Vol. 150, Supplement The American Naturalist July 1997

REPRODUCTIVE BRIBING AND POLICING AS EVOLUTIONARY MECHANISMS FOR THE SUPPRESSION OF WITHIN-GROUP SELFISHNESS

H. KERN REEVE[1,*] AND LAURENT KELLER[2,†]

[1]Section of Neurobiology and Behavior, Cornell University, Ithaca, New York 14853-2702; [2]Institut de Zoologie et d'Ecologie Animale, Université de Lausanne, Bâtiment de Biologie, 1015 Lausanne, Switzerland, and Zoologisches Institut, Bern University, Ethologische Station Hasli, 3032 Hinterkappelen, Switzerland

Abstract.—We show that a new, simple, and robust general mechanism for the social suppression of within-group selfishness follows from Hamilton's rule applied in a multilevel selection approach to asymmetrical, two-person groups: If it pays a group member to behave selfishly (i.e., increase its share of the group's reproduction, at the expense of group productivity), then its partner will virtually always be favored to provide a reproductive "bribe" sufficient to remove the incentive for the selfish behavior. The magnitude of the bribe will vary directly with the number of offspring (or other close kin) potentially gained by the selfish individual and inversely with both the relatedness r between the interactants and the loss in group productivity because of selfishness. This bribe principle greatly extends the scope for cooperation within groups. Reproductive bribing is more likely to be favored over social policing for dominants rather than subordinates and as intragroup relatedness increases. Finally, analysis of the difference between the group optimum for an individual's behavior and the individual's inclusive fitness optimum reveals a paradoxical feedback loop by which bribing and policing, while nullifying particular selfish acts, automatically widen the separation of individual and group optima for other behaviors (i.e., resolution of one conflict intensifies others).

Evolutionary conflicts of interest can exist among individuals in genetically heterogeneous animal societies (Hamilton 1964; West Eberhard 1981). A multilevel selection approach like those described in the other articles of this volume provides a natural framework for analyzing the evolutionary outcomes of such conflicts. We shall employ such an approach to illuminate social processes that limit the expression of selfishness within groups.

First, however, it is important to stress a point still overlooked in many discussions of multilevel selection theory (in both scientific and especially the semipopular literature). Multilevel selection approaches as exemplified by trait–group selection models (e.g., Wilson and Sober 1994) are not fundamentally different from "classical" individual selection approaches as represented by generalized inclusive fitness models (e.g., Queller 1992). It is possible in every instance to translate from one approach to the other without disturbing the

* E-mail: hkr1@cornell.edu.
† E-mail: Lkeller@ulys.unil.ch.

Am. Nat. 1997. Vol. 150, pp. S42–S58.

mathematics describing the net result of selection (Dugatkin and Reeve 1994). Multilevel selection approaches simply partition selection into different components (often into more components) than do classical individual selection models, and which approach is more useful depends on the theoretical aim.

Our aim is to examine evolutionary, socially mediated mechanisms that restrain within-group selfishness, so the multilevel selection strategy of partitioning fitness into within- and between-group components is particularly useful. The logical relationship of such a partitioning to classical or ''broad sense'' individual fitness (i.e., offspring number) is simple: an individual's offspring number is equal to the product of p, the organism's fraction of the group reproduction (also called the within-group component of fitness), and k, the total group output (also called the between-group component of fitness). The product kp therefore can be substituted for offspring number in ordinary individual selection models to analyze the selective fate of behaviors that affect k and p in various ways.

Our primary questions can be sharply defined in this multilevel selection framework. First, what social mechanisms will inhibit the expression of selfish, destructive acts, in which a selfish act is defined as a behavior that causes an increase in the personal share of reproduction p and a destructive act is defined as a behavior that reduces the group output k? The social inhibition of selfish, destructive acts should increase the degree to which group members appear to be maximizing group reproductive output. Second, under what circumstances will the use of one kind of inhibitory mechanism be favored over another?

Selfish, destructive behavior manifesting intragroup conflicts may be suppressed because such behavior is too harmful to kin (causing self-restraint or self-policing) (Hamilton 1964; Ratnieks and Reeve 1992) or will be made unprofitable by other group members (social suppression) (Ratnieks and Reeve 1992). The most frequently discussed form of social suppression is social policing (e.g., Trivers 1971; Clutton-Brock and Parker 1994; Frank 1995), which we subdivide into punishment and sabotage. In punishment, social policers physically interfere with or otherwise directly impose reproductive costs on selfishly behaving group members. For example, in the eusocial naked mole rat (*Heterocephalus glaber*), the breeding female aggressively shoves workers that are relatively lazy, causing the latter to work harder (Reeve and Sherman 1991; Reeve 1992). In sabotage, social policers reduce the profitability of selfishness more indirectly by undermining its benefits. For example, honey bee (*Apis mellifera*) workers are more likely to remove the male-destined eggs produced by other workers than the more highly related male-destined eggs produced by the mother queen (Ratnieks and Visscher 1989). This sabotage apparently helps prevent widespread attempted male production by workers in queen-right colonies (Ratnieks 1993).

We shall show that a new, simple, and robust general mechanism for the social suppression of within-group selfishness follows from Hamilton's rule combined with a multilevel selection approach to asymmetrical, two-person groups: if it pays a group member to engage in a selfish, destructive act (i.e., increase

its share of the reproduction at the expense of group productivity), then its partner will virtually always be favored to provide a reproductive "bribe" (i.e., to yield some reproduction) sufficient to remove the incentive for the selfish behavior. We shall demonstrate that this bribe principle greatly extends the scope for cooperation and provides a previously unrecognized theoretical linkage between the distribution of reproduction and potential for selfishness within groups. After exploring the conditions under which reproductive bribing will be favored over policing, we shall develop a general expression for the difference between the individual and group optimum for a behavior to study the consequences of social suppression of a particular selfish, destructive act for the social inhibition of other such acts.

