ELSEVIER

Evolution and Human Behavior 23 (2002) 203–231

Evolution and Human Behavior

Punitive sentiment as an anti-free rider psychological device

Michael E. Price[a,*], Leda Cosmides[b], John Tooby[a]

[a]*Department of Anthropology, Center for Evolutionary Psychology, University of California, Santa Barbara, CA 93106, USA*
[b]*Department of Psychology, University of California, Santa Barbara, CA 93106, USA*

Received 3 August 2000; received in revised form 24 August 2001; accepted 13 September 2001

Abstract

Those who contribute to a public good sometimes experience punitive sentiments toward others. But is the system that produces these sentiments an adaptation and, if so, which collective action problem was it designed to solve? Prior results from experimental economics show that acts of free riding are sometimes punished, that punishment deters free riding, and that the risk or actuality of punishment recruits higher levels of cooperation in a joint effort. This suggests that one function of punitive sentiments could be to recruit labor for collective actions. However, adaptations designed to cause participation in collective actions could not have evolved unless there were some mechanism that protected those who participated from having lower fitness than nonparticipating free riders. Therefore, a second possible function of punishment could be to eliminate or reverse fitness differentials that favor free rider designs over participant designs. To map the computational structure of this motivational adaptation (and hence identify its specific function) requires data that relate an individual's circumstances to his or her desire to punish. Herein, we report such data. The results indicate that the computational system that regulates one's level of punitive sentiment in collective action contexts is functionally specialized for removing the fitness advantage enjoyed by free riders rather than for labor recruitment or other functions. Results also support the hypothesis that a separate pro-reward motivational system exists that appears designed to handle the problem of labor recruitment. Rational choice counterexplanations for punitive sentiments were considered but eliminated on the basis of the evidence.

Keywords: Collective action; Altruism; Punishment; Morality; Cooperation; Public goods; Rational choice

* Corresponding author.
E-mail addresses: mep2@umail.ucsb.edu (M.E. Price), cosmides@psych.ucsb.edu (L. Cosmides), tooby@anth.ucsb.edu (J. Tooby).

1090-5138/02/$ – see front matter
PII: S1090-5138(01)00093-9

1. Introduction

Individual participation in collective action is one of the most intensively discussed issues in the behavioral sciences for a simple reason. It is, at present, a phenomenon in search of an explanation. On the one hand, it is clear that individual humans routinely and willingly participate in projects that require collective action: Sets of individuals will cooperate to achieve a common goal even when the rewards to individuals are not intrinsically linked to individual effort. Cooperation to provide a public good happens not just in agricultural and industrial societies but in hunter-gatherer and hunter-horticulturist ones as well. Plausible examples include cooperative hunting and food sharing (Gurven, Hill, Kaplan, Hurtado, & Lyles, 2000; Hawkes, 1993; Kelly, 1995; Lee & DeVore, 1968; Smith, 1985), offensive raids and collective defense (reviews in Keeley, 1996; Wrangham & Peterson, 1996), cooperative shelter building (Chagnon, 1997; Harner, 1984), and field clearings (Holmberg, 1969; Smole, 1976). On the other hand, game theoretical analyses in economics and biology have shown that the incentives individuals face in many collective action problems are insufficient to promote voluntary contributions to public goods and instead favor free riding and defection as the equilibrium outcome (Hardin, 1968; Olson, 1965; for reviews of studies related to collective action problems, see Ledyard, 1995; Ostrom, 1998). These impediments to collective action hold, whether the currency is monetary payoffs to a rational actor or fitness payoffs to alternative heritable neurocognitive design features. How, then, could selection have favored the spread of psychological design features that cause participation in collective action? That is, how should the fact that people often willingly sacrifice for their coalition, country, political party, or residential group be explained? Is this behavior an incidental byproduct of psychological mechanisms designed for some other purpose, or is it the signature of an adaptation that evolved specifically for collective action under ancestral conditions? (For discussion, see Alexander, 1987; Patton, 1996; Stern, 1995; Tooby & Cosmides, 1988; Wright, 1994.)

Recent models of the evolution of collective action have focused on the role of punishment (Boyd & Richerson, 1992; Gintis, 2000; Henrich & Boyd, 2001). These models show that willingness to contribute to a public good can be evolutionarily stable as long as free riders are punished, along with those who refuse to punish free riders. Moreover, research in experimental economics using public goods games has shown that higher levels of cooperation result when the probability that free riding will be punished is large enough (Fehr & Gächter, 2000a; Kurzban, McCabe, Smith, & Wilson, 2001). More puzzling from a game theory and selectionist standpoint, the results of these studies clearly show that individuals are willing to incur personal costs in order to punish free riders (Dawes, Orbell, & Van de Kragt, 1986; Fehr & Gächter, 2000a; Ostrom, Walker, & Gardner, 1992; Sato, 1987; Yamagishi, 1992). They do this even when it appears that they are unlikely to have future interactions with the individual they punished and, therefore, are unlikely to recoup their losses in the form of increased cooperation from that person in the future (for review, see Gintis, 2000).

Unfortunately, these results do not make the evolution of adaptations for collective action any less mysterious. Because punishing a free rider would generally have entailed some

nontrivial cost, each potential punisher has an incentive to defect — that is, to avoid this cost by not punishing acts of free riding. Thus, the provision of punishment is itself a public good: Each individual has an incentive to free ride on the punishment activities of others (Henrich & Boyd, 2001; Sober & Wilson, 1998; Yamagishi, 1986). Hence, second-order free riders should be fitter (or better off) than punishers. Without a way of solving this second-order free rider problem, cooperation should unravel, with nonparticipation and nonpunishment the equilibrium outcome. Even worse, this problem reappears at each new level, revealing an infinite regress problem: Punishment needs to be visited on free riders on the original public good, and on those who do not punish free riders, and on those who do not punish those who do not punish free riders, and so on. A number of models — some invoking group selection (Gintis, 2000), others not (Boyd & Richerson, 1992; Henrich & Boyd, 2001; Hirshleifer & Rasmusen, 1989; Tooby & Cosmides, 1988) — have been proposed to solve this problem. All have problems, and there is no consensus yet on which is most likely to be correct.

How adaptations for collective action could have evolved given the free rider problem is puzzling at present; 30 years ago, selectionists were wondering how sexual recombination could have evolved given the cost of meiosis. At the time, none of the existing models were fully adequate (Williams, 1975). Nevertheless, as Williams pointed out, one could still draw sound conclusions about the functions of many subcomponents of sexual reproduction, despite the fact that other aspects remained mysterious (including its ultimate function):

> The machinery of sexual reproduction in higher animals and plants is unmistakably an evolved adaptation. It is complex, remarkably uniform, and clearly directed at the goal of producing, with the genes of two parental individuals, offspring of diverse genotypes. How the production of diverse rather than uniform offspring contributes to the ultimate goal of reproductive survival may not be immediately obvious, but the precision of the machinery can only be explained on the basis of selection for efficiency in the production of offspring with the parental genes but not the parental genotypes. (Williams, 1966, p. 125)

The point is this: When selectionist theories have hit an impasse, one can often make headway by applying adaptationist tools (Williams, 1966, 1975). By exploring what sets of outcomes each subcomponent of an adaptive system seems narrowly designed to produce, one may be able to deduce their functions, i.e., the selection pressures that built the machinery.

We think that punitive and pro-reward motivational circuitry can be approached in this fashion. Unlike meiosis and gametogenesis, however, the motivational responses in humans to situations of collective action are not yet mapped with enough resolution to determine whether they are adaptations. But by mapping their internal interrelationships — what outcomes (motivations, decisions, and behaviors) they produce and what variables regulate those outcomes — we think progress can be made in discovering their design and, eventually, their adaptive function (if any).

Therefore, we wish to raise the following questions: Is there evolved neurocognitive circuitry that causes people to punish free riders in collective action contexts? More importantly, does our motivational system show sufficient evidence of special design for this outcome that we should conclude it is a solution to the problem of punishing free riders, rather than a byproduct or maladaptive misfiring of a mechanism designed for solving some

other kind of problem? If the answers to these questions are yes, then this raises a third, less explored question: What exactly was the selective advantage of punishment in ancestral collective action contexts — that is, what was the functional consequence of punishing?

The existing evidence, while suggestive, is not sufficient to answer these questions. Research in experimental economics shows that (1) people are willing to incur unreimbursed costs to punish others (a puzzle), (2) overall levels of cooperation are higher when opportunities to punish free riders exist (an unsurprising outcome that would be produced simply by a general ability to anticipate and respond to incentives, assuming punishment is anticipated), and (3) *average* levels of punishment are greater the more a free rider's contribution falls below the average contribution of the other group members (e.g., Fehr & Gächter, 2000a, 2000b). However, this research does not include analyses of individual choices to cooperate and punish, to see if and how the two are connected. Specifically, the data presented so far do not tell us whether a connection exists — and if so, how large it is — between an *individual's* willingness to contribute to the provision of a (first-order) public good and that individual's willingness to punish free riders. Nor do these data tell us whether one's willingness to punish is better predicted by one's willingness to contribute to a public good or by some other, possibly correlated variable, such as the perception that one will personally benefit from a collective action. The study reported herein was designed to address these issues.

1.1. Punitive sentiment as an anti-free rider psychological device

By a punitive sentiment, we mean the designed expression of evolved, reliably developing circuitry in the motivational system: specifically, a desire that the target of the sentiment be harmed. (Additional features could include the desires that (1) the target be aware of being harmed, (2) the target know why, and (3) others be aware of the reason for the punishment.) Consistent with the proposals of many researchers, we hypothesize that there is a motivational adaptation that evolved specifically to cause punitive sentiments toward free riders, with free riders being defined (by the computational adaptation) as individuals who (1) benefit from a collective action, (2) could have contributed to the success of the joint effort without incurring a cost disproportionately greater than their share of the expected benefit, and (3) did not expend adequate effort toward the collective action. The more punitive sentiment this circuitry generates, the more the experiencer is induced to take punitive action despite the existence of collateral costs or motivations that would otherwise deter such action.

