_1 \leqslant u_4 \leqslant v_2$ in (A.28b), whereas $v_2 < u_4 \leqslant 1$ in (A.28c). If $v_1 - v_2 = 0$, which can happen only if $v_1 = 0$ and $v_2 = 1$, then f_3 is maximized at $u_3 = (a+b)/(1+b)$ and any u_4 maximizes f_4. Taken together with (A.12), these results imply that the maximum of f_i is given by Table A3.

We have already established that when $v_3 \geqslant v_4$, f_r is maximized for $0 \leqslant u_1 \leqslant u_2 \leqslant 1$ where $u_1 = 0, u_2 = 1$ (unless $v_3 = v_4 = 1$, in which case both u_1 and u_2 are arbitrary). Moreover, it is clear from (A.5) and (A.16)–(A.17) that f_r is maximized where $u_1 = u_2$ if $v_3 < v_4 = 1$. Let us therefore assumed that $v_3 < v_4 < 1$, and hence that

$$b(1-v_4) < a + b(1-v_4) < a + b(1-v_3). \quad \text{(A.29)}$$

From (A.16), (A.18) and (A.20), we readily find that

$$f_1(u_1) = \tfrac{1}{2}Vu_1\{2(1+a+b)(v_4 - v_3) - 2t(1-v_4) - b(v_4 - v_3)u_1\} \quad \text{(A.30)}$$

if $0 \leqslant u_1 \leqslant v_3$, and that $V(u_1 - v_3)^2/2$ must be subtracted from (A.30) to obtain the correct expression for f_1 when $v_3 \leqslant u_1 \leqslant v_4$; whereas

$$f_1(u_1) = \tfrac{1}{2}V\{(v_4 - v_3)(v_4 + v_3 + u_1(2(a+b) - bu_1)) - 2tu_1(1-v_4) + t(u_1 - v_4)^2\} \quad \text{(A.31)}$$

if $v_4 \leqslant u_1 \leqslant 1$. In describing the shape of this function on $0 \leqslant u_1 \leqslant 1$, which depends on the relative magnitude of

$$\Delta = \frac{t(1-v_4)}{v_4 - v_3}, \quad \text{(A.32)}$$

it is convenient first to define θ_1, θ_2 by

$$\theta_1 = \frac{(1+a+b)(v_4 - v_3) - t(1-v_4)}{b(v_4 - v_3)} \quad \text{(A.33)}$$

$$\theta_2 = 1 - \frac{a(v_4 - v_3)}{t - b(v_4 - v_3)}. \quad \text{(A.34)}$$

Then routine application of the calculus shows that f_1 varies on $0 \leqslant u_1 \leqslant 1$ as follows. If $\Delta \geqslant 1 + a + b$, then f_1 decreases from $u_1 = 0$ to $u_1 = \theta_2 (> v_4)$ and increases again from $u_1 = \theta_2$ to $u_1 = 1$. If $1 + a + b > \Delta \geqslant 1 + a + b(1 - v_3)$, then f_1 increases from $u_1 = 0$ to $u_1 = \theta_1 (\leqslant v_3)$, decreases from $u_1 = \theta_1$ to $u_1 = \theta_2 (> v_4)$, and increases again from $u_1 = \theta_2$ to $u_1 = 1$. If $1 + a + b(1 - v_3) > \Delta > a + b(1 - v_4)$, then f_1 increases from $u_1 = 0$ to $u_1 = \omega$, where

$$\omega = \frac{(1+a+b)v_4 - (a+b)v_3 - t(1-v_4)}{1 + b(v_4 - v_3)} \quad \text{(A.35)}$$

satisfies $v_3 < \omega < v_4$, decreases from $u_1 = \omega$ to $u_1 = \theta_2 (> v_4)$, and increases again from $u_1 = \theta_2$ to

$u_1 = 1$. (Note that $\Delta > a + b(1 - v_4)$ and (A.29) imply $t > b(v_4 - v_3)$in the denominator of (A.34).) Finally, if $\Delta \leqslant a + b(1 - v_4)$ then f_1 increases monotonically from $u_1 = 0$ to $u_1 = 1$; its concavity is always downward for $0 \leqslant u_1 \leqslant v_4$ but is upward or downward on $v_4 \leqslant u_1 \leqslant 1$ according to whether $\Delta > b(1 - v_4)$ or $\Delta < b(1 - v_4)$.

