MANAGERIAL AND DECISION ECONOMICS
Manage. Decis. Econ. **27:** 103–129 (2006)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/mde.1287

Cognitive Adaptations for *n*-person Exchange: The Evolutionary Roots of Organizational Behavior

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

[a] *Center for Evolutionary Psychology, Department of Anthropology, University of California, Santa Barbara, CA 93106-3210, USA*
[b] *Indiana University Workshop in Political Theory and Policy Analysis and The Santa Fe Institute, Indiana University, 513 N. Park, Bloomington, IN 47408-3895, USA*
[c] *Olin School of Business, Washington University in St. Louis, Campus Box 1133, 1 Brookings Drive, St. Louis, MO, 63130-4899, USA*

Organizations are composed of stable, predominantly cooperative interactions or *n*-person exchanges. Humans have been engaging in *n*-person exchanges for a great enough period of evolutionary time that we appear to have evolved a distinct constellation of species-typical mechanisms specialized to solve the adaptive problems posed by this form of social interaction. These mechanisms appear to have been evolutionarily elaborated out of the cognitive infrastructure that initially evolved for dyadic exchange. Key adaptive problems that these mechanisms are designed to solve include coordination among individuals, and defense against exploitation by free riders. Multi-individual cooperation could not have been maintained over evolutionary time if free riders reliably benefited more than contributors to collective enterprises, and so outcompeted them. As a result, humans evolved mechanisms that implement an aversion to exploitation by free riding, and a strategy of conditional cooperation, supplemented by punitive sentiment towards free riders. Because of the design of these mechanisms, how free riding is treated is a central determinant of the survival and health of cooperative organizations. The mapping of the evolved psychology of *n*-party exchange cooperation may contribute to the construction of a principled theoretical foundation for the understanding of human behavior in organizations. Copyright © 2006 John Wiley & Sons, Ltd.

EVOLUTIONARY FORMS OF RATIONALITY: THE EVOLUTIONARY AND COGNITIVE BACKGROUND TO EXCHANGE AND MULTI-INDIVIDUAL COOPERATION

Voluntary exchange for mutual benefit improves the net welfare of participants. Consequently, traditional economic and social science theories often treat exchange simply as the straightforward product of human rationality plus self-interest, transparently based on general-purpose abilities to reason, pursue goals and select the highest payoffs. Although collective action has a similar ability to improve the net welfare of participants, economists and organizational theorists have long realized that there was a problem in attempting to explain collective action as a parallel expression of rationality plus self-interest (Olson, 1965; Ostrom, 1998; Price *et al.*, 2002). In *n*-party interactions that produce public and common goods (that is, produce outcomes where individual effort is initially unlinked to the ability to consume the benefits of a joint effort), free riding provides higher payoffs than contributing. As a result,

*Correspondence to: Center for Evolutionary Psychology, Department of Anthropology, University of California, Santa Barbara, CA 93106-3210, USA. E-mail: tooby@anth.ucsb.edu

rational agents—who according to standard economic theory will be essentially everyone—will choose not to contribute. Because of this free rider problem, it is puzzling why any rational, self-interested individual would contribute at all to a collective action. Yet, across all known cultures, throughout history, humans have commonly engaged in collective actions (Ostrom, 1990; Price *et al*., 2002; Tooby and Cosmides, forthcoming). Where interactions are small in scale, joint efforts are routine. How can this be?

Locating Rationality, Choice, and Organizational Behavior in an Evolutionary Psychological Framework

Although contributing to collective actions appears to violate standard rationality and the theory of choice behavior that emerges from it, rationality itself appears less straightforward when the attempt is made to locate formal principles of rationality and economic choice in the causal matrix of the physical world. First, rationality and choice are not, of course, mechanism-free, but are real solely to the extent they are embodied in the information-processing architecture of neural programs in the human brain. Economic and organizational behaviors do not flow from disembodied principles, but are the computed outputs of structured cognitive mechanisms incarnated in brain organization. So, cognitive science is one foundation needed to understand human economic and organizational decision-making. Second, species-typical mechanisms of reasoning and choice, in whatever form they exist, came into being because they (or their developmental bases) were produced by the evolutionary process. So, evolutionary biology provides a second foundation for understanding human decision-making.

Evolutionary psychology unites the two projects of evolutionary functionalism and cognitive science into a single integrated research program (Tooby and Cosmides, 1992a). It explicitly uses knowledge of what natural selection would have favored during human evolution to guide the empirical mapping of the computational architecture of the human mind/brain. Thus, it asks two sets of interrelated, complementary, and mutually illuminating questions. First, what are the actual computational procedures that the human mind/brain embodies in its reasoning, emotion, and motivational systems? Second, which aspects of these procedures constitute functional design features that led natural selection to incorporate them into the human species-typical architecture? That is, what adaptive problems are these procedures designed to solve, and in what way do their patterned outputs correspond to solutions? Evolutionary psychology has been effective as a research strategy because understanding the adaptive problems faced by our ancestors provides detailed predictions about the designs—the functional architecture—of our evolved psychological mechanisms, guiding new empirical initiatives.

Consequently, taking an evolutionary psychological approach offers a new framework for the study of rationality, choice, and organizational behavior that differs in certain respects from more traditional approaches. Because reliably developing species-typical mechanisms are the product of evolution, the functional logic they embody will be a species of evolutionary rationality—what we have called *ecological rationality* (Tooby and Cosmides, 1992b). Ecological rationality differs from ordinary concepts of rationality and functionality in several ways:

First, the functional product that evolved mechanisms are designed to produce is not utility maximization, social welfare or the general ability to realize goals (although it may overlap with these under many circumstances). Instead, the mechanisms were designed (in interaction with the other mechanisms in the architecture) to produce outputs that typically promoted genic fitness under ancestral conditions (that is, that increased the frequency of the mechanisms' genetic basis in subsequent generations *relative* to pre-existing or mutational alternatives). This definition of functionality often departs from both formalized normative theories of rationality and common sense intuitions about functionality. For example, mechanisms involved in dyadic or *n*-person exchange may, by design, cause choices that do not maximize absolute payoffs, violating standard rational principles (e.g. Hoffman *et al*., 1998; Price *et al*., 2002). Similarly, mechanisms underlying jealousy violate common sense notions of functionality. Jealousy often causes desperate unhappiness for the individual in which it is activated, as well as for its targets—a byproduct of the fact that its function is to spread the genetic basis of jealousy at the expense of the alternative design (indifference to the sexual behavior of one's mate), regardless of its impact on the jealous individual's happiness.