REPRODUCTIVE BRIBING: A SIMPLE MODEL

We assume that two individuals forming a simple group partition the total reproduction, with a fraction p of the pair's total reproductive output going to the potentially selfish individual (henceforth called the recipient) and $1 - p$ to the potential suppressor (henceforth called the suppressor). Our model will allow for the possibility that either of the group members can be in the recipient role. If $1 - p > p$, then the recipient can be considered the subordinate and the potential suppressor can be considered the dominant individual; the reverse holds true otherwise (thus, we are considering an asymmetrical, two-person game). The pair has a total reproductive output k, relative to an output equal to 1.0 for a single individual if the other leaves the group ($k > 1.0$). Now suppose that the recipient is favored to engage in a selfish, destructive act that would increase its share of reproduction from p to $p + z$ at the expense of the pair's total output, which would decrease from k to $k - c$, with $c > 0$ (the loss in group productivity c must not be large enough to destabilize the group, because, by assumption, belonging to the group is favored over not belonging to the group).

What action is favored either for the recipient or the suppressor will be governed by Hamilton's rule; that is, action i will be favored over action j if

$$(P_i - P_j) + r(K_i - K_j) > 0, \tag{1}$$

where r is the coefficient of relatedness (assumed symmetrical) between the interactants, P_i (or P_j) is the personal reproduction (offspring number) associated with action i (or j), and K_i (or K_j) is the other party's reproduction if action i (or j) is performed (Hamilton 1964; Grafen 1984). Use of the simple additive version of Hamilton's rule is appropriate because dominants and subordinates can be viewed as being in different contexts—the rule appears to work especially well if behaviors are highly context dependent (conditional) (Parker 1989).

First, we ask under what conditions the selfish act will be initially favored. If the selfish act is performed, then the recipient's personal reproduction is $(p + z)(k - c)$ and the suppressor's personal reproduction is $(1 - p - z)(k - c)$, so by inequality (1), the selfish act will be favored only if

$$(p + z)(k - c) - pk + r[(1 - p - z)(k - c) - (1 - p)k] > 0. \quad (2)$$

The expression on the left side of inequality (2) declines as p increases (its derivative with respect to p is $-c[1 - r]$), so as p increases, the temptation to act selfishly decreases. For sufficiently large p, selfishness may not be favored.

What minimum fraction e of the pair's reproduction would be sufficient to make a selfish act unfavorable for the recipient if it received this fraction in addition to its initial fraction p? The recipient's personal reproduction is $(p + z)(k - c)$ and the suppressor's personal reproduction is $(1 - p - z)(k - c)$ if the selfish act is performed, and the recipient's personal reproduction is $(p + e)(k)$ and the suppressor's is $(1 - p - e)(k)$ if the bribe (e) is given and the selfish act is thus not performed. Substituting the above expressions into inequality (1), converting the inequality into an equality, and solving for e yield the magnitude of a bribe at which selfishness begins to become unfavorable for the subordinate:

$$e = \frac{[(p + z)(k - c) - kp](1 - r) - rc}{k(1 - r)}. \quad (3)$$

A recipient would do as well to take such a bribe and refrain from the selfish act as to engage in the selfish act. A bribe that is even infinitesimally larger than e will be sufficient to suppress the act. Policing is not required for the act's suppression.

Next we ask under what conditions it will pay the suppressor to yield the fraction e (actually, e plus an infinitesimal quantity $\approx e$) of reproduction to the recipient to keep the latter from behaving selfishly. If the suppressor does not yield a fraction e to the recipient, then its personal reproduction is $(1 - p - z)(k - c)$ and that of the recipient is $(p + z)(k - c)$. If the suppressor does yield the fraction e (i.e., provides the bribe), then its personal reproduction is $(1 - p - e)(k)$ and that of the recipient is $(p + e)(k)$. Inequality (1) indicates that providing the bribe is favored for the suppressor if and only if

$$c(1 + r) > 0. \quad (4)$$

Since $c > 0$ by assumption, the suppressor thus is always favored to bribe the recipient, regardless of the relatedness between them. Condition (4) is independent of the initial reproductive share p, so it applies to either the dominant or subordinate member of the group. For example, subordinates will be favored to bribe dominants into avoiding selfish acts, as long as the subordinate's initial reproduction exceeds the minimum fraction necessary to make staying in the association favorable. Dominants also will be favored to bribe subordinates into avoiding selfish acts, as long as the dominant's initial reproduction exceeds the minimum fraction necessary to make retaining (not ejecting) the subordinate advantageous (fig. 1).

Two questions immediately arise regarding the evolutionary stability of bribing. First, why does the recipient not simply take the bribe and then act selfishly anyway? This is not problematic if the briber dispenses the bribe in sufficiently small packets over time and ceases dispensing it if the recipient initiates the

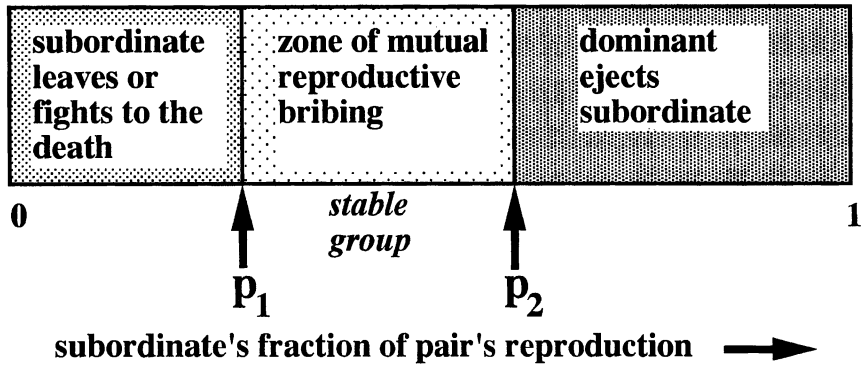

FIG. 1.—Mutual reproductive bribing zone. Here p_1 is the critical threshold of p for a subordinate above which it is favored to stay in the group and not fight to the death for complete control of the group's resources; p_2 is the critical threshold of p for the subordinate below which the dominant is favored to retain versus eject the subordinate.