An adaptation causing punitive sentiments toward nonparticipants in a collective action could have increased the probability that free riders were actually punished in three distinct but nonexclusive ways. First, it could motivate the experiencer to inflict punishment personally, independent of the actions of others and despite some personal cost. In this case, the sentiment would also act as a commitment mechanism (Frank, 1988; Hirshleifer, 1987). Indeed, there is some evidence that those who incur costs to punish in public goods games were angered by acts of free riding (Andreoni, 1995). Second, a punitive sentiment could have induced its experiencer to participate in or proportionately subsidize punishment efforts provided they are joint, since these levy a smaller personal cost than unilateral acts of

punishment. By definition, joint punishments reduce or eliminate the possibility of second-order free riding. Third, it could have induced the experiencer to advocate punishment of free riders without subsidizing it (although the functional dynamics of advocacy compared to other kinds of action depend on other assumptions about the social environment). Not all coalition members would have had to participate equally in every act of punishment to deter free riding. By supporting a punishment norm, advocacy could help advertise the high costs of free riding (thereby deterring it) and perhaps induce others to punish by making it clear that their efforts will not be opposed (e.g., Henrich & Boyd, 2001).

What features of special design should we expect in a system designed for motivating the punishment of nonparticipants? The answer differs depending on exactly which adaptive problem this punitive sentiment evolved to solve.

1.2. Recruiting participants or eliminating adverse fitness differentials?

An adaptation that causes punitive sentiments toward nonparticipants in a collective action could have two separate, though compatible, functions:

1. It could be designed to increase the probability that a public good will be successfully provided. Such an adaptation would be designed to accomplish this by encouraging otherwise reluctant parties to participate more fully in a collective action (Alexander, 1987; Cronk, 1994; Irons, 1991; Sato, 1987; Tooby & Cosmides, 1988). The threat of punishment could increase participation from three different labor pools: (a) free riders, (b) nonparticipants who are not free riders,[1] and (c) non-free riding participants who, through working even harder, could increase the public good.

[1] Indeed, given the definition of free riding that we propose is built into our psychology, it should be clear that one can refrain from participating in a collective action without being a free rider for several distinct reasons. Nonparticipants who do not take or accrue the rewards of a collective action are not free riders; nor are those who, due to injury, pregnancy, or other reasons, are incapable of contributing. If only free riders are the targets of punishment, then these other categories of individual should be seen as exempt. Moreover, it is worthwhile to distinguish actors according to the degree they benefit from the collective action, their costs of participating, and whether they participated. Those whose benefits from a collective action would have exceeded their costs of participating and yet do not participate are free riders by any useful definition. In contrast, one could take two views about whether the following set of individuals is usefully categorized as free riders: those for whom the cost of participating in a collective action would have exceeded their benefit from that collective action, and so do not participate. If they participated, they would be the victims of exploitation, contributing to others' net benefits while incurring a net deficit themselves. If they failed to participate, they would be getting benefits out of the efforts of others without contributing to their production. An interesting question is, therefore, which of the two definitions is embodied in the decision-making psychology of collective action: Are they categorized as free riders or exempt nonparticipants? We suspect that under more egalitarian conditions, our psychology does not categorize them as free riders. To exempt these individuals from punishment — that is, to categorize them as non-free riders — is more evolutionarily stable. After all, the alternative — a collective action system that punished these individuals — would be pitting itself against adaptations that resist exploitation. We suspect, however, that when a sufficient number of individuals interested in the collective action are powerfully situated, they switch to the more exploitive definition of free rider, and include those whose costs of participating exceed the benefits of the action as suitable targets of punishment.

2. Punitive sentiments toward nonparticipants could be designed to eliminate the fitness advantage free riders accrue over participants in collective action by inducing participants to proportionately damage the fitness of free riders (Tooby & Cosmides, 1996). By definition, both free riders and participants in collective action enjoy its benefits, but the participants pay a cost that free riders do not and so suffer from a relative fitness disadvantage. Design features that motivate participation in collective action could not have been systematically favored or be evolutionarily stable unless the free riders' default fitness advantage had been somehow eliminated or reversed.

Eliminating the fitness advantage of free riding is a separate adaptive function from maximizing the profitability of a collective action venture through optimizing the number of participants recruited. Punitive sentiments in collective actions could have been designed to accomplish either, both, or neither of these adaptive functions. Because each function carries different predictions, it is possible to test among these alternatives. Of course, because of the way the world is structured, punishment sufficient to eliminate the fitness benefits of free riding may also have the collateral effect of encouraging participation — it may motivate the participation of individuals who would otherwise free ride. Nevertheless, the predictions remain sufficiently different that progress can be made differentiating these two functions.

In addition, it is commonplace for different components of adaptive machinery to solve different but interlocking adaptive problems — the eye, for example, has a lens for focusing light and rhodopsin for registering its presence. Similarly, the two functions of optimizing labor recruitment and eliminating adverse fitness differentials need not be handled by precisely the same behavior-regulatory circuit logic. Furthermore, the adaptive problem posed by the potential proliferation of nonpunishers (as second-order free riders) may not be solved by the same means and circuitry as the adaptive problem created by the potential proliferation of nonparticipants or undercontributors (as first-order free riders). This means that we can treat these three adaptive problems independently, and proceed to test competing hypotheses about the adaptive function of punitive sentiments toward first-order free riders without knowing how the adaptive problem of second-order free riding has been solved. Once the function of punishing first-order free riders is clarified, we can see if this throws light on competing theories of how the problem of second- and higher-order free riding on punishers might have been solved.

1.2.1. Recruiting participants

There are many evolutionarily recurrent contexts — war being the most obvious — in which success in providing a public good (e.g., defense) may be a function of either how many individuals participate or the degree to which participants expend effort (Tooby & Cosmides, 1988; Wrangham & Peterson, 1996). This means that the problem of recruiting sufficiently many participants (or sufficiently effortful participants) would have often occurred. Moreover, the experimental work by Fehr and Gächter (2000a) shows that the same individuals contribute at higher levels when the possibility of punishment exists than when it does not, showing that punishment is in fact effective at mobilizing higher rates and levels of participation. Whether this is punishment's selectively designed function or merely a beneficial byproduct remains to be determined.

Listing all of the design features expected of a system whose function was to optimize recruitment is beyond the scope of this article, but certain predictions are relevant to this study.

(a) If encouraging participation by otherwise reluctant individuals is one adaptive function of a system that causes punitive sentiments toward nonparticipants, then those who are most likely to benefit from the achievement of a group goal should differentially act to induce others to achieve this goal (e.g., Alexander, 1987, pp. 191–192; Cronk, 1994; Irons, 1991). More specifically, the adaptation should be designed such that the greater an individual's expected benefit from a successfully executed collective action, the more punitive sentiment that individual will experience.

(b) The predicted relationship between own expected benefit and sentiment for punishing others should remain significant, even when one controls for the individual's own willingness to participate. After all, encouraging self-sacrificial participation by others provides the largest net benefit, even for a free rider.

(c) If encouraging participation by others is this adaptation's *only* function, then after controlling for perceived benefit, any relationship between the individual's willingness to participate and punitive sentiment toward others should disappear.

(d) Sentiment for rewarding participants should track sentiment for punishing nonparticipants. There is nothing inherent in the problem of labor recruitment that privileges punishment over reward as an incentive — each might serve to motivate recruitment. If the function of punitive sentiment is purely for motivating recruitment, then sentiment for rewarding participants should be correlated with sentiment for punishing nonparticipants, since they both serve the same function.

Equally important (although not testable in this study) are three further predictions:

(e) If optimizing labor recruitment were the overriding selection pressure designing punitive sentiments toward nonparticipants, then this system should be sensitive only to the labor needs of the collective action, not to the existence of free riders per se. The target of punitive sentiments ought to be those whose punishment-inducible participation would most help the collective action, not those nonparticipants who differentially benefit by the collective action. Once the manpower needs of the collective action are satisfied, a system with this function should be indifferent to the prospering of free riders, and their continued presence should not provoke punitive sentiments.

(f) Equally, a system designed to optimize recruitment should be indifferent to whether a nonparticipant is a free rider (i.e., someone for whom the collective action is beneficial) or not. What matters is whether a potential recruit's participation would be beneficial to the collective action, not whether the collective action helps or harms the recruit. This is the reverse of what would be expected if the function were to eliminate the fitness advantages of free riders. In that case, a nonparticipant who does not benefit by a collective action is not a free rider, regardless of how much his or her participation would help in provisioning the collective good. If punitive sentiments were designed exclusively to punish free riders, then nonparticipants who do not benefit from the collective action but who could have helped it succeed ought not to stimulate punitive sentiments.

(g) Finally, because the function of a participation-managing system is to induce others to contribute in a way that maximally benefits the incentive-manipulator, individuals who

contribute anything less than the optimal amount might be suitable targets for punishment. Indeed, if participation management is the function, those who benefit the most from a collective action should feel punitive even toward individuals who *do* contribute at an average level, if this would encourage them to contribute even more. (Evidence from other studies relevant to this prediction will be discussed in Section 4.)

