THE ROYAL SOCIETY

Received 7 August 2003
Accepted 15 September 2003
Published online 19 August 2004

Evolution of spite through indirect reciprocity

Rufus A. Johnstone* **and Redouan Bshary**

Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK

How can cooperation persist in the face of a temptation to 'cheat'? Several recent papers have suggested that the answer may lie in indirect reciprocity. Altruistic individuals may benefit by eliciting altruism from observers, rather than (as in direct reciprocity) from the recipient of the aid they provide. Here, we point out that indirect reciprocity need not always favour cooperation; by contrast, it may support spiteful behaviour, which is costly for the both actor and recipient. Existing theory suggests spite is unlikely to persist, but we demonstrate that it may do so when spiteful individuals are less likely to incur aggression from observers (a negative form of indirect reciprocity).

Keywords: spite; cooperation; image scoring; reputation; eavesdropping; indirect reciprocity

1. INTRODUCTION

Humans often help others in situations where direct reciprocation by the beneficiary of their altruistic act is highly unlikely (Alexander 1987). Recently, game theoretical models have demonstrated that such altruism can nevertheless prove evolutionarily stable, provided one allows for 'image scoring' by observers (Nowak & Sigmund 1998a,b; Lotem *et al.* 1999, 2003; Leimar & Hammerstein 2001). In these models, altruistic individuals improve their 'reputation' or 'image' (while selfish individuals impair theirs), increasing (or, in the case of selfishness, decreasing) their own probability of receiving help from others when needed. Most recently, Lotem *et al.* (2003) have linked indirect reciprocity with Zahavi's (1975) handicap principle: altruism may lead to an improved 'image' because it serves as a costly (and therefore honest) signal of quality, which may prove attractive to potential mates as well as encouraging altruistic responses from others. An experiment with students provides evidence supporting these ideas, showing that in humans, altruistic individuals may receive a higher pay-off than selfish individuals as a result of indirect reciprocity (Wedekind & Milinski 2000).

A basic requirement for image scoring is that interactions take place within a 'communication network', in which bystanders may readily 'eavesdrop' on encounters between others (McGregor 1993; McGregor *et al.* 2000). There is now growing evidence that eavesdropping does occur in a wide range of taxa, from primates to crabs (Cheney & Seyfarth 1990; Naguib & Todt 1997; Oliveira *et al.* 1998; Naguib *et al.* 1999; Peake *et al.* 2001; Earley & Dugatkin 2002), and some indications that animals may consequently adjust their behaviour in response to the presence of bystanders (Doutrelant *et al.* 2001; Bshary 2002). Intriguingly, however, most of these empirical studies have focused, not on altruistic or cooperative interactions, but on competitive or agonistic encounters. In this context, 'image scoring' is likely to promote greater levels of aggression rather than cooperation, because it means that a victorious individual will not only obtain the contested resource, but will also improve its reputation, reducing its chance of incurring aggression from others (Johnstone 2001).

Here, we further explore the consequences of eavesdropping in antagonistic contexts, focusing on the possibility of spite. We demonstrate that just as image scoring can support costly altruistic acts that would otherwise prove unstable, so too can it support costly spiteful acts that otherwise would not persist. Just as altruism may be favoured because it encourages observers to act altruistically towards the focal individual, so spite may be favoured because it discourages observers from acting aggressively towards the perpetrator (a form of 'negative indirect reciprocity').

2. A MODEL OF NEGATIVE INDIRECT RECIPROCITY

Consider a population in which individuals engage in random, pairwise encounters over a large number of 'rounds'. In each round, one member of every pair, chosen at random and referred to (for that round) as an active player, must decide whether or not to attack the other (who is referred to, for that round, as a passive player).

Individuals vary in their strength or competitive ability S, which is distributed according to the probability density function $f(S) = F'(S)$; for simplicity, we will describe them in terms of their relative competitive ability $s = F(S)$, so that an individual for whom $s = 0.25$ is stronger than 25% of the population, one for whom $s = 0.5$ is stronger that 50% of the population, and so on. Assuming that no two individuals have precisely the same competitive ability, s is thus evenly distributed between zero and unity (regardless of the form of $f(S)$).