selfish act. Such a temporal segmentation of the bribe coupled with retaliatory withdrawal of the bribe can make cheating unfavorable, as in the case of evolutionary stability of the tit-for-tat strategy in the iterated Prisoner's Dilemma game (Axelrod and Hamilton 1981). Such a possibility for retaliation is also assumed in the so-called optimal skew models (Vehrencamp 1979, 1983; Emlen 1982; Reeve and Ratnieks 1993) in which reproductive inducements are given by dominants to subordinates to form a stable group in the first place (Reeve and Nonacs 1992). Second, what prevents a recipient from "bluffing" a selfish act that is actually disadvantageous to it to receive a reproductive bribe? Whether the selfish act is favorable for the recipient is a fact that may be assessable by both recipient and potential briber; if this information is for some reason known only to the recipient, then there nevertheless can be an evolutionarily stable signaling system in which the recipient honestly communicates whether the selfish act will be advantageous—and thus the intention to act selfishly (Zahavi 1977; Grafen 1990). (As an example of the latter, whether the act will be favorable for the recipient may depend in part on some aspect of the recipient's quality or vigor that cannot be directly assessed by the briber. A recipient may nevertheless honestly signal its quality or vigor if the signals are both costly and differentially costly for lower-quality recipients; bribers would be selected to ignore signals that lack these features.) We stress that the key novel point of the bribe principle is that the potential briber will nearly always benefit from bribing instead of allowing a selfish act to go to completion; the separate problem of the evolutionary resistance of bribery to cheating is resolved with already established principles in reciprocity and communication theory.

Reproductive bribes are not expected to cause significant group-output costs, since the transfer of reproduction might simply involve the bribing individual's peaceful yielding of resources to the bribed individual. Suppose, however, that such a transfer of resources did cause a group-output cost equal to a (i.e., brib-

ing causes the group output to fall to $k - a$). Then, the new condition (4) under which bribing is favorable becomes $(c - a)(1 + r) > 0$, or just $a < c$. This means that even bribing that is not cost-free will evolve under the likely condition that group-level bribing costs are less than the group-level costs resulting from performance of the threatened selfish behavior.

IMPLICATIONS OF THE BRIBE PRINCIPLE: GROUP COOPERATION AND LEVELS OF INTRAGROUP AGGRESSION

Reproductive bribing enhances the likelihood of cooperation within animal societies since many potential selfish acts are predicted to be suppressed by the transfer of some reproduction from the potential victims to the potential perpetrators of such acts. This may be one reason that highly escalated aggression is infrequent within animal societies (Archer 1988; de Waal 1996). The bribe principle provides a straightforward explanation for the evolution of reconciliation in primates (de Waal 1996): affiliative behavior immediately following conflict may represent the first stage of reproductive bribing (and bribe reception) after selfish acts have been threatened.

The bribe principle also establishes a firm selective benefit for dominants that induce cooperative behavior by subordinates through reproductive payments. For example, it predicts that performance of increasingly cooperative acts by subordinates (such as foraging or group defense) will be accompanied by increasing reproductive payments from dominants, because the failure to cooperate will often be equivalent to a selfish act (as defined earlier). The bribe principle thus greatly extends previous optimal skew models, which show that, under certain ecological conditions, dominants will yield reproductive inducements (i.e., staying and peace incentives) to subordinates to ensure formation of a stable group (Emlen 1982; Vehrencamp 1983; Reeve 1991; Reeve and Ratnieks 1993; Reeve and Keller 1995). We show here that such inducements theoretically will be supplemented by additional inducements (bribes) to prevent selfish acts within a stable group (although, as will be discussed later, mutually negating selfish threats and bribing between subordinates and dominants can lead to no net change in the partitioning of reproduction).

The bribe principle also provides a precise, quantitative, and testable model of the link between selfish opportunities within animal societies and the distribution of reproduction within those societies. This is most clearly seen from an equivalent formula for the magnitude of a reproductive bribe:

$$e = \frac{\beta - \left(\frac{r}{1 - r}\right)c}{k}, \tag{5}$$

where β is the number of offspring gained by a selfish individual (obtained from eq. [3] by noting that $[p + c][k - c] - pk = \beta$ and dividing both the numerator and denominator by $[1 - r]$). This formulation shows that bribes will increase

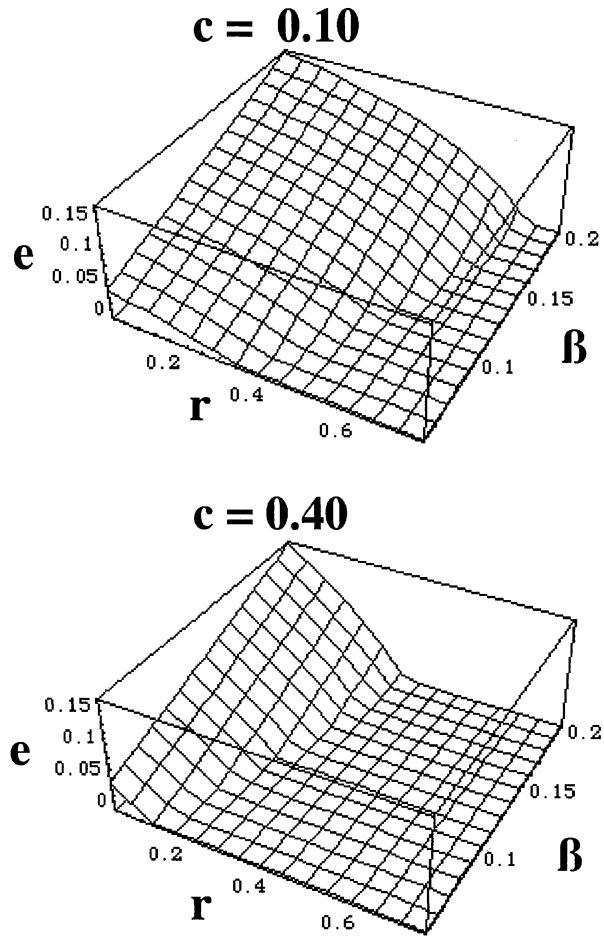

FIG. 2.—Magnitude of the reproductive bribe e as a function of the number of offspring β gained by a potentially selfish act, the relatedness between interactants r, and loss in group productivity due to selfishness c. For both graphs, $k = 1.2$. The reproductive bribe increases as loss in productivity because of selfishness c decreases.