1.2.2. Eliminating adverse fitness differentials

In evolving populations, heritable designs are selected for to the extent they exhibit a fitness advantage relative to their competitors. Because characteristics that create a relative fitness advantage are not always the same as those which maximize absolute returns, adaptationist predictions sometimes diverge from the predictions of rational choice theory. In the case of collective action, collective producers realize benefits from productive action that nonlocal nonproducers do not, and so enjoy higher fitness compared to them. Unfortunately, local designs that free ride enjoy even higher relative fitness and so outcompete simple producers. Reversing this relative fitness advantage is mandatory if designs that reap the benefits of producing through collective action are to prevail. Adaptations in producers might be expected to evolve, even if they lower absolute returns, provided they lower the returns to free riders even more, thereby creating a relative fitness advantage for producers over free riders.

Eliminating or reversing this adverse fitness differential is a specific, logically distinct adaptive problem that could have selected for neurocognitive circuitry narrowly specialized solely for this task. Moreover, the outputs of such circuitry might appear nonrational (i.e., individually costly) or even spiteful because their function is to reverse relative fitness orderings rather than to maximize returns (Tooby & Cosmides, 1996). Hence, an alternative hypothesis for the function of punitive sentiments is to motivate actions that remove or reverse the fitness differential that would accrue to free riders relative to producers in the absence of punishment.

The hypothesis that punitive sentiments toward nonparticipants were designed to eliminate the fitness advantage of free riders predicts the following:

(a) The individual's own participation is the specific factor that should trigger punitive sentiments toward free riders. This is because only those individuals who contribute to a collective action are at risk of incurring lower fitness relative to free riders.

(b) The more an individual contributes, the greater the adverse fitness differential s/he potentially suffers relative to free riders. A sentiment designed to redress adverse fitness differentials and prevent outcompetition by free riders should therefore key the *degree* of punitive sentiment toward free riders to the individual's own willingness to participate in a collective action. The more willing the individual is to participate, the more that individual should wish to see free riders punished.

(c) This relationship between participation and punitive sentiment should be specific and selective: Punitive sentiment should track willingness to participate strongly even when one controls for other variables. For example, individuals may differ in the extent to which they will benefit if a collective action succeeds. Accordingly, the perception that one will benefit from a successful collective action should be positively correlated with one's willingness to participate. Even when benefits are equal, however, the costs of participation will inevitably

vary from individual to individual, making the correlation between perception of benefit and willingness to participate imperfect at best. However, it is participation per se and not degree of benefit that makes the design vulnerable to free riding. Therefore, it is participation rather than the benefit derived that should — on this theory — predict punitive sentiment. If this is true, then punitiveness should track willingness to participate strongly even after any effects of perception of benefit are statistically removed.

(d) If preventing adverse fitness differentials were the adaptation's *only* function, then after controlling for willingness to participate, any relationship between perceived benefit and punitive sentiment should disappear.

(e) Willingness to participate should predict punishment, but not reward. Punishment is better suited to eliminating adverse fitness differentials than reward is (even though reward remains an effective way of solving problems of labor recruitment). When reward induces a free riding underproducer to join a collective action, this preserves the underproducer's relative fitness advantage compared to the producer design that is doing the rewarding. If redressing adverse fitness differentials created by free riding is an adaptive problem the human mind was designed to solve, then this function should be differentially linked to punishment over reward.

(f) This means that reward sentiments *should not* track punitive sentiments, especially among those most willing to participate. In contrast, if labor recruitment were an important adaptive function of punitive sentiments toward nonparticipants, then reward and punishment sentiments should be correlated, especially among those with the greatest interest in seeing the common goal achieved.

Note that distinguishing between these two functions — participant recruitment and eliminating adverse fitness differentials — would be difficult on the basis of behavioral data alone (especially aggregate data). Behavioral data tell *whether* an individual contributed, free rode, and/or punished, but not why. In contrast, survey data can assess the perceptions and attitudes necessary to test between these two hypotheses.

1.2.3. Both functions

If punitive sentiment toward nonparticipants evolved in the service of both functions — encouraging otherwise reluctant participants and preventing adverse fitness differentials — then it should be correlated with both predictor variables: perceived benefit to the individual of the collective action and that individual's willingness to participate. Moreover, both correlations should stand, even when one controls for any correlation between these two predictor variables.

1.2.4. Byproduct?

It is difficult to make a prediction without knowing which adaptation punitive sentiments toward free riders are supposed to be a byproduct of. In the absence of a specific proposal, the zero-level prediction is a random relationship between functional variables (Tooby & Cosmides, 1989). If punitive sentiment does not track the functional variables described above, then it might track demographic ones instead, such as ethnicity or birth order (as has been suggested by Carroll, Perkowitz, Lurigio, & Weaver, 1987; Davis, Severy, Kraus, &

Whitaker, 1993; Sulloway, 1996). Or, there might be a general appetite to punish wrongdoing, independent of type. For example, one might find that willingness to participate predicts punishment of free riders no more than punishment of other forms of wrongdoing. This would suggest that willingness to participate activates punitive sentiments *in general* rather than ones specifically designed to remove the fitness advantage of free riding.

In contrast, the adaptationist hypotheses above predict no *independent* relationship between demographic variables and punitive sentiment (i.e., unmediated by willingness to participate or perceived benefit). They also predict no particular relationship between punitiveness toward free riders and punitiveness toward criminals or other wrongdoers.

The most widely believed byproduct hypothesis is that behavior and sentiments are generated in accordance with the economic concept of rationality — that is, people have an adaptation that somehow calculates which course of action will maximize their individual payoffs, and choose their behavior and sentiments on this basis. Of course, punishment in collective action contexts is almost always irrational in this sense, because individuals incur costs to punish under circumstances where it is clear to a rational actor that no compensation for these costs could ensue (Fehr & Gächter, 2000a). For this reason, punitive sentiments cannot be explained as the expression of, or as a byproduct of, "rationality," "intelligence," or "rational choice." Nevertheless, we will dissect components of the rational choice hypothesis in Section 4 as well as a more relaxed form of this hypothesis: that people naturally favor the adoption of rules, norms, and incentives that benefit them (whether or not they always behave in a rationally cost-effective manner in support of them).

1.3. Sentiments for rewarding participation in collective action

The goal of encouraging others to participate in a collective action could be achieved either through the carrot or the stick. Above, we focused on how punitive sentiments might help encourage the reluctant to participate. However, researchers have long noted that cooperators "are sometimes motivated by a desire to win prestige, respect, friendship, and other social and psychological objectives" (Olson, 1965, p. 60). This suggests that one way to encourage participation in a collective action is to reward participants in a way that exceeds what they would get from the mere provision of the public good (Andreoni, 1990; Hawkes, 1993; Sober & Wilson, 1998). Many anthropologists have argued that in environments most similar to those in which humans evolved, reward (especially, increased social status) apparently does motivate individuals to provide public goods such as food (Hawkes, 1993; Lemonnier, 1996; Sugiyama, 1996) and military service (Chagnon, 1988; Patton, 1996, 2000; Watson, 1971).

Hence, it seems likely that there are motivational adaptations for providing rewards to those who contribute to collective actions. An obvious design feature is that these pro-reward sentiments should be keyed to the degree to which the sentiment holder is likely to benefit from a collective action. In fact, if recruiting more participation were a selection pressure designing pro-reward sentiments, then one might expect own interest in the goal to predict pro-reward sentiments, even after controlling for one's willingness to participate. For

example, due to injury, ineligibility, family obligations, or other circumstances, an individual may be unable to participate directly in a collective action that would be self-beneficial (Chagnon, 1997). Yet, that individual may be able to increase the endeavor's chances of success by providing rewards to those who can participate, but who otherwise might not.

Our goal herein is to expose the design of the motivational adaptations deployed in collective action. Each adaptive function discussed above implies a different design: Each makes different predictions about the conditions that trigger punitive (and pro-reward) sentiments in collective action. To test these predictions, a survey was conducted as described below.

2. Method

Data were collected by pencil-and-paper survey. Subjects were 18–25-year-old undergraduate US citizens at the University of California, Santa Barbara ($N = 287$), 53% ($n = 152$) of whom were female and 47% ($n = 135$) male. Fifty eight percent took the survey voluntarily in an anthropology class, and 42% were paid US$4 to take it after they approached a campus booth that was set up for the purpose of recruiting subjects. Subjects were asked to report their sex, ethnicity, age, birth order, and annual parental income.

Subjects read two different scenarios describing warfare between the USA and foreign countries (both scenarios are presented fully in the Appendix). The first scenario described the USA mobilizing *defensively* in reaction to a Russian invasion of Alaska. The second scenario described the USA mobilizing *offensively* in order to attack several Middle Eastern countries that had radically increased the price of oil. (Two scenarios rather than one were included so that we could confirm that a significant result in one scenario was not a fluke; both a defensive and offensive scenario were used so that we could detect any signs of "separate psychologies of offense and defense" hypothesized by Tooby and Cosmides, 1988, p. 9.) In both scenarios, subjects were told that the USA was going to have to start drafting citizens in order to have a chance of winning the war.

Following each scenario, subjects were presented with four statements and asked to respond on a 1–7 Likert-like scale from "disagree strongly" to "agree strongly." The first two items were the predictor variables. "If the USA won this war, it would be very good for me as an individual" measured how much subjects perceived that they would benefit from collective success, and will be referred to as SELF-INTEREST IN GROUP GOAL. "If I got drafted for this war, I would probably agree to serve" measured how willing subjects would be to participate in the collective action, and will be referred to as WILLINGNESS TO PARTICIPATE. Next came the two dependent variables. "If a US citizen resisted this draft, I'd think they should be punished" measured punitive sentiment toward nonparticipants, and will be referred to as PUNISH NONPARTICIPANTS. "If a drafted US citizen agreed to serve in this war, I'd think they should be rewarded" measured pro-reward sentiment toward participants, and will be referred to as REWARD PARTICIPANTS. The questionnaire also included a short battery of questions that assess attitudes toward punishment in general (focusing on crime; Carroll et al., 1987), which will be referred to as GENERAL

PUNITIVENESS; whether this battery appeared at the beginning or end was counterbalanced across subjects.