Second, these mechanisms were designed to be evolutionarily functional within the ecological (causal) structure of the ancestral world humans evolved in, but not necessarily in environments whose structure departs from ancestral conditions. On this view, modern 'irrationality' will commonly be the expression of ancestral functionality: A taste for salt evolved under conditions where salt—a necessary nutrient—was so chronically scarce that modern humans 'irrationally' overconsume it now that it is abundant.

Third, our cognitive architecture system has not been designed by evolution to reach with equal efficiency any goal that arbitrary preferences might nominate. Rather, our species-typical neurocomputational architecture evolved to solve particular families of adaptive problem that recurred frequently enough to select for mechanisms to solve them. Hence, the design of specific problem-solving systems is matched to the recurrent structure of ancestrally significant adaptive problems just as specific keys are machined to fit particular locks.

How does this apply to understanding multi-individual cooperation and collective action? We think that human evolutionary history has equipped the human mind with specialized psychological adaptations designed to realize gains in trade that occur both in 2-party exchanges and in *n*-party exchanges, including collective actions. We believe that the specific characteristics of these mechanisms (e.g. cheater detection circuits) reflect the ancestrally recurrent structure of these adaptive problems (the existence of payoffs to cheating), just as parts of the key shape complement detailed parts of the lock. We think the understanding of collective action and organizational behavior can be improved by exploring the properties of this evolved psychology. We expect that emerging evolutionarily psychological theories of *n*-person exchange and collective action will replace theories of standard economic rationality by explaining their puzzles. In particular, we think that violations of standard economic rationality that are often found in situations studied by researchers in organizational behavior, behavioral economics, and anthropology are expressions of well-engineered, ecologically rational mechanisms producing outputs that would have been advantageous in ancestrally typical conditions. Indeed, we think that many of the 'irrational' behavioral expressions of these mechanisms (such as voting behavior or donating blood) will come to be recognized as engineering byproducts of these functional designs when they are activated outside of the ancestral envelope of conditions for which they were designed.

A Surprise: Exchange is not Produced by General Rationality but by Evolved Neurocomputational Programs Functionally Specialized for Exchange Interactions

One of the first applications of the evolutionary psychological research program was to the phenomenon of dyadic exchange (i.e. where two parties deliver benefits to each other, each delivery being made conditional on the other). The primary questions addressed were: What do the actual computational procedures underlying exchange in humans look like, and how do their design features reflect solutions to the adaptive problems imposed by the recurrent structure of ancestral social exchanges? Perhaps the largest surprise to emerge from this research was the finding that exchange, despite being precisely the kind of thing that general rationality would guide rational agents to engage in, is not after all produced by a general-purpose set of reasoning abilities. That is, the traditional view that exchange is the expression of general economic rationality turns out to be false. Instead, exchange is produced by evolved reasoning specializations—social exchange procedures—tailored by natural selection to solve the computational problems specific to social exchange (Cosmides, 1989; Cosmides and Tooby, 1989, 1992, 2005; Hoffman *et al*., 1998). What leads to this conclusion? Exchange depends on conditionally delivered behavior (e.g. *I will do X if you do Y*), and so necessarily requires conditional reasoning for its regulation. In theory, of course, this could be accomplished by general reasoning abilities—the traditional assumption. Yet when methods were developed to study conditional reasoning performance in humans, subjects turned out to be very poor at detecting potential violations of conditional rules across a very broad range of familiar and unfamiliar contents (see review in Cosmides, 1989). One could not attribute the human ability to engage in exchange to a general ability to reason about conditionals.

Based on an adaptationist analysis of exchange, it was hypothesized that humans have an evolved cognitive specialization for reasoning about social

exchange, including a subroutine for detecting cheaters (Cosmides, 1989; Cosmides and Tooby, 1989). This led to a number of novel and specific predictions. For example, despite being generally poor at detecting potential violations of conditional rules, subjects should nevertheless detect them readily when the rule involves social exchange and looking for violations corresponds to looking for cheaters. Second, subjects' reasoning choices should correspond to the adaptively correct choices laid out in a specialized logic of social exchange, even when these choices conflicted with logically correct choices. These and related predictions were subsequently confirmed by numerous experiments conducted on subject populations drawn from many societies, from Harvard undergraduates to Amazonian hunter–horticulturalists (for review, see Cosmides and Tooby, 2005). Indeed, evidence from cognitive neuroscience shows it is possible through brain damage to have the mechanisms underlying social exchange reasoning selectively impaired, while other reasoning abilities remain intact—something that would not be possible if reasoning about different contents were accomplished using the same set of general-purpose rational mechanisms (Stone *et al*., 2002.)

Identifying the fitness advantages of exchange, of course, was not a deep evolutionary puzzle: As has been recognized at least since Adam Smith (if not for hundreds of thousands of generations), exchange increases the welfare of both parties. What is critical from an evolutionary perspective is whether the mechanisms that cause us to engage in exchange have been shaped in any specific ways by selection pressures that are particular to exchange interactions (Cosmides, 1989; Cosmides and Tooby, 1989). If exchange had turned out to be simply one out of an indefinitely large set of activities made possible by general learning abilities or general rationality, then an evolutionary approach would have had little to contribute. Of course, the evolutionary biology community has not widely addressed the superordinate category of exchange *per se* (Tooby and Cosmides, 1996), but primarily a narrower subset of exchange interactions: alternating, deferred (usually implicit) exchange. (Trivers, 1971) confusingly labeled this behavior reciprocal altruism, although it did not meet the biological or common sense definition of altruism. We will use *reciprocation* to refer to alternating deferred exchange and *exchange* or *reciprocity* to refer to the superordinate category of intercontingent interactions involving gains in trade).