Correspondingly, from (A.17), f_2 varies on $0 \leqslant u_2 \leqslant 1$ as follows. If $\Delta \geqslant 1 + a + b$, then f_2 increases from $u_2 = 0$ to $u_2 = \theta_2$ $(> v_4)$ and decreases again from $u_2 = \theta_2$ to $u_2 = 1$. If $1 + a + b > \Delta \geqslant 1 + a + b(1 - v_3)$, then f_2 decreases from $u_2 = 0$ to $u_2 = \theta_1 (\leqslant v_3)$, increases from $u_2 = \theta_1$ to $u_2 = \theta_2 (> v_4)$, and decreases again from $u_2 = \theta_2$ to $u_2 = 1$. If $1 + a + b(1 - v_3) > \Delta > a + b(1 - v_4)$, then f_2 decreases from $u_2 = 0$ to $u_2 = \omega$ (where ω is defined by (A.35) and satisfies $v_3 < \omega < v_4$), increases from $u_2 = \omega$ to $u_2 = \theta_2 (> v_4)$, and decreases again from $u_2 = \theta_2$ to $u_2 = 1$. Finally, if $\Delta \leqslant a + b(1 - v_4)$ then f_2 decreases monotonically from $u_2 = 0$ to $u_2 = 1$. Taken together with (A.15), these results imply that the maximum of f_r on $0 \leqslant u_1 \leqslant u_2 \leqslant 1$ is given by Table A4. Note that the maximizing strategies correspond to unconditional signalling if $\Delta \leqslant a + b(1 - v_4)$.

Now, if $\mathbf{v} \in S$ is an ESS then the maximum in Table A3 must occur where $u_3 = v_3$, $u_4 = v_4$; the maximum in Table A4 must occur where $u_1 = v_1$, $u_2 = v_2$; and all conditions on u must be satisfied. Let us first of all look for a strong ESS. Then $\mathbf{v} \in S$ must be the only best reply to itself. This immediately rules out the fourth and sixth rows of Table A4, where (u_1, u_2) is not unique; and although in the fifth row (0, 1) is a unique best reply, it corresponds to the bottom row of Table A3 where u_4 is not unique. Accordingly, we restrict our attention to the first three rows of Table A4. Then, for the maximum to occur at $u_2 = v_2$, each possibility requires $v_2 > v_4$. Thus the maximum at $u_4 = v_4$ in Table A3 must satisfy $v_4 < v_2$, excluding the third and sixth row of that table. Again, from (A.32), the relative magnitudes of v_3 and v_4 in the first three rows of Table A4 all imply $v_3 < v_4 < 1$, so that the maximum at $u_3 = v_3$ in Table A3 cannot satisfy $v_3 \geqslant v_4$, and hence (because $v_4 < v_2$) cannot satisfy $v_3 \geqslant v_2$; thus the fourth and fifth rows of Table 3 are excluded. The first row of the table is likewise excluded, because the maximum where $u_3 = v_3$ and $u_4 = v_4$ would have to satisfy $v_4 < v_1 \leqslant v_3$, which is impossible because $v_4 > v_3$. Only the second row of Table A3 now remains. Because the maximum at $u_3 = v_3$ must therefore satisfy $v_1 \leqslant v_3$, we cannot have $v_1 > v_3$, which excludes the third row of Table A4. But the maximum where $u_1 = v_1$, $u_2 = v_2$ in Table A4 cannot now occur where $u_1 = 0$, $u_2 = \theta_2$ because from (A.27) the second row of Table A3 would then imply $0 \leqslant a + b \leqslant b\theta_2$, which is impossible for $a > 0$. We have thus excluded the top row of Table A4, and only the second remains. We conclude that a strong ESS must correspond to the second row in each table.