A total of 122 subjects (43%) received surveys in which the order of the dependent variables was reversed, i.e., surveys in which WILLINGNESS TO PARTICIPATE was followed first by REWARD PARTICIPANTS and then by PUNISH NONPARTICIPANTS instead of vice versa. Item order was manipulated in this way to ensure that correlations between WILLINGNESS TO PARTICIPATE and the dependent variables were unaffected by the proximity of the dependent variables to WILLINGNESS TO PARTICIPATE.

3. Results

Two types of correlations are shown in Table 1: simple correlations between each predictor and each dependent variable, and partial correlations between each predictor and each dependent variable when controlling for the effects of the other predictor variable. Also shown are each variable's mean and standard deviation.

3.1. Is there a relationship between a person's willingness to participate in a collective action and that individual's belief that s/he would personally benefit from that action?

Yes. There was a significant and positive correlation between WILLINGNESS TO PARTICIPATE and SELF-INTEREST IN GROUP GOAL in both scenarios [$r_{\text{defensive}} = .379$, $r_{\text{offensive}} = .276$, P's $< .001$ (all reported P values are two-tailed)].

Table 1
Simple and partial correlations in each scenario

	Punish nonparticipants ($M = 2.77$, S.D. $= 1.83$)		Reward participants ($M = 5.22$, S.D. $= 1.71$)	
Defensive scenario	Simple r	Partial r	Simple r	Partial r
Self-interest in group goal ($M = 4.16$, S.D. $= 1.69$)	.296***	.092	.296***	.272***
Willingness to participate ($M = 3.27$, S.D. $= 2.06$)	.597***	.548***	.121*	.010
	Punish nonparticipants ($M = 2.37$, S.D. $= 1.73$)		Reward participants ($M = 4.72$, S.D. $= 1.88$)	
Offensive scenario	Simple r	Partial r	Simple r	Partial r
Self-interest in group goal ($M = 4.23$, S.D. $= 1.88$)	.246***	.092	.369***	.351***
Willingness to participate ($M = 2.56$, S.D. $= 1.86$)	.648***	.623***	.125*	.025

* $P < .05$.
*** $P < .001$.

3.2. Which predicts punitive sentiment better, willingness to participate in a collective action or self-interest in group goal?

WILLINGNESS TO PARTICIPATE and SELF-INTEREST IN GROUP GOAL were both correlated with PUNISH NONPARTICIPANTS. However, the simple correlation for WILLINGNESS TO PARTICIPATE was much higher than that for SELF-INTEREST IN GROUP GOAL (a result that holds for both the offensive and defensive scenarios).

Given that WILLINGNESS TO PARTICIPATE and SELF-INTEREST IN GROUP GOAL were correlated with each other leads to the more interesting question: How much of the variance in punitive sentiment is predicted by each variable, controlling for the other? When one controls for WILLINGNESS TO PARTICIPATE, the correlation between SELF-INTEREST IN GROUP GOAL and PUNISH NONPARTICIPANTS plummeted to .092 in both scenarios (simple r's were .246 and .296, respectively). In contrast, when one controls for SELF-INTEREST IN GROUP GOAL, the correlation between WILLINGNESS TO PARTICIPATE and PUNISH NONPARTICIPANTS remained high (partial r's: $r_{\text{defensive}} = .548$, $r_{\text{offensive}} = .623$, P's $< .001$). Indeed, the partial r's for WILLINGNESS TO PARTICIPATE were almost as high as the simple r's. (N.B. The relationship between willingness to participate and punitive sentiment held for each sex separately; see Section 3.8.)

In short, a large and significant amount of the variance in subjects' punitive sentiment was predicted by their willingness to participate in a collective action, whereas very little of this variance was predicted by their perceived self-interest in the group goal being achieved.

3.3. Does willingness to participate in a collective action predict a desire to punish wrongdoing in general, or does it specifically activate the desire to punish free riders in the collective action?

There was a small but significant positive correlation between WILLINGNESS TO PARTICIPATE and GENERAL PUNITIVENESS ($r_{\text{defensive}} = .148$, $r_{\text{offensive}} = .138$, P's $< .05$). This simple correlation does not answer the question, however, because GENERAL PUNITIVENESS and PUNISH NONPARTICIPANTS were also weakly correlated ($r_{\text{defensive}} = .175$, $r_{\text{offensive}} = .179$, P's $< .05$). Thus, the correlation between WILLINGNESS TO PARTICIPATE and GENERAL PUNITIVENESS could reflect nothing more than the robust correlation reported above between WILLINGNESS TO PARTICIPATE and PUNISH NONPARTICIPANTS.

That appears to be the case. When one statistically controls for the effects of PUNISH NONPARTICIPANTS, the correlation between WILLINGNESS TO PARTICIPATE and GENERAL PUNITIVENESS disappears (partial r's: $r_{\text{defensive}} = .065$, $r_{\text{offensive}} = .035$, n.s.). Moreover, when one statistically controls for the effects of WILLINGNESS TO PARTICIPATE, the relationship between PUNISH NONPARTICIPANTS and GENERAL PUNITIVENESS also dwindled to nonsignificance (partial r's: $r_{\text{defensive}} = .110$, $r_{\text{offensive}} = .124$, n.s.).

In sum, WILLINGNESS TO PARTICIPATE did not predict GENERAL PUNITIVENESS once the effects of PUNISH NONPARTICIPANTS were removed. Instead, willingness to

participate was specifically and selectively associated with a desire to punish nonparticipants in the collective action in which one was participating.

3.4. Which is better predicted by a person's willingness to participate in a collective action: punitive sentiment or a wish to reward?

Consistent with the notion that reward is unsuitable for removing the fitness advantage of free riders, the partial correlation between WILLINGNESS TO PARTICIPATE and REWARD PARTICIPANTS was low and not significant (partial r's: $r_{\text{defensive}} = .010$, $r_{\text{offensive}} = .025$). This is in marked contrast to the partial correlation between WILLINGNESS TO PARTICIPATE and PUNISH NONPARTICIPANTS, which was positive, large, and significant (partial r's: $r_{\text{defensive}} = .548$, $r_{\text{offensive}} = .623$). And indeed, WILLINGNESS TO PARTICIPATE correlated significantly more positively (P's of $\Delta F < .001$) with PUNISH NONPARTICIPANTS ($r_{\text{defensive}} = .597$, $r_{\text{offensive}} = .648$, P's $< .001$) than with REWARD PARTICIPANTS ($r_{\text{defensive}} = .121$, $r_{\text{offensive}} = .125$, P's $< .05$).

There were no order effects. In both scenarios, correlations between WILLINGNESS TO PARTICIPATE and PUNISH NONPARTICIPANTS and WILLINGNESS TO PARTICIPATE and REWARD PARTICIPANTS were unaffected by whether PUNISH NONPARTICIPANTS came before or after REWARD PARTICIPANTS in the survey (P's of ΔF due to item order $> .704$).

3.5. Does labor recruitment cause reward sentiments to track punitive sentiments?

No. If labor recruitment were one function of punitive sentiments, then punitive and reward sentiments should track each other strongly and be predicted by SELF-INTEREST IN GROUP GOAL. But in fact, punitive and reward sentiments were uncorrelated in the defensive scenario ($r = .079$, $P = .186$) and only weakly correlated in the offensive scenario ($r = .165$, $P = .005$). To see whether this latter correlation is a real consequence of labor recruitment adaptations, one must first remove those components of the correlation that are irrelevant to the labor recruitment hypothesis. [Each sentiment (reward and punishment) is triggered by a different predictor variable (see Sections 3.2 and 3.6), but these predictor variables are correlated with one another (see Section 3.1). This could create a spurious correlation between reward and punitive sentiments.]

When this is done, the (already weak) correlation between PUNISH NONPARTICIPANTS and REWARD PARTICIPANTS effectively disappears, showing that it was spurious — a side effect of the correlation between predictor variables. The partial correlations between PUNISH NONPARTICIPANTS and REWARD PARTICIPANTS, controlling for the effects of WILLINGNESS TO PARTICIPATE, approximate to zero (partial r's: $r_{\text{defensive}} = .011$, $P = .82$; $r_{\text{offensive}} = .085$, $P = .061$). Sentiments for punishment and reward were also uncorrelated when the effects of WILLINGNESS TO PARTICIPATE remained, but the effects of SELF-INTEREST IN GROUP GOAL were removed (partial r's: $r_{\text{defensive}} = .008$, $P = .89$; $r_{\text{offensive}} = .08$, $P = .168$) — as one would expect if reward were unsuitable for preventing outcompetition by free riders.

3.6. Does the perception that one will individually benefit from a collective action predict wish to reward?

Yes. SELF-INTEREST IN GROUP GOAL was positively and significantly correlated with REWARD PARTICIPANTS ($r_{\text{defensive}} = .296$, $r_{\text{offensive}} = .369$, P's $< .001$). This correlation remained strong, even after controlling for the effects of WILLINGNESS TO PARTICIPATE (partial r's: $r_{\text{defensive}} = .272$, $r_{\text{offensive}} = .351$, P's $< .001$). Thus, although willingness to participate in a group action does not predict a pro-reward sentiment, perceived self-interest in the outcome of the group action does, and this effect is independent of willingness to participate.

3.7. Were any demographic variables correlated with either punitive sentiment or wish to reward?

None of the demographic variables were useful predictors: neither PUNISH NONPARTICIPANTS nor REWARD PARTICIPANTS was significantly correlated with birth order (firstborns vs. laterborns), ethnicity (non-Whites vs. Whites), age, or annual parental income in either scenario (P's $> .05$).