linearly as β increases and decrease as c increases, k increases, and r increases (fig. 2). Thus, a recipient can maximize its received bribe by threatening a selfish act that has a high potential direct fitness payoff β and that would only slightly reduce group productivity by the amount c (if $r > 0$).

Thus, the bribe principle may explain both the maintenance and form of ritualized aggressive behaviors within animal societies. Such behaviors are predicted to be abbreviated forms of behaviors that would greatly increase β and cause only small group losses c if such behaviors were allowed to go to completion. Ritualized intention movements in aggressive displays (Tinbergen 1959) seem to have these features and thus receive a new functional interpretation under the bribe model.

Finally, a strikingly counterintuitive prediction of the bribe model as applied to aggressive threat displays is that such displays should be more frequent or intense within groups of relatives than within groups of nonrelatives, as indeed appears to be the case for multiqueen associations in social insects (Reeve and Ratnieks 1993; Bourke and Heinze 1994; Keller and Reeve 1994). Consider the case of mutual reproductive bribing as pictured in figure 1. Cooperative association of a pair is stable only if the subordinate's total fraction of reproduction both exceeds the critical threshold (p_1) at which it is favored to stay in the group (Emlen 1982; Vehrencamp 1983; Reeve and Ratnieks 1993) and not fight to the death for complete control of the group's resources and also lies below the critical threshold (p_2) at which the dominant is favored to eject the subordinate. Suppose, as predicted by optimal skew theory, that the subordinate's initial fraction p of reproduction $= p_1$. Inequality (4) indicates that the dominant will always be favored to yield a bribe e to the subordinate sufficient to prevent a selfish act that would increase the latter's offspring production by an amount β. Such a bribe will always be possible. The maximum selfish increment z for a subordinate without threatening group stability is $(p_2 - p_1)$, and it is easy to show that the corresponding required bribe e_{max}, given by equation (3) for $z = (p_2 - p_1)$, will be less than the amount available for the dominant to give $(p_2 - p_1)$ as long as $r > p_2/(p_2 - 1)$ (i.e., always). Thus, regardless of the fractions of reproduction necessary to keep subordinates from leaving the group and dominants from ejecting the subordinate, for any selfish act not threatening group stability, a reproductive bribe exists that will suppress the act. The subordinate will be favored to accrue reproductive bribes by threatening selfish acts until its total fraction of reproduction is at most p_2 (the maximum value p_2 is set by the assumption that above this value the dominant will eject the subordinate, which would be detrimental to the subordinate). However, inequality (4) also applies to bribes by the subordinate to prevent selfish acts by the dominant. The dominant will be favored to accrue reproductive bribes by threatening selfish acts until the subordinate's reproduction is at minimum p_1 (if the subordinate's share drops below p_1, then the dominant loses the assumed advantages of having the subordinate). Thus, at evolutionary equilibrium, the subordinate and dominant may both be threatening, but not performing, selfish acts, even if there is no net effect of bribes on their partitioning of the total reproduction.

It follows from the above arguments that the frequency or intensity of the threats of selfish acts (e.g., ritualized aggression) is expected to increase with the width of the mutual bribing zone between p_1 and p_2, since this width defines the maximum potential selfishness (i.e., the maximum cumulative selfish increment z for selfish acts). According to optimal skew theory, a nonzero p_1 will decrease with increasing relatedness (Emlen 1982; Vehrencamp 1983; Reeve and Ratnieks 1993). Moreover, if x is the expected success of an ejected subordinate (relative to 1.0 for a lone dominant) and f is its probability of winning a lethal fight, then, by Hamilton's rule, $p_2 = (k - 1 - rx)/k(1 - r)$ if an ejected subordinate reproduces solitarily and $(k - 1 + f[1 - r])/k(1 - r)$ if it engages the dominant in a lethal fight; both values increase with increasing relatedness r, given the necessary conditions for initial stability of the association ($x < k -$

1 and $k > 1$, respectively) (Reeve and Ratnieks 1993). It follows that increasing relatedness will increase $p_2 - p_1$ and consequently the overall expected frequency and/or intensity of selfish threats, in accordance with growing evidence on aggression among conesting social insect queens (Reeve and Ratnieks 1993; Bourke and Heinze 1994; Keller and Reeve 1994).

It is intriguing that reproductive bribing is not restricted to the transfer of direct reproduction (offspring production) from bribing to potentially selfish individuals; the principle also applies to the transfer of indirect reproduction (production of nondescendant relatives) between nonreproducing members of a society (e.g., sterile workers in a social insect colony). In particular, if the two members are each related to two subgroups of reared (potentially reproducing) nondescendant relatives by relatednesses r_1 and r_2, and r_2 and r_1, respectively, with $r_1 > r_2$, the magnitude of the reproductive bribe becomes equal to

$$e = \frac{r_1\beta - \left(\frac{r_2}{r_1 - r_2}\right)c}{k}, \tag{6}$$

where β is the gain in production of the closer relatives; the transfer of direct reproduction corresponds to the special case $r_1 = 1$ and $r_2 = r$. Thus, the bribe principle may also explain cases of surprising absence of conflict among even nonbreeding members (e.g., workers) of animal societies.

WHEN WILL BRIBING BE FAVORED OVER POLICING?