George Williams was, characteristically, the first to introduce the topic of reciprocation into evolutionary biology, perceptively identifying the key strategy as one of the conditionality of one act of benefit delivery on the other (Williams, 1966). Trivers (1971) elaborated Williams' insights into a far richer treatment, in the light of the already existing experimental game theory literature on iterated prisoners' dilemmas and the social psychology literature on reciprocity. This was followed by more formal and systematic analyses using evolutionary game theory, such as Axelrod and Hamilton's exploration of tit-for-tat and Maynard Smith's ESS analyses of reciprocation (Axelrod and Hamilton, 1981; Maynard Smith, 1982). One result that robustly emerges from most evolutionary game theory formalizations of reciprocation is that it cannot easily evolve unless reciprocation strategies avoid cheaters. This implicitly assumes that reciprocation strategies have the capacity to detect instances of cheating (noncompliance, defection). Thus, although the fitness advantages of exchange are no puzzle, and their elucidation is a ratification rather than a contribution to economics, the theoretical analysis of reciprocation highlighted an evolutionary vulnerability to the strategy that our cognitive and motivational machinery would necessarily have been selected to address: A reciprocation strategy, to be successful, must incorporate defenses against being outcompeted by cheaters. This result applies not only to reciprocation, but to the more encompassing category of exchange (Cosmides and Tooby, 1989, 1992). This conclusion led to the prediction and discovery that humans have a cognitive specialization that allows them to detect violations of conditional rules when the conditional relationship involves exchange and a violation would constitute an act of cheating (Cosmides and Tooby, 2005).

If 2-party exchange is not caused by general rationality, then contributing to collective action is not—after all—an anomaly because it departs from general rationality. Because general rationality (as a set of real, physical mechanisms) seemingly does not exist, departures from it do not require explanation.[1] Instead, both types of behavior are explained by the actual designs of the evolved neurocomputational programs that cause them. But what exactly are these designs, and how

could they have emerged over evolutionary time in the face of potentially fitter competitive challenges by alternative designs? On this view, both 2-person and *n*-person exchange behavior appear (at first examination) to be anomalous from an evolutionary perspective because both suffer from parallel vulnerabilities to evolved counterstrategies that should have outcompeted them: Cheaters and free riders both take benefits without making contributions, and so would have outcompeted from the beginning simple first-order designs for 2-party or *n*-party unconditional 'exchange' (benefit delivery). Consequently, what an evolutionary perspective predicts instead is that the mechanisms that evolved must necessarily embody more complex, *conditional* exchange strategies whose computational properties defend them against outcompetition by cheaters (in 2-party exchange) and free-riders (in *n*-person exchanges such as collective actions). These strategies do not just pick the highest *absolute* payoffs—they often sacrifice such payoffs in favor of enduring practices that generate higher *relative* payoffs against exploitive strategies, when averaged across all interactions. Antiexploitation computational elements are thus central and indispensable design features of a cooperative strategy—essential if the strategy is to successfully emerge and stably persist. Hence, the relevant question becomes: What exactly do these exploitation-resistant strategies look like, when described as psychological (information-processing) designs?

The Functional Logic that Regulates Motivation is an Integral Part of Evolved Social Rationality

Another cardinal difference between standard approaches to rationality and evolutionary approaches to rationality involve the role that motivation plays in decision-making. Instead of preferences being arbitrary, wholly acquired, or exogenous to rationality, evolved motivational specializations are (we believe) intrinsic components of more encompassing, ecologically rational problem-solving adaptations, such as adaptations for 2-party exchange and collective action. That is, motivational mechanisms are endogenous and indissoluble constituents to systems of rationality and cognition, with regulatory architectures that employ evolved, proprietary forms of representation in order to direct motivation so that individuals correctly implement solutions to adaptive problems (Tooby *et al*., 2005). Thus, a fear of snakes is not the product of a general rational system for avoiding negative payoffs, but rather an evolved system with evolved proprietary representations (*snake*) and motivational dispositions (*fear*/*avoidance*). This system is ecologically rational in that its specialized motivational circuits motivate behavior that was ancestrally functional—snake avoidance—even though it may sometimes lead to unnecessary ('irrational') avoidance of some harmless snakes (i.e. it is not perfect). The key claim is that for motivation to direct behavior adaptively, it requires guidance systems equipped with evolved representational elements capable of correctly specifying the targets and intensities of the motivation. Examples of such target-specifying 'innate ideas' that reliably develop within the *n*-person exchange motivational system include *contributor*, *free rider*, *benefit to the group*, *contribution required for entitlement*, *individual receipt of joint benefit*, *entitlement to benefit, undercontribution, exploitation*, and so on. These evolved conceptual elements are necessary to direct the specific motivations that can lead to the realization of potential gains in trade while simultaneously defending against exploitation.

Properly targeted and calibrated motivational circuits are central to defending against free riders (and cheaters), and hence essential to making multi-individual cooperation (and dyadic exchange) evolutionarily stable against exploitation. This regulation is delivered through a set of motivational specializations that, for example:

(1) calibrate willingness to devote effort to collective projects,
(2) direct punitive sentiment toward free riders, and
(3) mobilize reward sentiments towards contributors.

It is important to emphasize that the control system that regulates the deployment of these sentiments is cognitive or computational in nature, and must apply procedures to representations and inputs to determine, for example, under what conditions to feel punitive sentiment, how much to feel, and toward whom to feel it. Despite the frequent mystification of feeling, feelings are not blind, ineffable, randomly emerging forces, but are neurally computed outputs of evolved programs whose structure can be mapped and whose functions can be identified (Tooby and Cosmides, 2005).

THE EVOLVED FUNCTIONAL DESIGN OF THE PSYCHOLOGICAL MACHINERY UNDERLYING *N*-PERSON EXCHANGE

Using the foregoing framework, we would like to explore how multi-individual cooperation (*n*-person exchanges, including collective action) as a pervasive human activity can be accounted for. We would like to sketch out some of the cognitive and motivational design features that we propose underlie the human ability to engage in collective enterprises. We would like to link these design features in a general way to selection pressures and ancestral situations that we believe led to their incorporation into the human psychological architecture.

The evolution of the ability to engage in 2-party exchange would almost certainly have preceded the evolution of the ability to navigate the game theoretic complexities of *n*-party exchange and collective action. Consequently, we would like to consider what additional components would have had to be added to cognitive mechanisms for dyadic social exchange in order to extend the system to *n*-person exchange, including the special case of collective action. Finally, we will consider how such a model fares against criticism that multi-individual cooperation, including collective action, could not have evolved via individual selection for such exchange strategies.