Let us now write $\mathbf{v} = (I, J, K, L)$ as in Fig. 2 in the main text. Then we have shown that a strong ESS must satisfy the equations $I = \theta_1$, $J = \theta_2$, $K = \theta_3$ and $L = \theta_4$ where θ_1, θ_2 are functions of K, L defined by (A.33)–(A.34), and θ_3, θ_4 are functions of I, J defined by (A.23) and (A.26). If these equations have a unique solution, then a unique strong ESS exists. Note that with C defined by (A.18), the equations $I = \theta_1$, $J = \theta_2$, $K = \theta_3$ and $L = \theta_4$ correspond to eqns (1)–(4) of Section 4 in the main text, with $X = I$, $X = J$, $X = K$ and $X = L$, respectively, in C. The last equation can be rewritten in the form $I/(I + 1 - J) = a + b(1 - L)$. Thus, because (as is most easily seen from Fig. 4) $1 - L$ decreases with T, the proportion of threat displays that are bluffs also decreases with T as stated in Section 4.

Now $K = \theta_3$ and $L = \theta_4$ can be substituted into the equations $I = \theta_1$ and $J = \theta_2$ to yield a pair of equations for I and J. The first of these two equations has the form

$$tab(1 - J)^2 + d_1(1 - J) + d_0 = 0, \tag{A.36}$$

where

$$d_0 = (1 - a)\{(1 + t)(1 - bI + b) + a\}I \tag{A.37}$$

TABLE A4

Maximum of f_r defined by (A.6) on $0 \leqslant u_1 \leqslant u_2 \leqslant 1$

Relative magnitude of v_3, v_4	Value of (u_1, u_2) where maximum occurs	Conditions on (u_1, u_2)
$\Delta \geqslant 1 + a + b$	$(0, \theta_2)$	$u_1 \leqslant v_3$, $u_2 > v_4$
$1 + a + b > \Delta \geqslant 1 + a + b(1 - v_3)$	(θ_1, θ_2)	$u_1 \leqslant v_3$, $u_2 > v_4$
$1 + a + b(1 - v_3) > \Delta > a + b(1 - v_4)$	(ω, θ_2)	$v_3 < u_1 \leqslant v_4$, $u_2 > v_4$
$\Delta \leqslant a + b(1 - v_4)$, $v_4 > v_3$	(u_1, u_2)	$u_1 = u_2$, u_2 arbitrary
$v_3 \geqslant v_4$, $v_4 \neq 1$	(0, 1)	
$v_3 = v_4 = 1$	(u_1, u_2)	arbitrary

is a quadratic polynomial in I and

$$d_1 = -\{(a+b+at)(1-bI+b) + bt(1-a)I + a(a+b)\} \quad \text{(A.38)}$$

is linear is I. The second equation is cubic in J, and can be used in conjunction with (A.36) to express J as the quotient of cubic and quadratic polynomials in I. Substituting this expression back into (A.36) yields a sextic equation for I. Fortunately, three solutions of this equation, namely, $I=0$, $I=1+a/b$ and $I=1+(1+a)/b$ can be found by inspection. None satisfies $0<I<1$. Thus, removing the appropriate linear factors, we find that I must satisfy the cubic equation

$$b^2(1+b)(1+t)I^3 + c_2I^2 + c_1I + c_0 = 0, \quad \text{(A.39)}$$

in which the coefficients c_0, c_1 and c_2 are defined by

$$\begin{aligned} c_0 &= -a\{(a+b)(1+2a+b) + at(1+a+b)\} \\ c_1 &= (1+a+b)\{(1+t)\{1+a+(1+b)(1+t)\} \\ &\quad + 2ab\} + a(1+t)\{1+b+b(3a+2b)\} + a^2 \\ c_2 &= -b\{(1+t)\{(2+b)t + 2b^2 + (3b+2) \\ &\quad \times (1+a)\} + a(1+b)\}. \end{aligned} \quad \text{(A.40)}$$

Because $c_0<0$ and $b^2(1+b)(1+t)+c_2+c_1+c_0>0$, (A.39) always has a real solution satisfying $0<I<1$. It is not difficult (but somewhat tedious) to show that this solution is the only solution satisfying $0<I<1$; the other two solutions of (A.39) are either complex conjugates or, if they are real, satisfy $I>1$. Moreover, only one solution of the quadratic equation (A.36), the solution with a negative square root, satisfies $J>I$. Thus $\mathbf{v}=(I, J, K, L)$ is unique.