3.8. Were there any sex differences?

Although sex did not predict REWARD PARTICIPANTS in either scenario (P's $> .186$), females did score significantly lower in PUNISH NONPARTICIPANTS in both scenarios (males coded as 1, females as 2; $r_{\text{defensive}} = -.291$, $r_{\text{offensive}} = -.211$, P's $< .001$). Much of this difference had to do with the fact that females also scored significantly lower in WILLINGNESS TO PARTICIPATE, which, as reported above, predicts PUNISH NONPARTICIPANTS [WILLINGNESS TO PARTICIPATE: female vs. male means: 2.80 vs. 3.80 (defensive), 2.21 vs. 2.96 (offensive); both differences in means are significant (t test for equality of means, P's $< .002$)]. However, even when controlling for WILLINGNESS TO PARTICIPATE, the variance in PUNISH NONPARTICIPANTS predicted by sex was significant in the defensive scenario (partial $r = -.196$, $P = .001$) and marginally so in the offensive scenario (partial $r = -.115$, $P = .054$), that is, women were slightly less motivated to punish nonparticipants, even after controlling for participation. Nevertheless, it is interesting to note that, while sex was somewhat predictive of PUNISH NONPARTICIPANTS, the correlation between WILLINGNESS TO PARTICIPATE and PUNISH NONPARTICIPANTS was strong for both sexes and close to the value when the sexes are combined: .561 for females and .572 for males in the defensive scenario (.597 combined) and .635 for females and .631 for males in the offensive scenario (.648 combined; all P's $< .001$). There was no significant difference between the sexes in the size of these correlations (P's of ΔF due to sex $> .547$). The partial r's for these variables were also similar for both sexes (defensive: male = .558, female = .461, combined = .548; offensive: male = .598, female = .609, combined = .623). (Indeed, there was no significant difference between males and females for any of the 16 cells in Table 1.) Thus, although the scenarios used involved warfare, there was nothing male-specific about the computational connection between participation and punitive sentiment.

4. Discussion

The results were surprisingly clear cut. Subjects' punitive sentiments sensitively tracked their risk of suffering a fitness disadvantage relative to free riders in a collective action and did not track variables suggested by other functions or theories.

4.1. Adaptations for eliminating the free rider fitness advantage: evidence of special design

As any functional model of collective action would predict, the extent to which individuals believed they would benefit from a successful collective action predicted how willing they were to participate in this action. Moreover, both of these variables showed a simple correlation with support for punishing nonparticipants. Note, however, that the simple correlation was much higher for willingness to participate than for perceived self-interest. Moreover, the *only* predictor that made an independent contribution to punitive sentiment was willingness to participate. When we statistically controlled for the effects of perceived self-interest, the correlation between willingness to participate and punitive sentiment remained high (partial r's .548 and .623 compared to simple r's .597 and .648). In contrast, when we statistically controlled for the effects of willingness to participate, the correlation between perceived self-interest and punitive sentiment vanished (partial r's .092 and .092 compared to simple r's .296 and .246).

Precision in the match between the design of a system and a proposed adaptive function is a necessary condition for demonstrating that that system is an adaptation, and for demonstrating what that adaptation's function is (Williams, 1966). An adaptation engineered to prevent a design that causes participation in collective actions from being outcompeted by free riders should key willingness to participate in a collective action precisely to support for the punishment of free riders. That is, the relationship between these variables should be specific and selective. It was. For example:

4.1.1. Selectivity of the participation–punishment link

Although reward and punishment are both time-honored incentive systems, willingness to participate in a collective action predicted the desire to punish free riders but not the desire to reward those who participated. The correlation between willingness to participate and support for rewards was low, and it disappeared once we statistically controlled for the effects of perceived self-interest (partial r's .010 and .025 compared to simple r's .121 and .125). Thus, the effect of willingness to participate on moral intuitions and incentives was selective: It activated punitive sentiments toward free riders, without activating pro-reward sentiments toward participants.

4.1.2. Specificity of the punishment response

The punitive sentiment activated was specific to free riders in a particular collective action. Willingness to participate activated punitive sentiment toward free riders but not punitive sentiments in general (as measured by attitudes toward punishing crimes). Although there was a small correlation between willingness to participate in a collective action and general

punitiveness, this effect disappeared once we statistically controlled for the key variable, that is, the extent to which subjects supported punishment of nonparticipants in a collective action (partial *r*'s .065 and .035 compared to simple *r*'s .175 and .179).

4.1.3. Precision of response

Willingness to participate was the *only* variable that independently predicted the motivation to punish free riders. None of the demographic variables (some of which were suggested by alternative theories of punitive sentiment; Carroll et al., 1987; Davis et al., 1993; Sulloway, 1996) predicted support for punishing free riders. Nor did perceived self-interest, once the effects of willingness to participate were removed.

4.1.4. Uniformity of response

Although women expressed less overall support for punishment than men (even controlling for the fact that they were less likely to be willing to serve in the military), the correlation between willingness to participate in a collective action and support for punishment of free riders was just as strong in women as in men.

In other words, the relationship between willingness to participate in a collective action and desire to punish free riders was specific, selective, and uniform. This evidence of special design suggests the presence of an adaptation that was designed for eliminating the adverse fitness differentials that producers would otherwise incur relative to free riders in collective action contexts.

4.2. Is there evidence that punitive sentiment was also designed to encourage participation?

There was no support in these data for the hypothesis that punitive sentiment toward nonparticipants was *designed* to encourage the participation of others in a collective action. Punishment may sometimes have the *effect* of encouraging reluctant non-free riders to participate. But in this research, we are attempting to apply Williams' (1966) adaptationist program: We are trying to distinguish between an adaptation's design features and any incidental side effects it might have, whether beneficial or not.

If punitive sentiment were designed to encourage participation, one would expect the extent to which individuals perceive a collective action to be in their self-interest to be correlated with their degree of punitive sentiment. This was not the case: There was no correlation between individuals' self-interest in a group goal and their punitive sentiment, once we controlled for the effects of their willingness to participate.

That encouraging participation is a (beneficial) byproduct of punishment rather than its primary adaptive function receives further support from experimental economics. Fehr and Gächter (2000a) have shown that the possibility of punishment does encourage higher levels of participation in public goods games but only by discouraging contributions less than the average of the other group members (see also Kurzban et al., 2001). Individuals were punished to the extent that they contributed less than their "fair share" (i.e., the average contribution of other group members); but they were rarely punished for contributions that

were at or above that group average. This result is clearly consistent with the function of preventing outcompetition by free riders. It is not, however, what one would expect if the function of punitive sentiments was to recruit optimal participation.

If motivating optimal participation were the selection pressure responsible for designing punitive sentiments, then one should feel punitive toward anyone contributing below the optimum. In Fehr and Gächter's games, an individual benefits the most when others contribute everything. Thus, each individual should view the optimum level of participation by others as the maximum contribution. Nevertheless, there was no pattern of punishing all or even most deviations from this optimum. Punishment was rarely delivered upon those who had contributed their "fair share," even though this amount was almost always well below the optimal contribution.

Logically, why should "fair share" contributors be exempt from punishment? The optimal recruitment function predicts that subjects will feel punitive even toward individuals who are contributing their fair share, as long as they would benefit by these individuals contributing even more. There is no intrinsic reason that circuitry well designed for optimal recruitment would block punitive sentiments toward a person whose contribution is well below the optimum but above whatever the group average happens to be.

If future research confirms this result, then it would show that the trigger for punishment lies in making contributions lower than the average. But this specific trigger point emerges more naturally in an analytical sense from the function of preventing outcompetition by free riders than from the function of encouraging optimal contributions. An individual cannot be a fitness free rider with respect to the group if s/he is contributing at or above the average level[2], but can easily be contributing suboptimally. Although many notions of fairness are conceptually possible (and have been proposed by social theorists), we suggest that terms such as "fair share" and "exploitation" derive their special psychological resonance and rhetorical power from their connection to the evolved circuitry that defends against the threat of adverse fitness differentials posed by free riders.

In the absence of the ability to punish, reducing one's own contribution is the only defense left against outcompetition by free riders. We think that this motivation to reduce contributions is, like punishment, an evolved defense against the proliferation of free rider designs. Indeed, experiments where punishment is not an option show that participants who encounter free riders do indeed reduce their own contributions to the public good (Fehr & Gächter, 2000a; Kurzban et al., 2001). But in a world where punishment is impossible, an inevitable byproduct of this defense is that cooperation — which starts out fairly high in

[2] To switch from a group to an individual frame, the criterion for developing a punitive sentiment toward person *j* should be "did person *j* contribute less than me," given that person *j* sufficiently benefits from the group action. If other criteria need to be met (e.g., sufficient social support for punishing an individual) before the sentiment is prudently expressed in punitive action, then this would appear as punishment of those below the *group* average. If the function is redressing adverse fitness differentials, then the motivation should be proportionate to the fitness differential. Assuming that differences in benefit vary randomly across events, this may often reduce to the difference between own contribution and the other's contribution.

public goods games — will eventually unravel (Kurzban et al., 2001). In contrast, cooperation can be sustained when punishment is possible because one can defend against free riders without reducing one's own contribution. This provides a parsimonious explanation for the fact that participation in public goods games (and the net benefit received by each individual) is higher when punishment is possible (Fehr & Gächter, 2000a; Kurzban et al., 2001) — without invoking the additional hypothesis that punitive sentiment evolved to encourage participation.