An individual that chooses to attack inflicts an additive fitness cost c (> 0) on the target. If the victim is stronger than the attacker, however, then the latter also incurs an additive fitness cost of d (> 0); if the victim is weaker than the attacker, then the latter obtains an additive fitness benefit of b (a negative value of b implies that even victory entails a net cost for the attacker). For simplicity, we will scale these various costs and benefits so that $c = 1$ (d and b may therefore be interpreted as the cost or benefit of attacking stronger and weaker opponents, relative to the cost of being attacked).

* Author for correspondence (raj1003@hermes.cam.ac.uk).

Individuals may base their decision about whether or not to attack on their opponent's most recent action as an active player (i.e. on whether or not their opponent itself chose to attack on the last occasion when it had the opportunity to do so). Because the expected cost that the active player incurs if it attacks is a non-increasing function of its relative competitive ability, we assume that it will choose to attack only if this ability exceeds some threshold value. Given the option of taking into account an opponent's most recent action, a strategy can thus be defined by two threshold values, t_1 and t_2 ($< t_1$), the critical competitive abilities above which a player will attack an opponent that chose (at its last opportunity) to attack or not to attack, respectively. We wish to determine the evolutionarily stable strategy (ESS) or strategies (t_1^*, t_2^*), which maximizes an individual's long-term average pay-off per round (our measure of fitness), given that it is adopted by other members of the population.

(a) *Solving the model*

Consider a population of individuals that adopt the strategy (t_1, t_2). The proportion of players in this population that, at the end of round n, chose to attack when they last had the opportunity to do so, is denoted $a(n)$. This proportion changes from one round to the next according to the following difference equation

$$a(n+1) = \frac{1}{2}a(n) + \frac{1}{2}[a(n)(1-t_1) + (1-a(n))(1-t_2)],$$

converging to a, given by

$$a = a(1-t_1) + (1-a)(1-t_2),$$

$$a = \frac{1-t_2}{1+t_1-t_2}.$$

In this population, the relative competitive ability of an individual who chose not to attack when it last had the opportunity to do so must fall between 0 and t_1 (because those of greater strength always choose to attack). The probability that the relative competitive ability of such an individual falls below t_2, converges to p_0 given by

$$p_0 = \frac{t_2}{1-a} = \frac{t_2}{t_1}(1+t_1-t_2),$$

while the probability that the competitive ability of such an individual falls above t_2 converges to p_1 given by

$$p_1 = 1 - p_0 = \frac{(1-t_2)}{t_1}(t_1-t_2).$$

We can use the above results to determine the long-term average pay-off per round (our measure of fitness) to a mutant individual of relative competitive ability, s, that never attacks, denoted $W_n(s)$. Equally, we can also work out the pay-off to such a mutant that attacks only if its opponent failed to attack at its last opportunity, $W_c(s)$, and to a mutant that always attacks, $W_a(s)$:

$$W_n(s) = 0 - \frac{1}{2}(1-t_2),$$

$$W_c(s) = \begin{cases} \frac{1}{2}(1-a)\left(\frac{p_0 s}{t_2}b - \left(1-\frac{p_0 s}{t_2}\right)d\right) \\ \qquad -\frac{1}{2}(a(1-t_2)+(1-a)(1-t_1)), \text{for } s < t_2, \\ \frac{1}{2}(1-a)\left(\left(1-\frac{p_1(t_1-s)}{t_1-t_2}\right)b - \frac{p_1(t_1-s)}{t_1-t_2}d\right) \\ \qquad -\frac{1}{2}(a(1-t_2)+(1-a)(1-t_1)), \text{for } t_2 < s < t_1, \\ \frac{1}{2}(1-a)b - \frac{1}{2}(a(1-t_2)+(1-a)(1-t_1)), \text{for } t_1 < s, \end{cases}$$

$$W_a(s) = \frac{1}{2}(sb-(1-s)d) - \frac{1}{2}(1-t_1).$$

In each of the above equations, the first term on the right hand side represents the expected cost or benefit as a result of aggression by the focal player (the probability that the focal player is active, multiplied by the probability that it then chooses to attack, multiplied by the expected pay-off from attacking). The second term represents the expected cost as a result of aggression by others directed against the focal player (the probability that the focal player is passive, multiplied by the probability that its opponent then chooses to attack, multiplied by the cost of being attacked).