Dyadic Exchange Builds Many of the Components Necessary for *n*-Person Exchange

The first claim is that the evolution of mechanisms for 2-party exchange built a large part of the computational infrastructure necessary for engaging in *n*-party exchange (see, e.g. Cosmides and Tooby, 1989 for a detailed description of mechanisms; for a review of evidence, see Cosmides and Tooby, 2005). The evidence indicates that the social exchange algorithms that evolved for exchange use conceptual primitives such as *self*; *agent* (or *party*); the *welfare* or *interest* of a party; *exchange*; *entitlement* to benefit, *benefit to be gained*; *cost* or *requirement to be met* to gain entitlement to the rationed benefit; *cheater* (a party that has taken the benefit while having intentionally not met the requirement); *intended outcome* vs *accidental outcome*; *consent* to a *social contract* (an intercontingent plan of action in which each party agrees to undertake a course of action conditional on the other party's execution of a corresponding course of action); and so on. The exchange system also involves specialized procedures that interrelate these concepts, map actual situations into these representations, evaluate magnitudes, and compute necessary regulatory outputs. Two obvious examples are a system that scans the social and instrumental world for opportunities to increase welfare through exchange, and the look-for-cheaters algorithm.

Many of these components are not proprietary to the exchange system, but evolved because they are more broadly useful across a wide array of functions. For example, to plan an agent needs mechanisms of valuation—that is, procedures that can turn representations of a given situation (actual or imagined) into a magnitude representing the welfare that the situation holds for the agent. Similarly, agents must be equipped with procedures that represent potential or actual *changes* in situations as positive or negative magnitudes in *welfare* or *interest*, which generates the perspective-specific concepts of *benefits*, and *costs*. Humans can do this for themselves, and attribute this ability to other agents (i.e. have a capacity to interpret behavior in terms of an intuitive *theory of interests* that is, we think, one subcomponent of the evolved *theory of mind* system (Baron-Cohen *et al.*, 1985)). *Interests* are representations of the sets of changes to situations that positively or negatively modify an agent's welfare. Acts can simultaneously modify the welfare of more than one party. When this is foreseeable such behavior is commonly interpreted as expressing a stable internal variable that regulates how much the welfare of the self is traded off against the other party: a *welfare trade-off ratio* or *WTR* (Tooby and Cosmides, 2002, 2005, forthcoming). Thus, if one party reliably incurs high costs in order to deliver benefits to another, that behavior expresses a high welfare trade-off ratio toward the recipient.

When single exchanges morph into extended series or enduring exchange relationships, they may not be structured in the long run by tit-for-tat-like rules (which require alternating acts of reciprocation), but rather by a rule that matches one exchange partner's welfare trade-off ratio to the other's, scaled by the relative symmetry and frequency of each party's opportunities to help the other (e.g. Tooby and Cosmides, 1996, 2002). This exchange psychology represents what are, in effect, *accounts*—how much each party has done for the

other. Accounts and their interrelationships (such as a 'balance of trade' index) generate conceptual elements such as *owe*, *debt*, *obligation*, *exploitation*, *cheating*, and so on. Computations on these accounts and their interrelationships also regulate motivation: Others' kindness to us moves us (i.e. recalibrates our welfare trade-off ratio toward them so that it is higher than it was before). Their indifference, unwillingness to help, or their cheating harden our hearts (i.e. downregulates our welfare trade-off ratio with respect to them). If exploitive or deceptive enough, these acts may motivate anger and punitive sentiment towards the offender. Obviously, this neurocomputational machinery is beyond conscious awareness, although we are aware of some of its outputs.

Recursion as One Modification to 2-party Exchange Psychology that can Help to Adapt it to Solving the Problems Arising from *n*-party Exchange

Understanding the relationships that 2-party exchange embodies seems effortless and intuitive. In reality, these interactions involve a surprisingly large number of interrelationships, variables, steps, conditional branch points, unexpressed contingencies, and so on (see Cosmides and Tooby, 1989, for a list of the conditional relationships required to encapsulate a simple dyadic exchange). What makes this maze transparent, intuitive, and navigable is that humans are equipped with evolved machinery designed to automatically represent and track this intricate set of implicated relationships. What would need to be modified in the architecture of 2-party exchange to allow it to be able to coordinate the behavior of three, four, or more individuals to realize *n*-person gains in trade?

Theoretically, much of what is needed could be achieved through a relatively modest change to the architecture: adding recursion to the exchange structure. The pre-existing computational template used for navigating 2-party exchange could be leveraged to navigate *n*-party exchange through adding the ability to reduplicate an additional agent-centered unit for representing exchange relations for each added person. That is, if the 2-party form can be abbreviated as 'I will do x if you do y', then the generalization of this intercontingent regulatory structure to *n*-person social exchange could be abbreviated as 'For individuals 1 though n, I (individual$_1$) will do x_1 if individual$_2$ does x_2 and individual$_3$ does x_3 ...' (or 'I will if you$_1$, you$_2$, you$_3$...will.'). It is important to recognize that the gain that is produced need not be a public good, but can be a series of private goods (for benefit$_1$, for benefit$_2$, for benefit$_3$...).

Discoordination and Complexity as Factors that Sharply Limit the Scope of *n*-Party Exchange

One important limiting factor for *n*-person exchange is the problem of reaching and sustaining coordination among the interactants. Complexity in the set of possible alternative arrangements and behavioral interdependencies mounts explosively as the numbers in the exchange grow. The first problem is that of establishing a mutual understanding among many different minds. In dyadic exchange, this is relatively easy: If the proposed exchange is not worthwhile for one participant, then the other participant increases the value of the offer (or refuses to engage in the exchange). To persuade a holdout in a multisided negotiation, however, it is in every other party's interest to have others contribute the needed additional inducement. This leads to an $n-1$ sided game of chicken. Similar negotiative problems and instabilities emerge, for example, in deciding who is to do what, whenever complementary tasks with different costs need to be performed.

Another important limiting factor for *n*-person exchange is the rapidly increasing cognitive demand posed by the rapid growth of combinatorial complexity as the numbers in an exchange grow. For example, if a participant has to divide her attention among an increasing number of exchange partners, how are instances of undercontributing or other forms of exploitation to be detected? With increasing numbers of participants, the problems of negotiation, communication, coordination, and agreement become far more complex. Complexity is injected by the number of different roles individuals could take, by the number of different goals the group could pursue, by the possibility of different costs being incurred by different participants doing their respective parts, by the possibility of different individuals deriving different magnitudes of benefit from the *n*-sided exchange, and so on. Limitations on this type of exchange emerge from logistical constraints in negotiating and implementing *n*-sided exchanges, cognitive bottlenecks in attending and representing, and corresponding

limitations of the ability to defend oneself against exploitation within them. These problems are relatively small if the set of interactants is small, such as three, four, five, or six. But with increasing numbers of interactants, communication becomes noisy, mutual surveillance becomes difficult, cheating becomes easier, and transaction costs become prohibitive.