There is considerable a priori plausibility to the hypothesis that punitive sentiments evolved to serve both participation encouragement and free rider fitness reduction functions, and such a view may ultimately be vindicated in a more comprehensive research program. The existence of some punitive system for labor recruitment would plausibly be adaptive, and may be revealed by new methods. (It may be that while inverting adverse fitness differentials is the primary function, the cost of the punishment system is to some extent offset by labor recruitment. If this were true, punitive sentiments should be more easily acted upon when they also serve a labor recruitment function.) Nevertheless, taken as a whole, the data discussed cast doubt on the hypothesis that the motivation to punish free riders was designed, even secondarily, to accomplish labor recruitment. When the two functions are analyzed separately, the available evidence only provides support for the hypothesis that the motivation to punish free riders was designed for preventing the emergence of fitness advantages for free riders over contributors.

4.3. Evidence of a separate adaptation for producing pro-reward sentiments

This does not, however, mean that there are no adaptations for encouraging participation. Indeed, the data suggest that sentiments for rewarding participants in a collective action may have exactly this function. The perception that one will individually benefit from a successful collective action did predict support for rewarding participants, and this was true even after we statistically controlled for willingness to participate (partial r's .272 and .351 compared to simple r's .296 and .369). Yet, willingness to participate in a collective action did not predict support for rewarding participants, once we controlled for self-interest in seeing the goal achieved. This is consistent with the notion that people resort to positive incentives for participation when they have an interest in the goal being achieved yet cannot (or will not) participate in the collective action themselves.

The hypothesis that there are two independent adaptations at work — punishment circuitry designed to defend against free riders and a reward sentiment designed to encourage participation — is supported by the dissociation displayed in Fig. 1.

4.4. Rational choice fails to explain the design of punitive sentiments in collective action contexts

Evidence that people's sentiments and behavior appear well designed to solve an adaptive problem sometimes elicits the following response: *It is implausible and unparsimonious to argue that there are specialized adaptations for this purpose; instead, people just figure out*

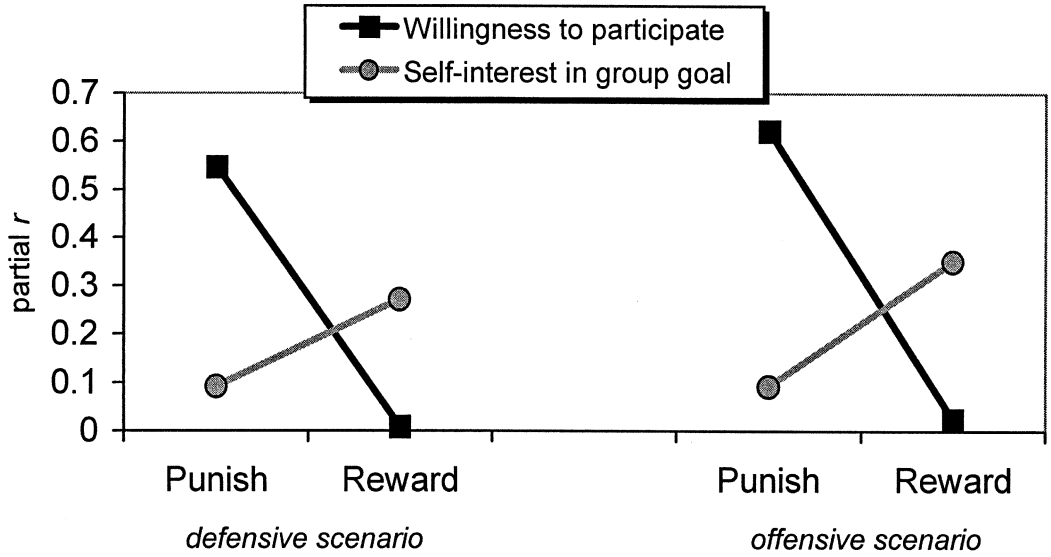

Fig. 1. Partial *r*'s indicate the effect of each predictor variable — controlling for the other — on each outcome variable. The crossover indicates a clear dissociation. Punitive sentiment is independently predicted by willingness to participate, but not by self-interest in the group goal. Pro-reward sentiment is independently predicted by self-interest in the group goal, but not by willingness to participate. This suggests that, in a collective action context, pro-reward sentiments and punitive sentiments are generated by two different adaptations.

which sentiments and choices are in their best interest. While seductive, this proposal is problematic as an alternative hypothesis. It fails to deliver on precisely what is at issue: What is the structure of the neurocognitive machinery whereby individuals "figure it out"? Consequently, it is too ill-defined to count as a proper alternative hypothesis about mechanisms. In fact, most rational choice style theories posit a form of unbounded rationality that is impossible to implement computationally (Cosmides & Tooby, 1987; Gigerenzer & Selten, 2001).

Nevertheless, it is illuminating to see how far one can go in empirically evaluating this family of hypotheses. This can be done by analyzing rational choice theory not as a (hopelessly) vague proposal about architecture but rather as a task analysis of what a general payoff-maximizing architecture, free of inherited specializations and defaults, should be designed to do. Rational choice theories posit that there is computational machinery that is able to assume whatever form is necessary to maximize individual payoffs, and which chooses (or deploys) sentiments (or other mental entities, such as decision rules) on the basis of whether they accomplish this function. Most critically, it does this solely on the basis of incentives and information available to the individual over the course of the lifespan: By hypothesis, this decisional machinery lacks inherited, evolved specializations and so does not make default assumptions that were ancestrally reliable (but may now be violated) about the structure of environments and likely payoffs. This alternative *can* be evaluated, because a system designed to maximize individual payoffs in current collective action contexts

would predict different sets of sentiments than one designed to eliminate adverse fitness differentials under ancestral conditions. This is true whether one assumes punitive sentiments operate by motivating the individual who has them (1) to personally inflict punishment or (2) to advocate a general regime of norms that are applied to everyone, including oneself. Indeed, we can also evaluate a more relaxed form of the rational choice viewpoint — that is, that these results are explained by the hypothesis that people favor acts, rules, and norms that are in their interest, even if they do not always do this in a rationally cost-effective manner.

The results that decide against the view that punitive sentiments evolved to optimize participant recruitment also falsify the view that they are produced by a general-purpose rational choice system. More specifically:

1. A rational agent should only punish noncooperation when the payoffs to that agent in terms of the benefits of increased future cooperation exceed the costs to that specific agent of punishing. This seems unlikely to be true in the great majority of cases, for example, when the social group is large, as it is in most residential groups, coalitions, political units, and so on (modern and ancestral). Testing the sensitivity of subjects to this condition, experimental economists have established that people regularly incur costs to punish free riders, even when they know they will not interact with these individuals again. The absence of future interactions means punishing others will not gain them benefits in the form of higher levels of future experienced cooperation, and so punishers in these contexts do not recoup the costs they expended on punishment (Fehr & Gächter, 2000a; Gintis, 2000). This form of spite, expressed even on the last move of an interaction, cannot be explained by a desire to maximize individual payoffs. It can, however, result from adaptations designed to reverse adverse fitness differentials and prevent outcompetition by free riders. In this case, the issue is preventing the malefactor from getting away with his or her unfair winnings. In contrast, for rational choice, the issue is maximizing present and future payoffs — and payoffs to the individual, not payoffs to one design relative to another. [Experimental economists have established that spiteful punishment also occurs in dyadic cooperation (Hoffman, McCabe, & Smith, 1998), suggesting that reversing adverse fitness differentials may be a significant factor in the evolution of dyadic reciprocation as well.]

2. Leaving aside the irrationality of incurring unreimbursed costs of punishment, rational choice predicts that the targets of punitive sentiments should be those individuals who could increase their levels of cooperation to the benefit of the rational agent. Clearly, this includes all individuals who contribute less than the optimum, including those who contribute at or above the group average. Yet, as discussed above, subjects in public goods games rarely punished contributions in this zone. This otherwise puzzling result makes sense if punitive sentiments were instead designed to eliminate ancestral fitness advantages of free rider designs rather than to maximize joint payoffs.

3. Most critically, if punitive sentiments were produced by a system designed to maximize individual payoffs, then rational choice theory straightforwardly predicts that self-interest in the group goal should regulate punitive sentiments. This is the most basic prediction, even for the relaxed form of the rational choice approach. However, in this study, self-interest in the group goal did not independently predict punitive sentiments.

4. Moreover, willingness to participate did independently predict punitive sentiments. Yet rational choice does not offer a good explanation for why participation per se rather than self-interest creates punitive sentiment toward nonparticipants.[3] On the contrary: To base punitive sentiments on willingness to participate *independent of expected gain* is to operate independent of self-interest, in a way that parallels the sunk cost fallacy. Sentiments are rationally deployed only as a function of one's expected gain from the project's success, and not as a function of how much effort one has or will expend, regardless of expected gains. In contrast, this relationship between participation and punitive sentiment is a direct prediction of the hypothesis that the function of punitive sentiments is to eliminate the fitness advantage of free riders.

5. Providing incentives to increase other people's participation is costly. However, a system designed to maximize individual payoffs should (aside from issues of cost-effectiveness) be intrinsically indifferent to whether a unit of cost is deployed to punish nonparticipants or to reward participants. This suggests that punishment and reward sentiments should track each other, and be triggered by the same variables. Neither was true in our data.

In interpreting our scenarios, subjects may have assumed that the punishments and rewards would be dispensed by the government. Can any of the difficult results noted above be explained away by assuming that the sentiments produced by a rational choice system operate by motivating one to advocate a general regime of norms?