Based on the formulae given above, the expected pay-off (averaging over all possible relative competitive abilities) to a mutant that adopts the strategy (t_1', t_2'), denoted $\bar{W}(t_1', t_2')$, is given by

$$\bar{W}(t_1', t_2') = \int_{t_2'}^{0} W_n(s)\,\mathrm{d}s + \int_{t_1'}^{t_2'} W_c(s)\,\mathrm{d}s + \int_{t_1'}^{1} W_a(s)\,\mathrm{d}s. \quad (2.1)$$

A necessary condition for the evolutionary stability of a strategy (t_1^*, t_2^*) is that

$$\bar{W}(t_1', t_2') \leqslant \bar{W}(t_1^*, t_2^*) \text{ for all } (t_1', t_2')(t_1^*, t_2^*),$$

assuming that typical members of the population adopt the strategy (t_1^*, t_2^*). This condition implies locally that (for $0 < t_2^* < t_1^* < 1$)

$$\frac{\partial \bar{W}(t_1', t_2')}{\partial t_1'} = \frac{\partial \bar{W}(t_1', t_2')}{\partial t_2'} = 0 \text{ for } t_1' = t_1^*, t_2' = t_2^*, \quad (2.2)$$

which, together with equation (2.1), can be used to identify potential equilibria (see Appendix A for details).

We can also examine the convergence stability of candidate equilibria under the adaptive dynamics described by Hofbauer & Sigmund (1998). Thus, we assume that evolutionary change in the strategy (t_1, t_2) adopted by a population depends on the slope of mutant fitness with respect to each component of the strategy, i.e.

$$\dot{t}_1 = \frac{\partial \bar{W}(t_1', t_2')}{\partial t_1'}, \quad \dot{t}_2 = \frac{\partial \bar{W}(t_1', t_2')}{\partial t_2'},$$

where both derivatives are evaluated at $(t_1', t_2') = (t_1, t_2)$. Ultimately, we are interested in equilibria that are both evolutionarily stable (i.e. immune to invasion) and convergence stable (implying that populations adopting a strategy that deviates slightly from the ESS will tend towards it under the influence of selection).

Having identified an equilibrium that features a non-zero frequency of aggression, we can ask whether (or how often) this aggression proves spiteful (i.e. entails an immediate

cost). Clearly, if $b < 0$, aggression must always be spiteful, because even victory then entails a net cost for an attacker. If $b > 0$, however, the sign of the immediate mean pay-off to attack depends upon the attacker's competitive ability (and upon the last action of the victim, because this conveys information about the latter's competitive ability). An individual of relative compeitive ability s, who attacks an opponent who refrained from attacking at the last opportunity it had to do so, obtains an immediate mean pay-off of $A_0(s)$, given by

$$A_0(s) = \begin{cases} b\left(p_0\frac{s}{t_2}\right) - d\left(1 - p_0\frac{s}{t_2}\right) \text{ for } s \leqslant t_2, \\ b\left(p_0 + p_1\left(\frac{s-t_2}{t_1-t_2}\right)\right) - d\left(1 - p_0 - p_1\left(\frac{s-t_2}{t_1-t_2}\right)\right), \\ \quad \text{for } t_2 < s \leqslant t_1 \\ b \text{ for } t_1 < s, \end{cases}$$

while the immediate mean pay-off from attacking an opponent who also chose to attack at the last opportunity it had to do so, denoted $A_1(s)$, is given by

$$A_1(s) = \begin{cases} -d \text{ for } s \leqslant t_2, \\ bq_0\left(\frac{s-t_2}{t_1-t_2}\right) - d\left(1 - q_0\left(\frac{s-t_2}{t_1-t_2}\right)\right) \text{ for } t_2 < s \leqslant t_1, \\ bq_1\left(\frac{s-t_1}{1-t_1}\right) - d\left(1 - q_1\left(\frac{s-t_1}{1-t_1}\right)\right) \text{ for } t_1 < s, \end{cases}$$