For these reasons, it is inevitable that in the regular course of human life immense numbers of mutually beneficial potential gains in trade from *n*-person exchange are never realized. Nevertheless, the human ethnographic record indicates that over our evolutionary history humans have been able to successfully carve out and master some subsets from the space of *n*-person exchanges (Ostrom, 1990; Price *et al.*, 2002). The persistence of these opportunities would have favored selection for cognitive and motivational mechanisms that could cost-effectively implement and increase the scope of the various families of solution.

Different Modes of *n*-Party Exchange Exploit Different Strategies for Solving Coordination Problems

We think that there exist a number of different but interrelated pathways to solving coordination problems, and that each has left its imprint on our evolved social psychology. For example, these coordinative problems selected for specialized cognitive and motivational adaptations that create the uniquely human interlocking roles of leadership and followership. Humans find these concepts and behaviors intuitive, and they appear spontaneously in all human cultures (within delimited social contexts). That is, we think a set of neurocomputational adaptations subserving leadership and followership evolved as solution to the coordination problems that emerge in complex exchanges and intercontingent social structures (Tooby and Cosmides, 2002, forthcoming). Leadership allows efficient coordination by routing coordinative decisions through a single individual mind so their mutual implications can be computationally integrated rapidly and dynamically. Leadership (when competent) simplifies the dynamics of determining new projects and allocating efforts to achieve them—something that becomes increasingly important when the numbers in a potentially cooperating group get larger than a handful. Cheating in leadership involves directing the group in excessively self-interested, group-injurious ways, while in followership it encompasses not only general undercontribution, but also the more specific concepts of unreliability, disloyalty and disobedience (discoordination). By effective use of joint attention, persuasion, directed disapproval and approval, threat, punishment, and reward, leaders can collapse discoordinating possibilities entertained by the other $n-1$ members of the group into a single shared representation of a plan of action and benefit allocation. The fact that such a plan is shared itself should excite greater willingness to engage in an *n*-person exchange, precisely because a lack of a shared representation of coordinated action is one of the primary obstacles to achieving *n*-person gains in trade.

Indeed, certain events that invite a common and mutually manifest response can sometimes substitute for leadership in solving coordination problems. If an event draws joint attention, and inherently invites the same kind of response from most observers, then the necessary coordination to trigger a collective effort can be triggered. The right kind of event can spontaneously lead to acephelous defense during attack, hunts precipitated by the sudden appearance of game animals, emergency-provoked rescues, and outrage-provoked mobbing, riots, and so on. The key feature of event-organized exchanges is existence of a single supersalient event that not only coordinates the minds of the co-participants, but makes it clear for each participant that this coordination is present in the minds of the other participants. Implicitly present is a shared appreciation of the potential for gains offered by coordination, which activates motivation once the near achievement or reality of coordination is perceived.

A third pathway that also collapses many of these problems into manageable units is the activation of a mutually enforced system of strict egalitarianism of external conditions and markers—what Alan Page Fiske calls *equality matching* (Fiske, 1991). This is not an equality of costs, benefits, and inner accounts, but of externally demarcated units of shares and burdens (e.g. one man, one vote), greatly simplifying monitoring against exploitation. We think this intuitive method of social organization is anchored in our evolved exchange circuitry as one natural mode of *n*-person exchange. In general, *n*-person exchanges are far more cognitively tractable to the extent that

contributions from the participants consist of parallel behaviors: If everyone is expected to contribute in the same way, then it is vastly easier to communicate and to monitor such an exchange.

Another natural mode involves collective action to produce public goods, policed by punitive sentiment against free riders (Price *et al*., 2002; Price, 2003, in press A, B). The entire destabilizing negotiation over the allocation of a jointly produced resource or desirable state of affairs is eliminated when what is produced is a public good-like diffused benefit. Collective actions for public goods simplify the issue of defending against exploitation by focusing surveillance on the issue of undercontribution. If the good to be produced is not excludable, then producers cannot be cheated by being denied consumption of the product of the joint effort. As we will explain, many groups that may initially emerge to pursue nonpublic goods eventually also produce public goods. This makes the evolved psychology of collective action relevant to most and perhaps all cooperating groups.

The Resuscitation or Perpetuation of a Pre-existing *n*-Person Exchange System is Another Pathway to Solving the Problems of Coordination

Because successfully coordinated *n*-person exchange is so difficult to achieve, its existence once attained becomes a kind of resource to be valued and effortfully perpetuated by its participants. That is, another condition that greatly simplifies subsequent coordination and the fruits that it makes available is the prior existence of some previously existing set of understandings that arose from a previous coordinated interaction. This means that it is in the interests of participants to give the precedents set by a previous *n*-person coordination a weight and an inertia that they would not necessarily give those expressed in previous 2-person exchanges, which can be far more easily generated, modified, and discarded as needed. The coordination problems inherent in establishing *n*-person exchanges make it far too costly simply to create them *de novo* as 1-shot arrangements in the face of every new opportunity for gains in trade. Because the great difficulty of arriving at coordination usually prevents mutually beneficial *n*-person exchanges from coming into existence where *n* is large, the default willingness to invest effort in their creation for a specific short-term end should be low. Correspondingly, however, the existence of recent and mutually known precedents gives a coordinative starting point that allows potential interactants to circumvent considering, negotiating, and sorting through the huge number of logically possible exchange arrangements that arise in situations involving more than a few people. Hence, the prior existence of *n*-person exchange structures should facilitate the motivation to engage in such exchanges by offering the opportunity for lower cost coordination. The magnitude of this effect should be a function of the number of individuals involved. The larger the set of minds to be coordinated, the greater the cost of arriving at a new set of mutually consistent representations. This means that the greater the number of participants is, the greater the comparative advantage of conservatively perpetuating pre-existing arrangements will be, however beneficial or flawed those pre-existing arrangements were.[2]

The key point is that our evolved coalitional psychology is designed to appreciate that it is easier to resuscitate or perpetuate successful coordinative arrangements than to create them.[3] In reality, groups do not objectively exist—they only exist to the extent that they are represented in mutually consistent ways in the minds of assorted individuals. So, groups only persist if individuals represent them as having continuity over time. Our cognitive circuits for representing groups are advantageously designed to endow them with continuity.