The limits of the strong ESS thresholds I, J, K and L can be found analytically both as $a\to 0$ and as $a\to 1$. As $a\to 0$ we have

$$I\to 0, \quad J\to 1, \quad K\to \frac{b}{1+b}, \quad L\to \frac{b+t}{1+b+t}. \quad \text{(A.41)}$$

As $a\to 1$ we have

$$I\to \frac{2}{1+(1+b)(1+t)+\sqrt{(1+b^2)(1+t)^2+2(1+t)(1+bt)+1}}, \; J\to 1, \; K\to 1, \; L\to 1. \quad \text{(A.42)}$$

For all intermediate values of a, the ESS thresholds are readily found numerically. A sample calculation for $b=0.4=t$ appears in Fig. 3 in the main text.

The strategy (I, J, K, L) does not, however, remain evolutionarily stable in the limit as $b\to 0$. Then $(I, J, K, L)\to \mathbf{u}^*$, where $\mathbf{u}^*$ is the four-dimensional vector defined by

$$\{1+a+t\}\mathbf{u}^* = (a^2, 1+a^2+t, \times a+a^2+at, a+a^2+t). \quad \text{(A.43)}$$

To see this, observe that the second rows of Tables A3 and A4 imply $\delta=0$ and $\Delta=1+a$ in the limit as $b\to 0$. Then θ_4 and θ_1 in (A.26) and (A.33) are indeterminate quantities, and $\mathbf{v}=(I, J, K, L)$ is instead determined from the equations $\Delta=1+a$, $J=\theta_2$, $K=\theta_3$ and $\delta=0$, which are linear, and readily yield (A.43).

Now, from (A.3), (A.12), (A.15), (A.17), (A.21), (A.30) and the appropriate modification of (A.24), we find on setting $b=0$ that $f(\mathbf{v}, \mathbf{u}^*)=f(\mathbf{u}^*, \mathbf{u}^*)$ for any

$$\mathbf{v}=(v_1, u_2^*, u_3^*, v_4) \quad \text{such that} \quad 0\leqslant v_1\leqslant u_3^*, \; u_1^*\leqslant v_4\leqslant u_2^*, \quad \text{(A.44)}$$

and that $\mathbf{v}$ then satisfies

$$f(\mathbf{u}^*, \mathbf{v}) - f(\mathbf{v}, \mathbf{v}) = -V(1+a+t)v_1(v_4-u_4^*)p_r - V(1-a)(v_1-u_1^*)(1-v_4)p_i \quad \text{(A.45)}$$

Because $f(\mathbf{u}^*, \mathbf{u}^*)=f(\mathbf{v}, \mathbf{u}^*)$, no such $\mathbf{v}$ satisfies (A.10b). On the other hand, if

$$u_1^* < v_1 \leqslant u_3^*, \quad u_4^* < v_4 \leqslant u_2^* \quad \text{(A.46)}$$

then the right-hand side of (A.45) is negative. Thus $\mathbf{v}\neq\mathbf{u}^*$ exists such that, although (A.10a) is satisfied, neither (A.10b) nor (A.10c) is satisfied. Therefore, although $\mathbf{u}^*$ defined by (A.43) is an equilibrium strategy (i.e. satisfies (A.10a)) when $b=0$, it is not an ESS.

To complete our analysis for $b>0$, we now establish that the ESS $\mathbf{u}^*=(I, J, K, L)$ is unique. We have already established that (I, J, K, L) is the only strong ESS, but from Tables A3 and A4 there are several candidates for a weak ESS. First, from the last row of Table A3 and the fifth row of Table A4, we find that

$$\mathbf{u}^* = (0, 1, (a+b)/(1+b), \lambda) \quad \text{(A.47)}$$

satisfies (A.10a) for any $\lambda\leqslant(a+b)/(1+b)$; however, (A.47) does not satisfy (A.10b), because $f(\mathbf{v}, \mathbf{u}^*)=f(\mathbf{u}^*, \mathbf{u}^*)$ for

$$\mathbf{v} = (0, 1, (a+b)/(1+b), v_4), \quad \text{(A.48)}$$

where v_4 is any number between 0 and 1. On using (A.6)–(A.7) and (A.47)–(A.48), we then find that $f(\mathbf{u}^*, \mathbf{v})-f(\mathbf{v}, \mathbf{v})=0$, so that (A.10c) does not hold. Thus $\mathbf{u}^*$ defined by (A.47) is not a weak ESS. Intuitively, *never threatening* cannot be an evolutionarily stable behavior because in equilibrium the threshold λ is irrelevant; even if the population strategy $\mathbf{v}$ satisfies $v_4\leqslant(a+b)/(1+b)$ to begin with, there is nothing to prevent v_4 from drifting to $v_4>\lambda$, in which

case *never threatening* is no longer a best reply. (In particular, there is nothing to prevent v_4 from drifting to 1, and *never threatening* cannot be a best reply to an opponent who never attacks when threatened.)