where q_0 denotes the probability that the competitive ability of an opponent who chose to attack at the last opportunity falls between t_2 and t_1, and q_1 the probability that it falls aove t_1. These values are given by

$$q_1 = \frac{1-t_1}{a} = \frac{1-t_1}{1-t_2}(1 + t_1 - t_2), q_0 = 1 - q_1.$$

We can use the above expressions to determine the critical levels of competitive ability, s_0 and s_1, below which any decision to attack a non-aggressive or an aggressive opponent entails an immediate mean cost. These values then allow us to determine what fraction of attacks are spiteful in nature. For instance (assuming $b > 0$), if $t_2 < s_0 < t_1 < s_1$, as turns out to be the case at any equilibrium featuring aggression (see below), the proportion of attacks that are spiteful is given by

$$\frac{a(s_1 - t_1) + (1-a)(s_0 - t_2)}{a(1-t_1) + (1-a)(1-t_2)} = (s_1 - t_1) + \frac{(1-a)}{a}(s_0 - t_2).$$

3. RESULTS

As detailed in Appendix A, when $b > 0$ (so that victory yields an immediate positive pay-off), the model yields a single strategy that is both evolutionarily and convergently stable. This strategy features values of t_1^* and t_2^* that are both less than unity (precise formulae are given in Appendix A), so that it leads to a positive frequency of attack, as illustrated in figure 1. At this equlibrium, a non-zero fraction of aggressive acts are spiteful, in the sense that their immediate mean pay-off is negative. The agressors involved nevertheless choose to attack because this immediate cost is outweighed by the benefits of an aggressive image: a player that is seen to attack others is less likely to be attacked itself, providing a reputation benefit that helps to maintain aggression at higher levels than would otherwise occur. In other words, spitefully attacking others can prove stable simply because it reduces the chances of being attacked oneself.

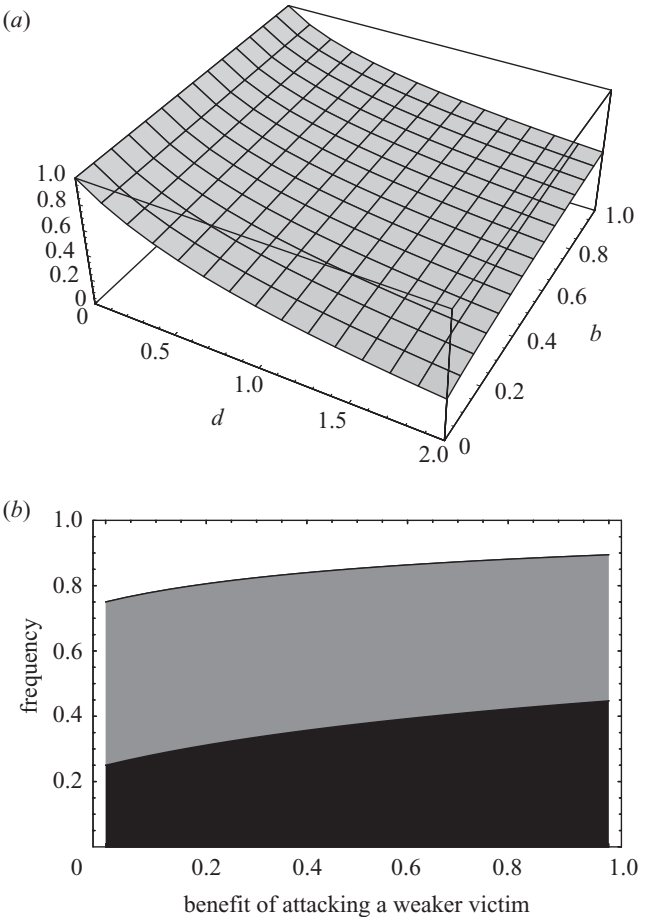

Figure 1. (*a*) The frequency of attack at the equilibrium described in the main text as a function of d, the cost of attacking a stronger victim (relative to the cost of being attacked), and b, the benefit of attacking a weaker victim (relative to the cost of being attacked). (*b*) The proportion of individuals at the aggressive equilibrium who never attack (white area), who attack conditionally (if their opponent was not seen to attack itself when it last had the opportunity to do so) (grey area), and who always attack (black area), as a function of b (assuming that $d = 1$).