The coordinative arrangements of an existing *n*-person exchange system themselves constitute a useful resource or instrumentality. This is because they make it possible for the individuals in the exchange system (*coalition*) to consistently take advantage of the flow of opportunities to generate joint benefits that uncoordinated individuals could not. That means that collective efforts do not simply emerge rationally at the time individuals come to anticipate the joint benefits that could be attained from reaching a specific goal, and cease when the goal has been reached. Instead, our minds evolved to inherently value the existence of the groups for their own sake, so that we have a suitable coordinative structure ready and waiting when the need strikes. Just as humans value any other intrinsic good, we are designed to prize the existence of and continuation of the groups we are members of. We feel pleasure upon becoming a

valued member of a group, satisfaction in its creation and successes, and sadness at its dissipation. Our evolved psychology includes a motivational variable that tracks the social sufficiency of the group or groups the individual is a member of. This variable regulates the appetite for finding or developing groups if the individual lacks a sufficient number of sources of support. We seek to acquire, create, and perpetuate group identities.

Groups Tend to Turn into Coalitions, Imposing Exchange Obligations on their Members

For this reason, there is a tendency for groups to outlive whatever rationales, circumstances, or events gave rise to their creation,[4] and to gradually morph into what we call *amplification coalitions* (Tooby and Cosmides, 2002, forthcoming). That is, the aboriginal *n*-party exchange system is one whose purpose is the amplification of the ability of each of its members to realize her interests in daily events by cost-effectively combining welfare trade-offs and joint efforts from the other members. Alexandre Dumas memorably encapsulated the nature of the amplification coalition as *all for one and one for all* (Dumas, 1844). Dyadic alliances or friendships are minimal 2-party instantiations of such a coalition, but they can be easily extended to include more individuals. Gangs, cliques, factions, friendships, alliances—these coalitions exist whenever individuals feel motivated to provide general support to someone else by virtue of implicit or explicit co-membership. What earns you the right to the support of others is the contribution you make to supporting others when they could cost-effectively benefit from assistance. Reciprocally, when others observe a pattern of undercontribution on your part, this erodes the willingness of others to support you.

For example, disputes are endemic in small-scale social interaction, and are often decided in ways that reflect the relative power of the disputants. Other things being equal, two individuals have more power than one, three more than two, and so on, so support for an individual in a dispute by others is often a considerable help in resolving the matter in a way that favors the supported individual. Everything can be taken from a powerless individual. So in addition to the other joint resources that were produced by collective effort ancestrally, ordinary social competition in hunter–gatherer interactions would have selected for a coalitional psychology. Indeed, experimental evidence indicates that our minds contain an evolved alliance detector that scans for the membership in coalitions bound by mutual alliance. This evolved cognitive specialization, as a routine part of person representation, takes the pattern of cues of mutual alliance present in individual behavior in the social world as input, and deduces coalitional categories in the mind as output (Kurzban *et al.*, 2001b). That is, our minds are designed to interpret the social world in terms of coalitions, spontaneously and automatically.

Trade-offs arise from the opposing costs and benefits to the individual of belonging to larger versus smaller coalitions: The larger the coalition, the stronger it is, and the more resources and support it can supply to an individual. However, the larger the coalition, the less joint attention and value the 'group mind' can give to an individual, and the less indispensable she is to it. Even more important, the larger the coalition, the more frequently disputes involving welfare trade-offs will occur among ingroup members rather than between members and nonmembers. Whenever disputes are internal to the coalition, the individual cannot draw on it for support. At the logical extreme, if everyone belongs to a single encompassing coalition, then it is not a coalition that usefully amplifies individual power in the disputes that emerge in daily life. Help from the coalition cannot be dependably asymmetric when both disputants happen to be in the same coalition and so qualify for the same support. Reciprocally, the smaller the coalition, the more differentially valued and attended to a member will be. Moreover, the smaller the coalition, the more individuals lie outside of it, and so more disputes with others will occur between ingroup members and outgroup members. As a result, the smaller the coalition, the more often the coalition can be used to amplify one individual's power against another in a dispute. Consequently, within each larger scale coalition, individuals benefit by calving off a smaller scale coalition to side with them when there are internal disputes—support that is earned by siding with others in other internal disputes. (Indeed, more marginal individuals may covertly welcome or encourage disputes among others that can give them opportunities to assist others and hence cement their status as someone entitled to defense and support.[5]) As a result, individuals are usually members of many concentric amplification

coalitions from dyads up to the largest scale of intergroup conflict. Coalitions existing at different scales are fractally organized by links of mutual alliance. This chronic set of problems suggests that the psychological system that carves out a coalitional identity includes regulatory circuitry to assess the local fractal structure of groups, and to allocate effort to forging alliances across a range of scales. For similar reasons, each individual in an amplification coalition should spontaneously disapprove of the formation of strong individual loyalties toward outgroup members by other individuals in the ingroup: Such loyalties will contract the set of circumstances in which the compromised individual can be expected to amplify the power of other ingroup members.[6]

These considerations extend beyond help in disputes. If, for example, responsibility for material assistance is diffused among too many individuals, no one helps (leading to the well-known bystander effect; Darley and Latane, 1968). To be cheated in this context is to fail to receive assistance when one has given it to other ingroup members. Cheating among large undifferentiated groups is more likely because the blame for the absence of assistance is distributed among too many others. For assurance of help from others to be strong, there must be bonds of mutual dependence among a small enough number of individuals that withdrawal by a cheated individual strongly hurts a few specific others. Individuals accordingly feel a sense of precariousness among homogeneity, and a desire to be individually valued by a small number of local others (see also, Tooby and Cosmides, 1996). Indeed, Marilyn Brewer proposed that individual identification with groups is shaped by a motivation for group optimal distinctiveness, which may also serve this function (Brewer, 1991). In short, it is not enough just to be nominally a member of a coalition—a sufficient number of local individuals in the coalition must know you, differentially value you, monitor your welfare, and support you when you need it.