The norms advocated would have to be self-beneficial, even though they are applied to everyone, including oneself. If self-beneficial norm establishment is the function of punitive sentiments, then perhaps nonparticipants should avoid advocating punishment, lest this community-enforced punishment norm be visited on them. On first pass, this seems to explain one otherwise puzzling finding: that punitive sentiments are high only in those willing to participate. On closer inspection, however, it becomes clear that this line of argument cannot rescue the rational choice view. The following results argue against the notion that sentiments for punishment and reward are produced by a system designed to advocate norms that are self-serving in the rational choice sense:

6. The self-serving norms approach predicts that punitive sentiment will be high only in individuals who feel they would benefit from the group goal's attainment *and* are willing to participate. That is, the two predictor variables should predict high punitive sentiment only when they co-occur (there should be an interaction, but no main effects). This was not the case: after all, willingness to participate predicted punitiveness *independent of self-interest in the group goal*, and there was no two-way interaction (P's $>.50$). (This finding was not due to high covariance between the predictor variables in the data set: The correlation between self-interest in group goal and willingness to participate was only .28–.38).

[3] At least insofar as punitive sentiments are designed to motivate the personal infliction of punishment on others, rather than norm creation. A motivation to punish others would not punish itself, and so could easily be designed to punish others for nonparticipation while being a nonparticipant.

7. The norm approach presumes that those who benefit from a public good are better off when there is a punishment norm and will therefore advocate one, unless by doing so they risk being punished themselves. Note, however, that if others believe you are exempt from participation, then you do not risk advocating your own punishment by supporting the punishment of free riders. The correlation between willingness to participate and punitive sentiment should therefore be lower — or even nonexistent — for those who are exempt: Exempt nonparticipants who will benefit from a public good can — and should — advocate punishment of nonexempt nonparticipants. But our data do not support this prediction. In the USA, women have traditionally been exempt from military conscription. Female subjects should therefore have a greater rational expectation that they will be exempt from participating in a war as soldiers than men will, i.e., that they will not be punished for being nonparticipants. Yet, the correlation between participation and punitive sentiment was just as strong in female subjects as in male subjects. This undermines the self-serving norm explanation. In contrast, this result makes sense if women's responses are not based on incentives in the modern world, but instead reflect a psychological design that ancestrally prevented out competition by free riders.

Indeed, although not probed in this study, it seems very likely that our psychology of collective action recognizes the distinction between free riding and nonparticipation for excusable reasons. The self-serving norm theory predicts that, to the extent that participants and the exempt both benefit from the provision of a public good, they should experience equal intensities of punitiveness toward free riders. In contrast, the hypothesis that punitiveness is a defense for participant designs against outcompetition by free riders predicts that participants — but not nonparticipants (exempt or not) — should experience the most intense punitive sentiment.

8. The notion that the mind contains a system that deploys sentiments insofar as they create payoff-maximizing norms is most damaged by the results pertaining to pro-reward sentiments. If such a system existed, it would surely use degree of participation to trigger sentiments for rewarding participants (whether or not it uses participation as a trigger for punitive sentiments). Those willing to participate should favor community rewards for participation because this norm is in their self-interest: They would be among the beneficiaries, thereby receiving payoffs above and beyond the provision of the public good itself. Yet willingness to participate did not independently predict pro-reward sentiment. In contrast, this otherwise puzzling result is a prediction of the outcompetition prevention function. If eliminating adverse fitness differentials is the dominating adaptive problem in collective action contexts, then willingness to participate should inhibit the desire to reward those who contribute, but contribute less. This is because transferring rewards from some participants to participants who contribute less preserves or even amplifies the fitness differentials accruing to free riders.[4]

[4] We predicted no relationship, rather than a negative one, between willingness to participate and pro-reward sentiment because in this study the rewards are coming out of a communal pocket rather than the individual's. This lessens the cost incurred by an individual of providing rewards; nevertheless, the provision of rewards to encourage more participation still preserves the relative fitness advantage of free riders.

In short, the rational choice hypothesis fails, in eight different ways, to explain the pattern of punitive and pro-reward sentiments elicited by collective action contexts in this and other studies.

4.5. Will the results of this study generalize?

Although the results reported here are robust and unambiguous, it will be important to see how well they generalize to other collective action contexts. This survey used a collective action problem involving a nation-state, because in the population tested, this was a common point of reference and also likely to elicit enough variation in willingness to participate to see whether willingness would predict punitive sentiment. Obviously, however, the hypothesis should be tested among people living in conditions more representative of the ancestral social environments in which these adaptations evolved (e.g., small-scale or foraging societies). If punitiveness toward nonparticipants operated as an anti-free rider device in ancestral collective action contexts, then its operation as such should be clearly observable in social environments that resemble those of the ancestral past more closely. A study of this kind is underway in the Ecuadorian Amazon, combining survey data with observations of behavioral choices and outcomes. The results so far are highly consistent with those presented herein.

Some argue that behavioral data are inherently reliable, while survey data — even when responses are anonymous — are worthless because subjects may lie. Our own view is that both kinds of data are valuable and complementary. Indeed, we hope the present study illustrates how survey data can fill in the blanks left by aggregate behavioral data and vice versa. There are several problems with the view that our results could be explained as deception. First, it carries no directional prediction; at best, it would introduce noise in the data (for the questions we asked, there is no consensual, "socially desirable" answer). Second, because we are interested in whether an increase in one variable is associated with an increase in another (controlling for a third!), subjects would have to be clairvoyant in order to deceive *in just the right way* that their answers would mesh with those of others to create the partial correlations found. Third, the possibility of lying or other forms of invalidity does not distinguish surveys from behavioral data in experimental games: People can "lie" in experimental games, that is, behave playfully, maliciously, or unrepresentatively, in the sense of making choices they would not make in real life. The idea that people indulge in these motivations and values in surveys but not in experimental games because of payments — often quite small — seems unlikely, given the large sums people routinely pay to indulge these motivations in other contexts. [Moreover, experienced experimentalists know all too well that subjects often respond to narrow features of the experimental situation differently than they would to those features of the world to which the experimenters imagine the task corresponds (Hoffman et al., 1998).] Even worse, in an experimental game — or in real life — behavioral indices of the predictor variables are hopelessly confounded: One's degree of participation should be correlated with one's self-interest in the group goal, preventing the experimenter from discovering whether punitive sentiment is triggered by one variable but not the other. In short, different approaches are complementary and powerful when combined,

with valid conclusions emerging from converging lines of evidence developed from applying a diversity of methods.

4.6. Can the results help decide between alternative selectionist theories of the evolution of collective action?

Evidence of special design in an adaptation can sometimes allow one to choose between alternative theories about selection pressures. Although the results presented herein are broadly consistent with both individual and group selection theories, it does seem fair to say that the fit between design and selectionist model is tighter in most respects for individual than group level theories.

Most theories of the evolution of collective action that emphasize selection at the individual level *require* adaptations that adjust an individual's willingness to punish free riders so that it reflects that individual's risk of being outcompeted by them (Tooby & Cosmides, 1988, 1996). Such theories assume that humans evolved in small social groups in which they would have had repeated interactions with the same individuals and that adaptations for collective action will reflect this fact, generating behavior that, while perhaps not adaptive in modern nation states or experimental laboratories, would have been adaptive ancestrally (for an application of this line of reasoning to results in experimental economics, see Hoffman et al., 1998).

Some group selection models also assume that punishment will be delivered by those who cooperate; the "strong reciprocity" model proposed by Gintis (2000) is an example. However, this model requires that *group* benefits outweigh *group* costs and assumes that strong reciprocators cooperate and punish even when this involves a penalty to their within-group fitness. Indeed, the model as presented only allows for discrete, and not graded, responses (i.e., cooperate or not). Thus, it does not require adaptations that calibrate *how much* an individual is willing to punish to the *amount* of his or her own personal sacrifice. Moreover, while the Gintis model does assume that cooperation and punishment are provided by the same individual at the phenotypic level, it is not clear that his or other group selection models require this. As the case of warriors and workers within the same species of ant illustrates, a uniform genotype can give rise to different phenotypic morphs, each with a differentiated function. Thus, for one group to outcompete another in fitness production, such models require that the group includes both cooperators and punishers, but it is not clear that these functions need to be localized within the *same* individuals.

One fundamental question is left unaddressed in this analysis: Why don't second-order free riders (i.e., those who do not punish free riders) proliferate at the expense of those equipped with punitive sentiments? This will be addressed in further work.

The results presented herein are suggestive, not conclusive. Future studies, providing more detailed information about the design of adaptations for collective action, may be better able to adjudicate between different selectionist theories. Nevertheless, a complex pattern of evidence has been developed which supports the hypothesis that punitive sentiments in collective action contexts evolved to reverse the fitness advantages that accrue to free riders

over producers. This suggests that an adaptationist analysis of the psychology of collective action may prove illuminating.

5. Conclusion

In sum, we think that the existing evidence, on balance, supports the following conclusions: There is a suite of adaptations that evolved to allow individuals to benefit from engaging cooperatively in collective actions. This suite includes a motivational subsystem that produces punitive sentiments specifically targeted at free riders. The primary function of this circuitry is to prevent free rider designs from having higher fitness than cooperator designs. Moreover, the regulatory effects of the variable underlying punitive sentiments appear narrowly tailored to this anti-free rider function: Willingness to participate did predict punitive sentiments, but did not predict support for rewarding collective action participants or for punishment outside a specific collective action context. This punitive subsystem lacked evidence of special design for alternative adaptive functions, such as optimizing participation in collective actions. While the secondary effect this system sometimes has of increasing contributions to collective efforts may be one way the system offsets its cost, there is no evidence that it is the primary function of punitive sentiments. However, existing evidence supports the view that a separate motivational subsystem evolved to solve the problem of labor recruitment through pro-reward rather than punitive sentiments. Finally, despite the consistency of these results, it nevertheless remains possible that experimenting with other contexts will elicit evidence of a motivational subsystem that is designed to deploy punitive sentiments in order to recruit labor into collective actions.