The frequency of spite at equilibrium (i.e. the proportion of agonistic encounters in which the aggressor chooses to attack even though doing so yields an immediate mean pay-off less than zero) varies in relation to the parameters b and d, as illustrated in figure 2. If the benefits of victory are low, the majority of aggression may qualify as spite. Nevertheless, there is always a non-zero frequency of straightforward aggression, in which the immediate mean pay-off to attack is positive. We therefore refer to this outcome as one of 'occasional spite'.

When $0 > b > -1/(4d)$ (so that attacking always entails an immediate cost, even if the attacker wins), the model yields two convergently stable ESSs: the one discussed above (and detailed in Appendix A), and the alternative strategy $(1, 1)$ which specifies a complete absence of aggression; between these two lies an ESS that is not convergently stable. Which of the two stable endpoints a population attains depends on the initial conditions, as

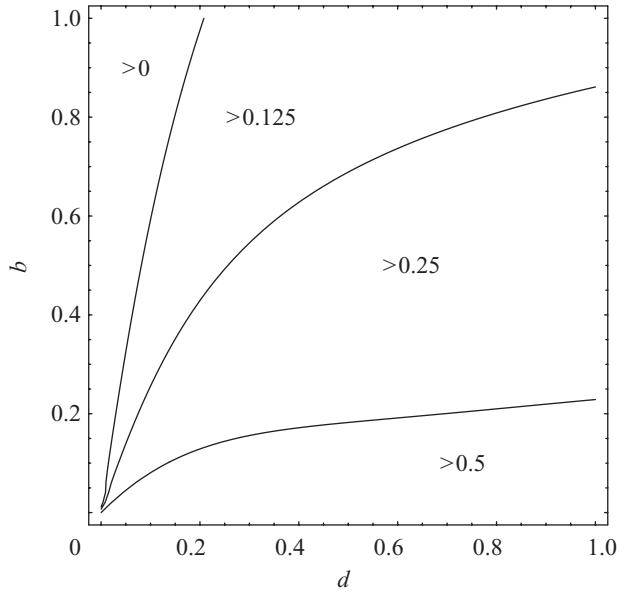

Figure 2. Contour plot of the frequency of spite (i.e. the proportion of agonistic acts in which the aggressor attacks despite the fact that doing so yields an immediate mean pay-off that is negative) at the equilibrium described in the main text, as a function of d, the cost of attacking a stronger victim (relative to the cost of being attacked), and b, the benefit of attacking a weaker victim (relative to the cost of being attacked).