It is possible to predict when reproductive bribing versus policing will be the preferred mechanism for suppressing a destructive, selfish act, but we first need a simple model of policing that contains the same variables as the bribing model. As for bribing, policing is assumed to be an act by the suppressor that is conditional on the performance of a selfish, destructive act by the recipient (the latter satisfying condition [2]). Policing, however, occurs only if the destructive act is performed, whereas bribing occurs only if the destructive act is not performed. Policing reduces the selfish recipient's reproduction from $p + z$ to $p + z - y$ (e.g., by recovering part of a "stolen" amount of group resource) and also reduces the group output from $k - c$ to $k - c - a$, where a is the group cost of policing. In the case of policing, it cannot be safely assumed that the group-level cost of suppression is negligible (as is the case for bribing; see above), because the physical interference or sabotage that is the mechanism for policing might reasonably entail significant reductions in group productivity.

What minimum value of y, the increment in the policer's fraction of reproduction because of policing, would be sufficient to make a selfish act unfavorable for a recipient? The group-level cost of policing $a(y)$ is assumed to be an increasing function of y. The recipient's personal reproduction is $(p + z - y)(k - c - a[y])$ and the policer's personal reproduction is $(1 - p - z + y)$

$(k - c - a[y])$ if the selfish act is performed, and the recipient's personal reproduction is pk and the policer's is $(1 - p)k$ if the selfish act is not performed.

For the threat of policing to be effective, the policing action must genuinely benefit the policer (i.e., satisfy Hamilton's rule, eq. [1]), even if the recipient continues to perform and completes the selfish act. If the policing action were only a bluff (i.e., unfavorable for the policer when the selfish act is carried out), then suppression of the act would not be evolutionarily stable—recipients ignoring the bluff would invade the population, which subsequently would lead to selective elimination of the bluffing. As for the case of bribing (see above), whether policing is favorable for the potential policer is a fact that may be assessable by both policer and recipient; if this information is for some reason known only to the policer, then there nevertheless can be an evolutionarily stable signaling system in which the policer honestly communicates whether policing will be favorable and thus the intention to police (Zahavi 1977; Grafen 1990).

Thus, we seek the conditions under which performance of an effective level of policing y in the presence of a selfish recipient will be favorable for the potential policer. If the potential policer does not police, then its personal reproduction is $(1 - p - z)(k - c)$ and that of the recipient is $(p + z)(k - c)$. If it does police, then its personal reproduction is $(1 - p - z + y)(k - c - a[y])$ and that of the recipient is $(p + z - y)(k - c - a[y])$. Equation (1) indicates that policing is favored for the potential policer if and only if

$$a(y) < \frac{y(k - c)(1 + r)}{1 - (1 - r)(p + z - y)}. \tag{7}$$

Condition (7) may or not be satisfied, depending on the values of the parameters. It is more likely to be satisfied as the group cost of policing $a(y)$ decreases, the recipient's selfish increment z increases, the recipient's baseline fraction of reproduction p increases, the group cost of selfishness c decreases, genetic relatedness r decreases, and baseline group output k increases. The right-hand side of inequality (7) is the critical level of group cost below which policing is favored; we shall denote this threshold a_{fav}. The threshold a_{fav} continuously increases as the policer's increment y increases (for $y \leq z$) (fig. 3).

When condition (7) is satisfied, policing behavior will evolve, but will it be effective enough to cause recipients to refrain from their selfish, destructive acts? To answer this question, we first need to apply Hamilton's rule to the recipient's decision to continue being selfish, despite policing, versus to refrain from the selfish act. If the recipient continues to be selfish, then its personal reproduction is $(p + z - y)(k - c - a[y])$ and that of the policer is $(1 - p - z + y)(k - c - a[y])$. If the recipient does not behave selfishly, then its personal reproduction is pk and that of the policer is $(1 - p)k$. Hamilton's rule (1) indicates that refraining from selfishness will be favored when

$$a(y) > \frac{(k - c)(1 - r)(z - y) - c(p + r - rp)}{(1 - r)(z - y) + (p + r - rp)}. \tag{8}$$

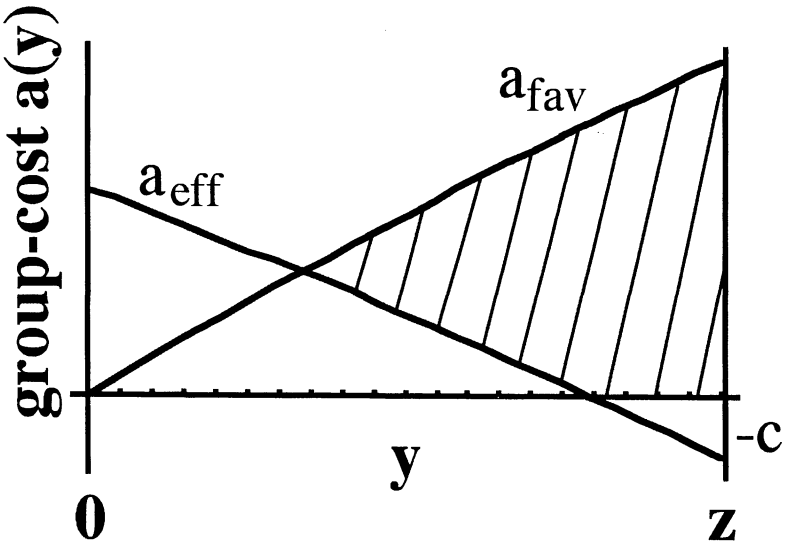

FIG. 3.—Conditions under which policing favored in the presence of selfishness will also be effective in deterring the selfish act. Here a_{fav} is the critical group cost below which policing is favored, a_{eff} is the critical group cost above which policing is effective, y is the policer's selfish increment in reproduction, z is the recipient's selfish increment in reproduction, and c is the group cost due to the recipient's selfish behavior (see the text). All of the points $(y, a[y])$ in the hatched region represent conditions under which favored policing is also effective; a_{fav} is nearly linear, and $a(y)$ is assumed linear or concave for y not close to z.