Second, from the last row of Table A4, we must investigate the possibility that there is a weak ESS of the form $\mathbf{u}^*=(v_1, v_2, 1, 1)$, where v_1 and v_2 are arbitrary. Because $a<1$, however, we see from Table A3 that this possibility requires $\theta_3=1$, which from (A.23) implies $(1-a)v_1+av_2=1$, and hence $v_1=v_2=1$. But then, from (A.28) and the first three rows of Table A3, either $v_4=(a+b)/(1+b)$ or $v_4=1-(1-a)/b$, contradicting $v_4=1$. Hence there is no such ESS.

The remaining possibility for a weak ESS is an *always-threatening* equilibrium with $v_1=v_2=\mu$, say, which corresponds to the fourth row of Table A4, and therefore satisfies $v_4>v_3$. This equilibrium cannot correspond to the first row of Table A3, because $v_1\leqslant v_3\leqslant v_2$, $v_4\leqslant v_1$ then implies $v_4<\mu\leqslant v_3$, contradicting $v_4>v_3$. For similar reasons, the equilibrium cannot correspond to either the second row of Table A3 (which would require $v_4=\mu=v_3$) or the fourth or fifth row (each of which would require $v_3>v_4$). Thus the equilibrium must correspond to either the third or sixth row of Table A3, and hence have the form

$$\mathbf{u}^*=(\mu, \mu, \zeta, (a+b)/(1+b)), \tag{A.49}$$

where

$$\zeta=\max\{\mu, 1-(1-a)/b\} \tag{A.50}$$

satisfies $\zeta<(a+b)/(1+b)$. Then $f(\mathbf{v}, \mathbf{u}^*)=f(\mathbf{u}^*, \mathbf{u}^*)$ and $f(\mathbf{u}^*, \mathbf{v})-(f(\mathbf{v}, \mathbf{v})=0$ for any

$$\mathbf{v}=(\bar{v}, \bar{v}, v_3, (a+b)/(1+b)). \tag{A.51}$$

Although $\mathbf{u}^*$ satisfies (A.10a), it fails to satisfy (A.10b)–(A.10c), and so $\mathbf{u}^*$ defined by (A.49) is not a weak ESS. Intuitively, *always threatening* cannot be an evolutionarily stable behavior because in equilibrium the threshold ζ is irrelevant; even if the population strategy $\mathbf{v}$ satisfies $v_3<(a+b)/(1+b)$ to begin with, there is nothing to prevent v_3 from drifting to $v_3\geqslant(a+b)/(1+b)$, in which case *always threatening* is no longer a best reply. (In particular, there is nothing to prevent v_3 from drifting to 1, and *always threatening* cannot be a best reply to an opponent who never attacks when not threatened.) This completes our proof that (I, J, K, L) is the only ESS.

It is interesting to note, however, from (A.32) and the fourth row of Table A4, that $\mathbf{u}^*$ defined by (A.49) is not even an equilibrium unless

$$t\leqslant\frac{a+b}{1-a}\left(\frac{a+b}{1+b}-\zeta\right). \tag{A.52}$$

Intuitively, if the handicap is sufficiently large, then the equilibrium defined by (A.49) is selected against; however, even if the handicap is small enough to satisfy (A.52), the equilibrium can still be invaded by random drift.