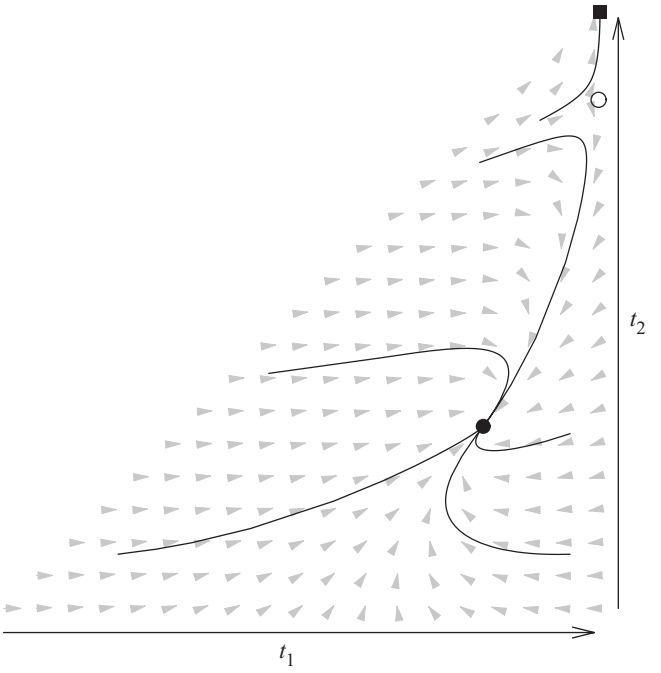

Figure 3. Evolutionary trajectories, in the permissible triangle of (t_1, t_2)-values for which $t_2 \leqslant t_1$, under the adaptive dynamics. Results are for the illustrative case in which $b = -1/8$ (implying that even victory entails an immediate fitness cost for an attacker) and $d = 1$. Under these circumstances, as discussed in the text, the model yields two convergently stable ESSs: the aggressive ESS defined in Appendix A (marked on the figure as a filled circle), which under these circumstances represents a purely spiteful outcome, and the alternative ESS $(1, 1)$ (marked as a filled square), which leads to a complete absence of aggression. The open circle marks the convergently unstable ESS that lies between the two stable endpoints. The pale arrows desribe the vector field of the adaptive dynamics, while the solid curves represent sample trajectories. Clearly, the stable endpoint for an evolving population depends on the starting conditions.

illustrated in figure 3. In a population that settles at the aggressive equilibrium, it is clear that all aggression must be spiteful, because a negative value of b implies that attack always entails an immediate cost. We therefore refer to this as outcome as 'universal spite'.

Finally, when $-1/(4d) > b$ (so that even victory entails a substantial fitness cost), the non-aggressive strategy $(1, 1)$ is the only ESS (and is convergently stable).

The reason why eavesdropping is stable in our model, i.e. the reason why (at an aggressive equilibrium) it pays to adjust one's threshold for attack according to the opponent's previous actions, is that aggression is a reliable signal of strength. It does not pay weaker individuals to act aggressively because they are more likely to incur substantial costs as a result; hence, those opponents that are seen to attack are likely to be of greater competitive ability, and are less likely to become the target of aggression.

4. DISCUSSION

The most influential adaptive accounts of spiteful behaviour, developed by Hamilton (1970) and Wilson (1975), are based on inclusive fitness arguments. Despite a few possible examples (Foster *et al.* 2001), however, these ideas are widely thought to be of limited applicability (Hamilton 1970; Foster *et al.* 2001). Here, by contrast, we have shown that spite can evolve readily in agonistic contexts.

Suppose that there are immediate benefits (however small) to be gained by attacking and defeating a rival (in terms of the model, $b > 0$). Selection will then favour some frequency of non-spiteful aggression. Where this aggression carries potential costs, weaker individuals are less likely to attack than are stronger competitors. As a result, it pays observers to 'eavesdrop' on encounters between others, and subsequently to exercise greater caution in attacking more aggressive opponents. In turn, this favours greater levels of aggression, as a means of deterring subsequent attack (see Johnstone 2001). The net result, as our model reveals, is 'occasional spite': an equilibrium at which levels of aggression are elevated (perhaps substantially) beyond the point where attack yields a positive mean pay-off. Individuals at this equilibrium will sometimes attack opponents likely to defeat and inflict costs on them, simply in order to gain reputation benefits. Indeed, if the immediate benefits of victory are small, the model demonstrates that such spiteful acts may constitute the majority of aggressive encounters. The costly nature of this spiteful behaviour in the short term is outweighed by long-term advantages because of negative indirect reciprocity.

Even when victory entails a net fitness cost for an attacker ($b < 0$), aggression can nevertheless persist. The model shows that 'universal spite' can be stable under these circumstances, yielding an equilibrium at which individuals engage in spiteful aggression simply in order to deter future spite directed at themselves. We suggest, however, that this outcome is unlikely: the same circumstances that

yield a universally spiteful equilibrium also permit a non-aggressive one, at which all individuals do better. Because mean fitness is lower at the spiteful equilibrium, it seems unlikely that populations which have become trapped at this 'paradoxical' ESS will persist (and, in addition, it is not clear what selective pressures could drive the transition from a non-aggressive to a universally spiteful equilibrium). It seems more likely that spite will arise in conjunction with straightforward aggression, under the circumstances described above.