Both for reasons of social competition and material insurance, when homogeneous coalitions get too large, our evolved coalitional psychologies prefer to give them internal factional structure. As Madison pointed out in the Federalist Papers, the 'latent causes of faction are thus sown in the nature of man' (Madison, 1788/1982). More broadly, modern humans feel naked without the protection of a sufficient set of group identities because the well-designed hunter–gatherer was unprotected against threats from the social and natural world without the support of others (Tooby and Cosmides, 1996). Individuals have an evolved appetite for creating coalitional identities independent of and prior to a real need for them. If individuals were designed to wait until need struck to build a coalition, the organizational difficulties and latencies would make it too late. So, to serve these appetites, modern humans inflect their lives with notional and recreational groups that often seem detached from utilitarian need, such as churches, clubs, sports teams, political groups, and the like (see, e.g. Tiger, 2004). Moreover, recurrently spending time with others in group settings is itself a primary generator of the expectation that one is in an amplification coalition. This means that whatever the formal or legal theory present, life in the workplace invites implicit assumptions of mutual amplification and support that transcends work-specific activities. Conformity to these expectations is considered elementary decency, and their repudiation or discouragement is experienced as an alienating feature of many corporate and nonprofit workplace cultures.

The Status of Coalitions is a Public Good, Making Coalitions Collective Actions

Amplification coalitions are intuitively understood, spontaneously self-organizing, and may emerge into awareness when one or more individuals are challenged by outsiders. One crucible that led to the evolution of the psychology underlying amplification coalitions was their ancestral operation in zero-sum competitions over access to resources by contending sets of individuals. Individuals take or relinquish resources to the extent that they feel empowered or overmatched. Public signals of support (or its absence) lead individuals and sets of individuals to revise what they attempt to possess, consume, or do. In consequence, our species-typical psychology evolved to represent coalitions as having status (Brase, 1998; Sidanius and Pratto, 1999). Individuals (where cost-effective) expend effort to defend, advertise or enhance the status of the coalitions they belong to, and to deflate the status of rival groups. This is one set of individual behaviors that humans are designed to interpret as expressions of the group as an entity. For a given

action by individuals, this interpretive system licenses the transformation of *He did this to me* (one individual taking action with respect to another) into *They did this to us* (one group taking action with respect to another).

A central feature of interactions over group status is that their product is a public good. That is, anyone who is a member of an amplification coalition automatically partakes of its status, and gets the benefits in daily transactions of the greater weight nonmembers will place on members' interests out of respect for how members coordinate to support each other in conflicts. Acts of alliance cause observers to categorize individuals together as members of the same coalition (Kurzban *et al.*, 2001b). The representations in the minds of observers of the status of a coalition are a common resource that has the properties diagnostic of a public good. That makes the work of maintaining group status intrinsically a collective action (Tooby and Cosmides, 2002, forthcoming). Because amplification coalitions arise easily and spontaneously from the mere existence of any cooperative precedent, and individuals tend to assume their existence and continuity, it follows that all groups will gravitate toward being experienced, to some extent, as collective actions by their members. This will be especially true around issues that affect the status of the group in the mind of outsiders.

Amplification coalitions by their nature are assumed to continue indefinitely as ongoing operations, so they are formally different from collective actions that exist to achieve a realizable, finite goal, with a specific termination point. Extended cooperative interactions invite the implicit assumption of the existence of an amplification coalition, binding its members into an exchange relationship. The individual is never done with the obligation to contribute to the group so long as the group persists. One can cheat by not contributing one's share of support to others in the coalition, including not contributing one's share to maintaining the representation of group status in the minds of nonmembers. Individuals can also cheat by assuming membership and signaling it to outsiders without having earned it (Brase, 1998). This makes policing the boundaries of such a coalition a sensitive issue. Natural questions that express this intuitive coalitional psychology include: *Is individual i really a member of the group*? *Is i a contributing member of the group* (*fully paid up*; *not a free rider*)? If the group has been in existence for some time, new members at the threshold of joining will intrinsically strike existing members as resembling cheaters in that they are attempting to gain the benefits of membership, without having yet paid any costs (Tooby and Cosmides, 2002). Social practices will emerge to express or disarm the punitive sentiment as well as the distrust recruits spontaneously attract from veterans. A strong feature of coalitional psychology will be organized around issues of membership, group identity, the price of entry, initiation, genuineness of membership (loyalty, commitment), exclusivity, and the pressure on new members to make contributions to bring them into the implicit exchange relationship in the minds of veteran members. This exchange view predicts that organizations in all human cultures will spontaneously tend to pass costlier, less attractive tasks to younger, newer recruits, and recruits will be motivated to sacrifice more in order to make more dramatic and observable contributions than established members will. The punitive sentiment veterans feel towards new entrants, combined with the advantages for veterans of creating strong representations of the dominance of the coalition over new members, together often lead to the coordinated infliction of costs by veterans on new entrants. Correspondingly, when recent or potential entrants assume increasing costs attendant upon membership, the motivational systems of established members will increasingly recategorize them as potentially worthwhile recipients of group benefits in an *n*-person exchange. The motivation to police the boundary will be calibrated by the magnitude of the benefits that go with membership, the labor recruitment needs of the coalition, the levels of trust and interdependence needed to function as a coalition, and the frequency with which the welfare of the coalition (or its participants) depends upon substantial sacrifices by members.

The Addition of the Cognitive Ability to Represent a Group as an Individual Allows an Additional Approach to Applying Dyadic Exchange Machinery to *n*-person Exchange

Although the addition of recursion is one important computational modification to 2-party exchange psychology, a second important modification to this exchange psychology involves

widening the exchange system's definition of what kinds of entities can be computationally treated as a *party* or *agent*. During the initial evolution of exchange circuitry, *party* would have meant *individual human*. The next step would have been widening *party* or *agent* so that open arguments in procedures that formerly would have referred strictly to human individuals become able to refer also to coalitions, alliances, communities, and other entities as well. This allows humans to represent, understand, engage in, enforce, and cheat on dyadic exchanges where one party is an individual, and the other party is a group, as well as where both parties are groups.