The right-hand side of inequality (8) is the critical level of group cost above which policing is effective in deterring selfishness; we shall denote this threshold a_{eff}. The threshold a_{eff} continuously decreases as the policer's increment y decreases (for $y \leq z$) (fig. 3). To show that favored policing also will be effective in deterring selfishness, we need to demonstrate that, if condition (7) is satisfied for at least some values of y, then at least some of these values of y will also satisfy inequality (8). We can do so graphically (fig. 3). We seek to show that, if $a(y)$ lies below a_{fav} for at least some values of y (with $y \leq z$), then there will exist values of y for which the group cost $a(y)$ also exceeds a_{eff}. Such a region must exist if $a_{fav} > a_{eff}$ at $y = z$ (fig. 3). Indeed, at $y = z$, $a_{eff} = -c$, which is negative, and $a_{fav} = z(k - c)(1 - r)/(1 - p + pr)$, which is positive. Thus, the required region exists: favored policing also will be effective in deterring selfishness. (It might appear that continued policing behavior might sometimes be favored even if the recipient terminates its selfish act, but such a behavior would then itself be a selfish act subject to policing or bribing by the recipient; such is not true for "true" policing behavior, which is strictly conditional upon recipient selfishness.)

Now we can return to the original question: When will policing be favored over bribing? If effective policing occurs, then the policer's output is $(1 - p)k$ and its partner's is pk. If bribing occurs, then the policer's output is only

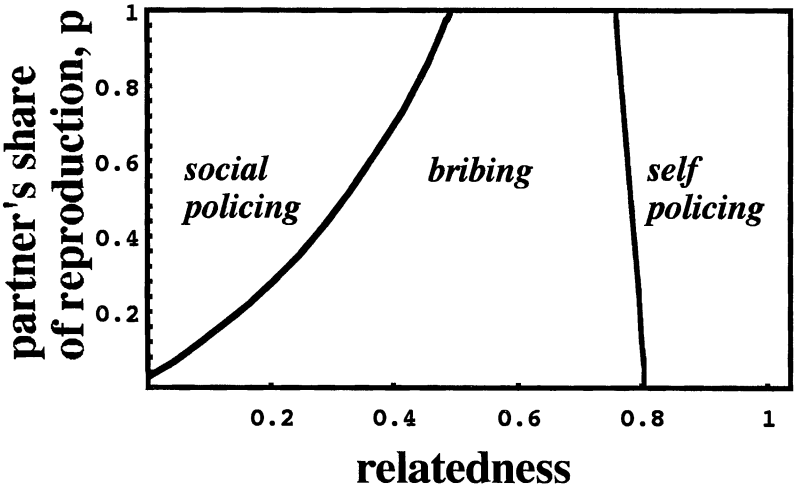

FIG. 4.—Conditions favoring reproductive bribing versus social policing and self-policing (see condition [7], the text). The ordinate is p, the recipient's fraction of reproduction, and the abscissa is r, the genetic relatedness between suppressor and recipient. Here $a(y) = 1.5y$, $y = 0.03$, $c = 0.03$, $z = 0.09$, and $k = 1.4$. Self-policing is favored when inequality (2) for selfish behavior by the recipient is not satisfied.

$(1 - p - e)k$ and its partner's is $(p + e)k$. Effective policing clearly will always be favored over bribing if it is favored at all. Successful policers prevent destructively selfish acts without having to donate part of their reproductive output to their partners. Thus, policing will be favored over bribing when condition (7) is satisfied; otherwise, bribing will be favored. This immediately generates the predictions that bribing will be most expected over policing when the group cost of policing a is high, the recipient's selfish increment z is small, the recipient's baseline fraction of reproduction p is low, the group cost of selfishness c is large, the genetic relatedness r is high, and baseline group output k is small (fig. 4). In other words, bribing is most expected between genetic relatives, when the threatened acts are mildly selfish but quite destructive, and for dominants instead of subordinates (i.e., it should be directed toward recipients with low p).

The study of intragroup conflict in animal societies is still in its infancy, but the simple theory of conflict suppression developed above from a multilevel selection approach provides a rich source of predictions for future research. For example, the prediction that policing will be more likely by subordinates than by dominants may help explain why simulated reproductive cheating (random egg removal) causes markedly increased aggression in subordinate, but not dominant, social wasp foundresses (Reeve and Nonacs 1992). The most sweeping prediction is that reproductive bribing should be widespread, particularly among genetic relatives. This prediction will be testable only with careful observations

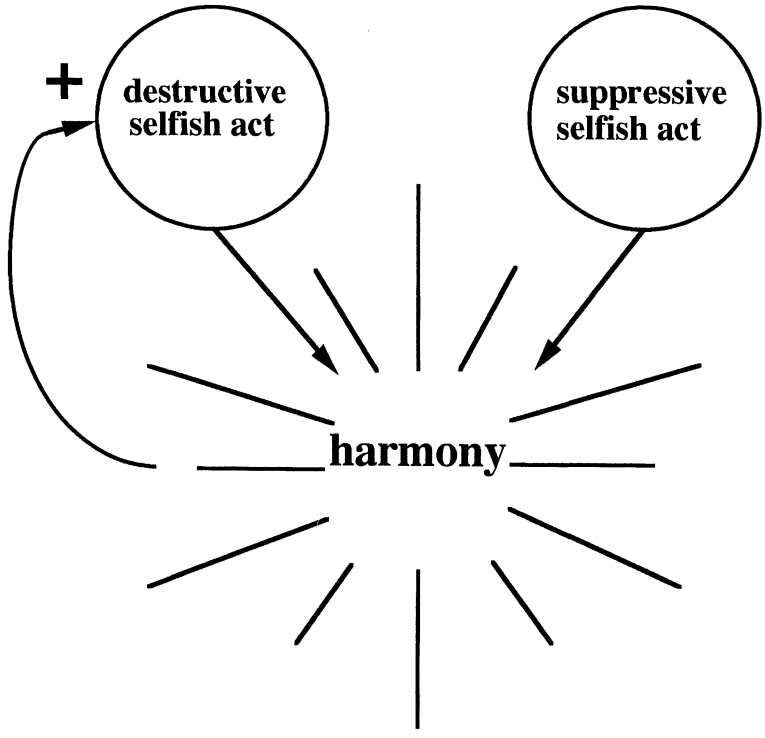

FIG. 5.—"Collision" of a selfish, destructive act with another selfish act (social suppression) producing harmony. The resultant increase in group productivity automatically increases the temptation for other selfish, destructive acts, which reduces the degree to which individuals appear to maximize group reproduction.

of potentially quite subtle behavioral interactions representing transfers of reproduction within groups.