The ESS does not persist if payoffs are changed so that a handicap T is paid not only by animals that threaten and lose, but also by animals that threaten and win. There are two possibilities. First, suppose that a resident who threatens does not pay the handicap if the intruder flees. Then $f(\mathbf{u}, \mathbf{v})$ is modified because $\rho=V-C-T$ if $X>Y$ in (A.4), whereas $\tau=-T$ if $X>Y$ in (A.14). These modifications have no effect on f_3 of f_4; but [for g defined by (A.20)] the second integral in (A.16) becomes $-Tu_1(1-v_4)$, while the third integral in (A.17) becomes $-T(1-u_2)(1-v_4)$. These changes affect R if $v_3<v_4<1$. In that case, although (A.30) is unaltered, $-Tu_1(1-v_4)$ replaces the two terms involving t in (A.31), so that the concavity of f_1 is always downward for $u_1\geqslant v_4$ (it no longer depends on Δ). Now f_1 decreases monotonically from $u_1=0$ to $u_1=1$ if $\Delta\geqslant1+a+b$; f_1 increases monotonically from $u_1=0$ to $u_1=1$ if $\Delta\leqslant a$; and, if $a<\Delta<1+a+b$, then f_1 increases to a maximum and then decreases again, the maximum occurring where $u_1=\theta_1$ if $1+a+b>\Delta\geqslant1+a+b(1-v_3)$, where $u_1=\omega$ if $1+a+b(1-v_3)>\Delta>a+b(1-v_4)$, and where $u_1=1-(\Delta-a)/b$ if $a+b(1-v_4)\geqslant\Delta>a$. From (A.17) and the corresponding variation of f_2, we deduce that the maximum of f_r on $0\leqslant u_1\leqslant u_2\leqslant1$ must occur where $u_2=1$ if $\Delta>a$. Setting $\mathbf{u}=\mathbf{v}$ in Tables A3 and A4 as before, we find that a strong ESS must correspond to the first three rows of Table A3. But $v_2=1$ implies $\delta=a+b-1<b$, which eliminates the third row; $v_4>v_3$ eliminates the first; and $v_4>v_3$ also eliminates the second, because $v_2=1$ implies $(1+b(1-v_1))\theta_3=v_1+(a+b)(1-v_1)$ and $\theta_4=1-(1-a)/b$, so that $\theta_3<\theta_4$. There is therefore no strong ESS; and the proof that there is no weak ESS is unchanged. Because the condition $\Delta\leqslant a+b(1-v_4)$ for an *always-threatening* equilibrium is replaced by $\Delta\leqslant a$, however, the coefficient of $(a+b)/(1-b)-\zeta$ in (A.52) is reduced from $(a+b)/(1-a)$ to $a(1+b)/(1-a)$. Thus the equilibrium is selected against for smaller handicaps.

Second, suppose that a threatening resident pays the handicap regardless of whether the intruder attacks. Then, in addition to the modifications described in the

previous paragraph, V is replaced by $V-T$ in the second row of Table A1, in the third and fourth integrals of (A.6), and in the second integral of (A.8), whereas the second integral of (A.16) and the third of (A.17) become $-Tu_1$ and $-T(1-u_2)$, respectively [for g defined by (A.20)]. Now $-Tu_1$ replaces the two terms involving t in (A.31), but virtually everything else in the preceding paragraph still holds, provided only that we replace $T(1-v_4)$ by T in the expressions for Δ, θ_1 and ω. The only difference, which does not affect the existence of an ESS, is that the coefficient of $(a+b)/(1-b)-\zeta$ in (A.52) is further reduced from $a(1+b)/(1-a)$ to a; that is, the equilibrium is selected against for even smaller handicaps.

Finally, we note that the limits of the strong ESS thresholds I, J, K and L can also be found analytically both as $t\to 0$ and $t\to\infty$, which has relevance to Fig. 4. As $t\to 0$ we have $I\to\eta$, $J\to 1-(1-a)\eta/(a+b)$, $K\to(a+b)/(1+b)$ and $L\to(a+b)/(1+b)$, where η is defined by

$$\eta=\frac{2a(a+b)}{b(a+b)+(1+a)(1+b)+\sqrt{(1+a)^2(1+b)^2+2b(1-a)(1+b)(a+b)+b^2(a+b)^2}}; \tag{A.53}$$

and as $t\to\infty$ we have $I\to 0$, $J\to 1$, $K\to(a+b)/(1+b)$ and $L\to 1$. With regard to Fig. 4, note that K is not a constant, although its variation with t is negligible (it decreases slowly to a minimum and thereafter slowly rises).