By some definitions (e.g. Pierotti 1980), the behaviour we focus on (attacking an opponent even when doing so yields an immediate mean pay-off that is negative) might not be described as truly spiteful: if the immediate costs of attack are outweighed by the long-term benefits of establishing an aggressive image or reputation, then aggression could be said to prove self-serving rather than spiteful. By the same logic, however, altruism based on positive reciprocity (direct or indirect) should not qualify as true altruism: if the immediate costs of altruistic actions are outweighed by the benefits of establishing a positive image, then once again such behaviour is simply self-serving. The point we wish to emphasise (as do models of positive indirect reciprocity) is that immediate costly actions can be favoured through their long-term impact on the image or reputation of the actor. We have focused here on reputation benefits in the form of a reduced risk of attack, but just as Lotem *et al*. (2003) argue that altruism may prove attractive to potential mates as well as encouraging altruism from others, we too could argue that spite may bring mating benefits as well as discouraging aggression, because it serves as a signal of strength.

Finally, if spite (in the sense explained above) may so readily evolve in the context of agonistic encounters, why have more instances not been reported? Although it is accepted that humans frequently inflict costs on others at their own expense (Wilson 1975; and see Fehr & Gächter (2002) for an experimental demonstration), reports of spiteful behaviour in other animals are rare and contentious (Pierotti 1980; FitzGerald 1992; Gadagkar 1993; though see Foster *et al*. 2001). We suggest, however, that spiteful aggression has often been overlooked. Seeking an adaptive explanation of aggressive behaviour, ethologists and behavioural ecologists have naturally focused on those occasions when attack yields immediate benefits (a contested resource item, territory or mate). Even if victory is rare, attackers often lose, and sometimes incur substantial injuries as a result, it is generally assumed (without quantitative calculation) that the expected benefits of victory must always outweigh these costs: otherwise why would agressive behaviour have evolved? However, in many instances we suggest the immediate expected pay-off to attack may be negative. Instead, aggression may be favoured in part by indirect reciprocity, proving spiteful in the short term. Without a proper calculation of the balance between expected costs and benefits of attack (something that is rarely attempted), one cannot dismiss the possibility that aggression is actually spiteful. We thus predict that future studies may well yield evidence for frequent occurrences of spite and indirect negative reciprocity in nature.

We thank Tim Clutton-Brock, Andrew Russell and two anonymous referees for helpful comments and discussion. R.B. was funded by a Marie Curie Fellowship from the EU.

APPENDIX A

When $b > -1/(4d)$, equations (2.1) and (2.2) yield the solution

$$t_1^* = \frac{(1-A)(1+b^2-b(1-d))+2d(1+(1+b)d+d^2)}{2(1+d+d^2+d^3-b(1-2d-d^2)+b^2(1-d)-b^3)},$$

$$t_2^* = \frac{1+d+d^2+d^3+d(3b+2bd+d^2)-A(1+(1+b)d+d^2)}{2(1+d+d^2+d^3-b(1-2d-d^2)+b^2(1-d)-b^3)},$$

where

$$A = 1+4bd.$$

This strategy is both evolutionarily stable and, as may be shown by numerical evaluation of the eigenvalues of the Jacobian matrix for the adaptive dynamics at (t_1^*, t_2^*), is also convergence stable. When adopted by a population, the strategy leads to a positive frequency of aggression equal to

$$\frac{(1+A)(1+(1+b)d+d^2)-2b+2b^2(1-d)-2b^3}{(1+d-b)(2+(1+A)d+2d^2+b(A+4d-1)+2b^2)}.$$

For $b < 0$, the strategy $(1, 1)$, which implies that individuals should never attack, is also evolutionarily and convergently stable. For $0 > b > -1/(4d)$, the model thus yields two alternative, stable endpoints; in this case, a third ESS that is not convergently stable lies between the two stable equilibria, as illustrated in figure 3.

REFERENCES

Alexander, R. D. 1987 *The biology of moral systems*. New York: Aldine de Gruiter.

Bshary, R. 2002 Biting cleaner fish use altruism to deceive image scoring reef fish. *Proc. R. Soc. Lond.* B **269**, 2087–2093. (doi:10.1098/rspb.2002.2084.)

Cheney, D. L. & Seyfarth, R. M. 1990 *How monkeys see the world*. University of Chicago Press.