THE CONSEQUENCES OF SOCIAL SUPPRESSION: A FEEDBACK LOOP PROMOTING OTHER CONFLICTS

In both reproductive bribing and policing, social suppressors prevent selfish acts with behaviors that are themselves selfish (even in reproductive bribing, the briber ends up with a larger share of reproduction than would have resulted if the original selfish act were allowed to occur, and thus bribing is still selfish as defined earlier). In an analogy from particle physics, two selfish acts "collide," with the suppressive act "annihilating" the originally threatened, destructive selfishness to produce harmony (i.e., avoidance of a potential reduction in group reproductive output; fig. 5). However, this conflict resolution can have the somewhat paradoxical effect of promoting other within-group conflicts. To see this, it is useful to derive a general expression for the difference in the group-level

optimum for an individual's behavior and the individual's inclusive fitness optimum for the same behavior.

We begin with a continuous behavior described by the variable x. The discrete form of Hamilton's rule (eq. [1]) implies that selection will favor the value of x that maximizes the inclusive fitness quantity

$$I(x) = p(x)k(x) + r[1 - p(x)]k(x).$$

Now the value of $x = x_0$ that maximizes group output will have the properties $k'(x) = 0$ and $k''(x) < 0$ at $x = x_0$ (for an internal maximum). We can derive an approximation for the value of x that maximizes the inclusive fitness $I(x)$ in the following way: we take the derivative of $I(x)$ with respect to x and perform a first-order Taylor expansion of this derivative around the group optimum x_0. We then obtain

$$\begin{aligned} dI(x)/dx \approx\ & (1 - r)k(x_0)p'(x_0) + (x - x_0) \\ & \{rk''(x_0) + (1 - r)[p(x_0)k''(x_0) + k(x_0)p''(x_0)]\}. \end{aligned} \tag{9}$$

Now the inclusive fitness maximum will occur at the value of x^* at which $dI(x)/dx = 0$. Substituting expression (9) into the latter equation and solving for x, we obtain

$$x^* = x_0 + \frac{(1 - r)k(x_0)p'(x_0)}{-\{rk''(x_0) + (1 - r)[p(x_0)k''(x_0) + k(x_0)p''(x_0)]\}}. \tag{10}$$

In other words, the approximate difference between the individual (inclusive fitness) optimum and the group optimum is equal to the right-hand term of the right side of equation (10). It can be shown that the inclusive optimum will tend to be higher than the group optimum if $p'(x_0) > 0$ and less than the group optimum if the reverse is true. This makes sense. If $p'(x) > 0$, then the selfish payoffs for higher values of the behavior drives the individual optimum to a higher value than the group optimum. It follows from these considerations that the overall sign of the denominator of the right-hand expression must be positive, a fact that will be used later.

Equation (10) allows us to analyze which factors affect the deviation of the individual optimum from the group optimum. As is intuitive, the deviation declines with increasing relatedness and is 0 in the extreme case of $r = 1$. However, the divergence might be 0 even if $r = 0$ (i.e., for unrelated group members) if $p'(x_0) = 0$ (i.e., if there is no change in selfish share of the reproduction as the behavior deviates in small steps from the group optimum). Policing mechanisms might be seen as producing the latter effect (nullifying selfish gains for deviating from the group optimum), thereby causing individuals to act more like they were maximizing group output, even in the absence of genetic relationship (see also Frank 1995). A bribe principle also is implicit in equation (10), because increasing the fraction $p(x_0)$ of reproduction at the group optimum (i.e., bribing) will cause the absolute deviation of the individual optimum from the group optimum to decrease (since $-k''[x_0] > 0$, and the overall sign of the denominator is positive).

The feedback effect promoting new conflicts is related to the term $k(x_o)$ in the right-hand expression. The absolute deviation between the individual and group optimum increases with $k(x_o)$. Policing and bribing mechanisms will increase k by suppressing threatened destructive selfishness. Increases in k in turn will automatically increase the payoffs for other destructive selfish acts. From equation (2), the inclusive fitness payoff for selfishness in the absence of suppression is equal to $(p + z)(k - c) - pk + r[(1 - p - z)(k - c) - (1 - p)k]$. This payoff increases as k increases since the derivative of this payoff with respect to k is $z(1 - r) > 0$. Thus, a paradoxical effect of closing the gap between the group optimum and individual optimum for one behavior will be an increasing separation of these optima for other behaviors. The reason for this is that as group output due to cooperation increases, the payoff for enhancing one's share of this output through other routes also increases: harmony begets conflict (fig. 5). This may be one reason that moderate reproductive conflicts can be so frequent in large, highly productive social groups with strongly skewed reproduction (like societies of naked mole rats), even despite high levels of intracolony genetic relatedness (Reeve and Sherman 1991; Reeve 1992). A possible experimental test for this effect would be experimentally to augment (or reduce) group output and then look for increases (or decreases) in policing- or bribing-related behaviors specific to different kinds of conflicts. The general implication of this theoretical feedback effect is that although it is useful to view social suppression as eliminating certain kinds of conflicts, researchers must be mindful of other conflicts that might be enhanced by such suppression. The interaction between cooperation and conflict may have a more complex dynamic structure than has been previously appreciated.