Acknowledgments

We gratefully acknowledge Don Brown, Napoleon Chagnon, Nancy Collins, Brad Duchaine, Adam Fox, Ed Hagen, Nickie Hess, Rob Kurzban, Hassan Lopez, David Price, Don Symons, and Tina Wells for their kind assistance. This research was supported by a Jacob K. Javits Fellowship from the U.S. Department of Education, by the James S. McDonnell Foundation, the National Science Foundation (#BNS9157-449), the Harry Frank Guggenheim Foundation, and the UCSB Office of Research (Research Across Disciplines Program: Evolution and the Social Mind).

Appendix

Defensive scenario: Imagine that a few years from now, the Russian people elect a new, warlike dictator who claims that Alaska should rightfully belong to Russia. Under this dictator, Russia invades and conquers Alaska. There is good evidence that Russia also intends to conquer more US territory, in addition to Alaska. In response to this invasion, the USA

declares war on Russia. But because this war was unexpected, the USA has allowed its army to get relatively small, and it must start drafting US citizens in order to have a chance of winning this war. How would you feel about this war?

Offensive scenario: Imagine that a few years from now, several oil-rich Middle Eastern countries get together and decide that to increase profits, they will dramatically raise the price of their oil. This price increase devastates US industry and causes high inflation in the USA. US gas prices triple, and several US oil companies go bankrupt. After talks with these Middle Eastern countries fail, the USA declares war on them. But war was unexpected, so the USA has allowed its army to get relatively small, and it must start drafting US citizens in order to have a chance of victory. How would you feel about this war?

References

Alexander, R. D. (1987). *The biology of moral systems*. Hawthorne, NY: Aldine de Gruyter.

Andreoni, J. (1990). Impure altruism and donations to public goods — a theory of warm-glow giving. *Economic Journal*, *100*, 464–477.

Andreoni, J. (1995). Cooperation in public goods experiments: kindness or confusion. *American Economic Review*, *85*, 891–904.

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

Carroll, J. S., Perkowitz, W. T., Lurigio, A. J., & Weaver, F. M. (1987). Sentencing goals, causal attributions, ideology, and personality. *Journal of Personality and Social Psychology*, *52*, 107–118.

Chagnon, N. (1988). Life histories, blood revenge, and warfare in a tribal population. *Science*, *239*, 985–992.

Chagnon, N. (1997). *The Yanomamo: case studies in cultural anthropology*. New York: Harcourt.

Cosmides, L., & Tooby, J. (1987). From evolution to behavior: evolutionary psychology as the missing link. In: J. Dupre (Ed.), *The latest on the best: essays on evolution and optimality* (pp. 277–306). Cambridge, MA: MIT Press.

Cronk, L. (1994). Evolutionary theories of morality and the manipulative use of signals. *Zygon*, *29*, 81–101.

Davis, T. L., Severy, L. J., Kraus, S. J., & Whitaker, J. M. (1993). Predictors of sentencing decisions: the beliefs, personality variables, and demographic factors of juvenile justice personnel. *Journal of Applied Social Psychology*, *23*, 451–477.

Dawes, R. M., Orbell, J. M., & Van de Kragt, J. C. (1986). Organizing groups for collective action. *American Political Science Review*, *80*, 1171–1185.

Fehr, E., & Gächter, S. (2000a). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*, 980–994.

Fehr, E., & Gächter, S. (2000b). Fairness and retaliation: the economics of reciprocity. *Journal of Economic Perspectives*, *14*, 159–181.

Frank, R. (1988). *Passions within reason: the strategic role of the emotions*. New York: Norton.

Gigerenzer, G., & Selten, R. (Eds.) (2001). *Bounded rationality: the adaptive toolbox*. Cambridge, MA: MIT Press.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, *206*, 169–179.

Gurven, M., Hill, K., Kaplan, H., Hurtado, A., & Lyles, R. (2000). Food transfers among Hiwi foragers of Venezuela: tests of reciprocity. *Human Ecology*, *28*, 171–218.

Hardin, G. (1968). The tragedy of the commons. *Science*, *162*, 1243–1248.

Harner, M. J. (1984). *The Jívaro, people of the sacred waterfalls*. Berkeley: University of California Press.

Hawkes, K. (1993). Why hunter-gatherers work — an ancient version of the problem of public goods. *Current Anthropology*, *34*, 341–361.

Henrich, J., & Boyd, R. (2001). Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, *208*, 79–89.

Hirshleifer, D., & Rasmusen, E. (1989). Cooperation in a repeated prisoners dilemma with ostracism. *Journal of Economic Behavior and Organization*, *12*, 87–106.

Hirshleifer, J. (1987). On the emotions as guarantors of threats and promises. In: J. Dupre (Ed.), *The latest on the best: essays on evolution and optimality* (pp. 307–326). Cambridge, MA: MIT Press.

Hoffman, E., McCabe, K., & Smith, V. (1998). Behavioral foundations of reciprocity: experimental economics and evolutionary psychology. *Economic Inquiry*, *36*, 335–352.

Holmberg, A. (1969). *Nomads of the long blow: the Siriono of eastern Bolivia*. Garden City, NY: Natural History Press.

Irons, W. (1991). How did morality evolve? *Zygon*, *26*, 49–89.

Keeley, L. H. (1996). *War before civilization: the myth of the peaceful savage*. Oxford: Oxford Univ. Press.

Kelly, R. L. (1995). *The foraging spectrum: diversity in hunter-gatherer lifeways*. Washington: Smithsonian Institution Press.

Kurzban, R., McCabe, K., Smith, V., & Wilson, B. (2001). Incremental commitment in a real-time public goods game. *Personality and Social Psychology Bulletin*, *27* (12), 1662–1673.

Ledyard, J. (1995). Public goods: a survey of experimental research. In: J. H. Kagel, & A. E. Roth (Eds.), *The handbook of experimental economics* (pp. 111–194). Princeton, NJ: Princeton Univ. Press.

Lee, R., & DeVore, I. (Eds.) (1968). *Man the hunter*. Chicago: Aldine.

Lemonnier, P. (1996). Food, competition, and the status of food in New Guinea. In: P. W. Wiessner, & W. Schiefenhövel (Eds.), *Food and the status quest: an interdisciplinary perspective*. Providence: Berghahn Books.

Olson, M. (1965). *The logic of collective action: public goods and the theory of groups*. Cambridge: Harvard Univ. Press.

Ostrom, E. (1998). A behavioral approach to the rational choice theory of collective action. *American Political Science Review*, *92*, 1–22.

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: self-governance is possible. *American Political Science Review*, *86*, 404–417.

Patton, J. Q. (1996). *Thoughtful warriors: status, warriorship, and alliance in the Ecuadorian Amazon*. Doctoral dissertation, Department of Anthropology, University of California, Santa Barbara.

Patton, J. Q. (2000). Reciprocal altruism and warfare: a case from the Ecuadorian Amazon. In: L. Cronk, N. A. Chagnon, & W. Irons (Eds.), *Adaptation and human behavior: an anthropological perspective* (pp. 417–436). New York: Aldine de Gruyter.

Sato, K. (1987). Distribution of the cost of maintaining common resources. *Journal of Experimental Social Psychology*, *23*, 19–31.

Smith, E. A. (1985). Inuit foraging groups: some simple models incorporating conflicts of interest, relatedness, and central-place sharing. *Ethology and Sociobiology*, *6*, 37–57.

Smole, W. J. (1976). *The Yanoama Indians: a cultural geography*. Austin: University of Texas Press.

Sober, E., & Wilson, D. S. (1998). *Unto others: the evolution and psychology of unselfish behavior*. Cambridge: Harvard Univ. Press.

Stern, P. C. (1995). Why do people sacrifice for their nations? *Political Psychology*, *16*, 217–235.

Sugiyama, L. S. (1996). *In search of the adapted mind: a study of human cognitive adaptations among the Shiwiar of Ecuador and the Yora of Peru*. Doctoral dissertation, Department of Anthropology, University of California, Santa Barbara.

Sulloway, F. (1996). *Born to rebel*. New York: Pantheon.

Tooby, J., & Cosmides, L. (1988). *The evolution of war and its cognitive foundations*. Institute for Evolutionary Studies Technical Report 88-1. Reprinted in Tooby, J. & Cosmides, L. (in press). *Evolutionary psychology: foundational papers*. Cambridge, MA: MIT Press.

Tooby, J., & Cosmides, L. (1989). The innate versus the manifest: how universal does universal have to be? *Behavioral and Brain Sciences*, *12*, 36–37.

Tooby, J. & Cosmides, L. (1996). Groups in mind: the evolution of cognitive adaptations for coalitions and status. *Ciba Foundation*, London, England. Symposium #208: *Characterizing human psychological adaptations*.

Watson, J. B. (1971). Tairora: the politics of despotism in a small society. In: R. M. Berndt, & P. Lawrence (Eds.), *Politics in New Guinea* (pp. 224–275). Nedlands: University of Western Australia.

Williams, G. C. (1966). *Adaptation and natural selection*. Princeton: Princeton Univ. Press.

Williams, G. C. (1975). *Sex and evolution*. Princeton: Princeton Univ. Press.

Wrangham, R., & Peterson, D. (1996). *Demonic males: apes and the origins of human violence*. Boston: Houghton Mifflin.

Wright, R. (1994). *The moral animal*. New York: Pantheon.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, *51*, 110–116.

Yamagishi, T. (1992). Group size and the provision of a sanctioning system in a social dilemma. In: W. Liebrand, D. Messick, & H. Wilke (Eds.), *Social dilemmas: theoretical issues and research findings* (pp. 267–287). Oxford: Pergamon.