Doutrelant, C., McGregor, P. K. & Oliveira, R. F. 2001 The effect of an audience on intrasexual communication in male Siamese fighting fish, *Betta splendens*. *Behav. Ecol.* **12**, 283–286.

Earley, R. L. & Dugatkin, L. A. 2002 Eavesdropping on visual cues in green swordtail (*Xiphophorus helleri*) fights: a case for networking. *Proc. R. Soc. Lond.* B **269**, 943–952. (doi:10.1098/rspb.2002.1973.)

Fehr, E. & Gächter, S. 2002 Altruistic punishment in humans. *Nature* **415**, 137–140.

FitzGerald, G. J. 1992 Egg cannibalism by sticklebacks: spite or selfishness? *Behav. Ecol. Sociobiol.* **30**, 201–206.

Foster, K. R., Wenseleers, T. & Ratnieks, F. L. W. 2001 Spite: Hamilton's unproven theory. *Ann. Zoo. Fenn.* **38**, 229–238.

Gadagkar, R. 1993 Can animals be spiteful? *Trends Ecol. Evol.* **8**, 232–234.

Hamilton, W. D. 1970 Selfish and spiteful behaviour in an evolutionary model. *Nature* **228**, 1218–1220.

Hofbauer, J. & Sigmund, K. 1998 *Evolutionary games and population dynamics*. Cambridge, UK: Cambridge University Press.

Johnstone, R. A. 2001 Eavesdropping and animal conflict. *Proc. Natl Acad. Sci. USA* **98**, 9177–9180.

Leimar, O. & Hammerstein, P. 2001 Evolution of cooperation through indirect reciprocity. *Proc. R. Soc. Lond.* B **268**, 745–753. (doi:10.1098/rspb.2000.1573.)

Lotem, A., Fishman, M. A. & Stone, L. 1999 Evolution of cooperation between individuals. *Nature* **400**, 226–227.

Lotem, A., Fishman, M. A. & Stone, L. 2003 From reciprocity to unconditional altruism through signaling benefits. *Proc. R. Soc. Lond.* B **270**, 199–205. (doi:10.1098/rspb.2002.2225.)

McGregor, P. K. 1993 Signalling in territorial systems: a context for individual identification, ranging and eavesdropping. *Phil. Trans. R. Soc. Lond.* B **340**, 237–244.

McGregor, P.K., Otter, K.A., & Peake, T.M. 2000 Communication networks: receiver and signaller perspectives. In *Animal signals: signalling and signal design in animal communication* (ed.), pp. 329–340. Trondheim, Norway: Tapir Academic Press.

Naguib, M. & Todt, D. 1997 Effects of dyadic interactions of other conspecific receivers in nightingales. *Anim. Behav.* **54**, 1535–1543.

Naguib, M., Fichtel, C. & Todt, D. 1999 Nightingales respond more strongly to vocal leaders of simulated dyadic interactions. *Proc. R. Soc. Lond.* B **266**, 537–542. (doi:10.1098/rspb.1999.0669.)

Nowak, M. A. & Sigmund, K. 1998*a* Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577.

Nowak, M. A. & Sigmund, K. 1998*b* The dynamics of indirect reciprocity. *J. Theor. Biol.* **194**, 561–574.

Oliveira, R. F., McGregor, P. K. & Latruffe, C. 1998 Know thine enemy: fighting fish gather information from observing conspecific interactions. *Proc. R. Soc. Lond.* B **265**, 1045–1049. (doi:10.1098/rspb.1998.0397.)

Peake, T. M., Terry, A. M. R., McGregor, P. K. & Dabelsteen, T. 2001 Male great tits eavesdrop on simulated male-to-male vocal interactions. *Proc. R. Soc. Lond.* B **268**, 1183–1187. (doi:10.1098/rspb.2001.1648.)

Pierotti, R. 1980 Spite and altruism in gulls. *Am. Nat.* **115**, 290–300.

Wedekind, C. & Milinski, M. 2000 Cooperation through image scoring in humans. *Science* **288**, 850–852.

Wilson, E. O. 1975 *Sociobiology*. Cambridge, MA: Harvard University Press.

Zahavi, A. 1975 Mate selection-A selection for a handicap. *J. Theor. Biol.* **53**, 205–213.