In conclusion, a multilevel selection approach demonstrates that reproductive bribing can be added to policing (punishment and sabotage) as a theoretically important mechanism by which intragroup selfish behavior can be suppressed, leading to increased convergence between individual and group behavioral optima for that behavior. Reproductive bribing provides surprisingly robust benefits for bribers, and the subtle transfers of reproduction involved in bribing are predicted to be widespread, especially when potential bribers are sufficiently highly related to potential recipients, are dominants rather than subordinates, and/or when the selfish act suppressed by bribing causes moderate selfish gains but is quite destructive. Although suppression of selfishness via bribing and policing has the theoretical effect of increasing the degree to which a particular behavior will appear to maximize group reproduction, it should also paradoxically increase the temptation to engage in other selfish, destructive behaviors.

ACKNOWLEDGMENTS

We thank the Swiss National Science Foundation (L.K.) and the U.S. National Science Foundation (H.K.R.) for support. We especially thank D. S. Wilson for inviting us to participate in the multilevel selection symposium. We also

thank P. Christe, L. Dugatkin, S. Emlen, J. Maynard Smith, N. Perrin, A. Pusey, M. Scott, J. Shellman-Reeve, P. Sherman, and P. Taylor for valuable comments.

LITERATURE CITED

Archer, J. 1988. The behavioural biology of aggression. Cambridge University Press, Cambridge.

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

Bourke, A. F. G., and J. Heinze. 1994. The ecology of communal breeding: the case of multiple-queen leptothoracine ants. Philosophical Transactions of the Royal Society of London B, Biological Sciences 345:359–372.

Clutton-Brock, T. H., and G. A. Parker. 1994. Punishment in animal societies. Nature (London) 373: 209–216.

de Waal, F. B. M. 1996. Good natured: the origins of right and wrong in humans and other animals. Harvard University Press, Cambridge, Mass.

Dugatkin, L., and H. K. Reeve. 1994. Behavioral ecology and the levels-of-selection debate: dissolving the group selection controversy. Advances in the Study of Behavior 23:101–133.

Emlen, S. 1982. The evolution of helping. I, II. American Naturalist 119:29–53.

Frank, S. A. 1995. Mutual policing and repression of competition in the evolution of cooperative groups. Nature (London) 377:520–522.

Grafen, A. 1984. Natural selection, kin selection and group selection. Pages 62–89 *in* J. R. Krebs and N. B. Davies, eds. Behavioural ecology: an evolutionary approach. 2d ed. Sinauer, Sunderland, Mass.

———. 1990. Biological signals as handicaps. Journal of Theoretical Biology 144:517–546.

Hamilton, W. D. 1964. The genetical evolution of social behavior. I, II. Journal of Theoretical Biology 7:1–52.

Keller, L., and H. K. Reeve. 1994. Partitioning of reproduction in animal societies. Trends in Ecology & Evolution 9:98–102.

Parker, G. 1989. Hamilton's rule and conditionality. Ethology, Ecology, and Evolution 1:195–211.

Queller, D. C. 1992. A general model for kin selection. Evolution 46:376–380.

Ratnieks, F. L. 1993. Egg-laying, egg-removal, and ovary development by workers in queenright honey bee colonies. Behavioral Ecology and Sociobiology 32:191–198.

Ratnieks, F. L., and H. K. Reeve. 1992. Conflict in single-queen hymenopteran societies: the structure of conflict and processes that reduce actual conflict. Journal of Theoretical Biology 158: 33–65.

Ratnieks, F. L., and P. K. Visscher. 1989. Worker policing in the honey bee. Nature (London) 342: 796–797.

Reeve, H. K. 1991. The social biology of *Polistes*. Pages 99–148 *in* K. Ross and R. Matthews, eds. The social biology of wasps. Cornell University Press, Ithaca, N.Y.

———. 1992. Queen activation of lazy workers in colonies of the eusocial naked mole-rat. Nature (London) 358:147–149.

Reeve, H. K., and L. Keller. 1995. Partitioning of reproduction in mother-daughter versus sibling associations: a test of optimal skew theory. American Naturalist 145:119–132.

Reeve, H. K., and P. Nonacs. 1992. Social contracts in wasp societies. Nature (London) 359:823–825.

Reeve, H. K., and F. A. Ratnieks. 1993. Queen-queen conflict in polygynous societies: mutual tolerance and reproductive skew. Pages 45–85 *in* L. Keller, ed. Queen number and sociality in insects. Oxford University Press, Oxford.

Reeve, H. K., and P. W. Sherman. 1991. Intra-colonial aggression and nepotism by the breeding female naked mole-rat. Pages 337–357 *in* P. Sherman, J. Jarvis, and R. D. Alexander, eds. The biology of the naked mole-rat. Princeton University Press, Princeton, N.J.

Tinbergen, N. 1959. Comparative studies of the behavior of gulls (Laridae): a progress report. Behavior 15:1–70.

Trivers, R. L. 1971. The evolution of reciprocal altruism. Quarterly Review of Biology 46:35–57.

Vehrencamp, S. L. 1979. The roles of individual, kin and group selection in the evolution of sociality. Pages 351–394 *in* P. Marler and J. Vandenbergh, eds. Social behavior and communication. Plenum, New York.

———. 1983. Optimal degree of skew in cooperative societies. American Zoologist 23:327–335.

West Eberhard, M. J. 1981. Intragroup selection and the evolution of insect societies. Pages 3–17 *in* R. D. Alexander and D. Tinkle, eds. Natural selection and social behavior. Chiron, New York.

Wilson, D. S., and E. Sober. 1994. Reintroducing group selection to the human behavioral sciences. Behavioral and Brain Sciences 17:585–608.

Zahavi, A. 1977. The cost of honesty: further remarks on the handicap principle. Journal of Theoretical Biology 67:603–605.