This cognitive ability to interpret a coalition as an agent is not just speculation: Two decades of research on social exchange reasoning demonstrates that subjects can understand and reason about 2-party exchanges just as readily when the exchange rule is a social rule entered into with a group as they can about exchanges between two individuals (Cosmides, 1989; Cosmides and Tooby, 1992, 2005). Such a broadening of the meaning of agent creates alternative ways for individuals to represent and enter into *n*-person exchanges. An interaction could be represented as a single multilateral *n*-person exchange with every individual as a separate node; or, it could be represented for each individual as a dyadic exchange between the individual and the group, with two nodes for the two agents—*self* and *group*. Mathematically, an *n*-person multilateral exchange involves $n(n-1)$ directional links, while an *n*-person exchange conceptualized as a series of individual to group exchanges involves $2n$ links (or n exchanges). Seeing an *n*-person exchange in terms of a series of dyadic exchanges between each group member and the group as a single entity allows for great cognitive simplification and transparency (when n is greater than 3): Simply using pre-existing dyadic machinery, each individual can clearly understand and reason about what she herself owes the group, and what the group owes her. She does not need to simultaneously represent links to every individual in the group. She can also monitor cheating on the part of other members applying dyadic reasoning parallel to what the individual applies to the self: Is a given individual taking the benefit of the joint effort? Is the individual holding up their part? A person can comprehend and cheater detect an *n*-person exchange as a series of dyadic exchanges by going individual by individual through the group, first taking the individual's perspective, and then the group's (represented as a single party). This allows dyadic exchange psychology to generate and cognitively treat many kinds of social rules as social contracts (Cosmides and Tooby, 1989).

The widening of the agent slot to include groups as well as individuals is a cognitive expansion that applies beyond exchange relationships, although that may have been its selective crucible. In human psychology, groups are not treated simply as categories or aggregations of individuals. Groups are conceptualized and experienced as entities to which we can properly attribute mental states, and which can have propositional attitudes (Cosmides and Tooby, 2000). That is, humans have no trouble interpreting claims about the mental states of groups as if a group constituted an individual agent with a single mind. (In reality, obviously, they are not.) For example: Our university or department or company can want us to do things; our community can be angry, or disgusted or pleased with us; our nation can be intimidated or vindictive; our village can mistakenly believe we killed someone; we can be afraid to disappoint our group. Nelson signaled at Trafalgar, '*England expects that every man will do his duty*' without wondering whether his fleet would be philosophically puzzled about how a nation can entertain an expectation.

Moreover, humans also unproblematically treat groups according to an intuitive *theory of interests* that we believe originally evolved to interpret individual action. Humans not only represent individuals as having interests, but typically represent groups has having welfares or interests as well. Accordingly, groups (like individuals) are experienced as having a unitary exchange account that we can trade off our welfare against (and vice versa). In so doing, we can owe the group, supply help to the group so that it becomes indebted to us, feel entitled to group assistance, or punish the group for not taking our welfare sufficiently into account; reciprocally, the group can owe us, supply help to us so that we become indebted to it; it can feel entitled to our assistance, or punish us for not taking its welfare sufficiently into account. Groups can have status, rank, stigma, and dominance relations, not to mention friendships and enmities. We can see them as exploitive (favoring their own welfare above ours) or generous (injuring their own welfare to benefit us). A group can be dominant over us, whether we

are a member or not. (Indeed, leading members of male fraternal organizations often want new members to appreciate the power the coalition has over them; Tiger, 2004). Many cognitive scientists are persuaded by the developmental, cognitive, and neural evidence that humans come equipped with what has been called a *theory of mind*, a specialized set of mechanisms that interprets individual behavior in terms of mental states (Baron-Cohen *et al*., 1985). We propose that a set of augmentations and modifications to the theory of mind system constitute *a theory of group mind* system (Tooby and Cosmides, 2002, forthcoming). This system generates representations about what a group thinks and feels, triggers the individual's feelings with respect to imagined mental states of groups, and treats groups as entities that can be the object of emotions and motivations that initially evolved for individuals.

In reality, of course, groups (as sets of individuals) do not have a single mind, body, or voice. In order to treat groups as agents, the mind must contain procedures that can interpret the observable acts of individuals as the behavior of groups or the expression of group mental states and intentions. If groups are mentally represented as amplification coalitions, then this predicts that one major set of acts of individuals that invite reinterpretation as acts of the group are behaviors by one or more individuals in one coalition that negatively impact the welfare of one or more individuals in another coalition. When members of the Los Angeles African–American community saw footage of several Los Angeles police officers clubbing Rodney King, they did not treat it as an affair just between the interactants—it was seen as one *community* (coalitional entity) doing something to another *community* (coalitional entity). In accepting or attacking the acts of individuals from other coalitions, coalitions act as if they are negotiating general precedents specifying how much members of each coalition are obliged to trade off their welfare on behalf of members of the other coalition (an equilibrium welfare trade-off ratio expressing relative power or dominance). Some individual acts—*outrages*—serve as coordinative signals for a group to act together in a collective action to obtain what is usually a public good: to act together to change the representations in the minds of the other group as to the relative power of the two groups, by threatening, attacking, or dispossessing them.

How humans represent the organizations they work within will have a major impact on how they behave in these organizations. Consequently, it is hard to overestimate the magnitude of the impact that the ability to represent groups as individuals has had on human sociality. This key expansion in the human ability to engage in *n*-person exchange populates our mental world and our social world with groups, using much of the same machinery that governs how we think and feel about individuals. Yet, we maintain in parallel the ability to interpret groups as sets of individuals engaged in multilateral exchange. These alternative ways of representing *n*-person exchanges are different, and mobilize different ways of thinking and feeling: In one case, I am in a multilateral exchange with everyone else in the group. In the other, I am in a dyadic exchange relationship between myself and the group. It is possible for the same individual to activate both of these representational modalities toward the same object, at least at different times. A corporation might for some purposes be represented by the employee as a single agent (the company) in a dyadic exchange relationship with the employee, and on other occasions as a team or amplification coalition (a multilateral exchange) that the employee is a member of. These alternative representational methods will activate different motivational responses.

It is perhaps not an overstatement to say that corporations (and many other formal organizations) as human social institutions are founded on the evolved representational ability to represent a group (i.e. a corporation) as a unified agent or person, while collectives, associations, and unions more often invite representation as multilateral exchanges among (relatively) equal members. If I interpret my employment with a corporation as dyadic relationship, then my attention is focused on allocational conflict over benefits and costs (e.g. salary, effort, risk) between the two parties, myself and the corporation. Ordinarily, in dyadic exchange, the marginal cost of moving a unit of benefit from one individual to the other is compared, with some modulated expectation that units will go to the individual to whom it does the most good. If the corporation is viewed as an individual, however, it strikes the employee as a wealthy, dominant, and potentially exploitive individual—something not significantly harmed by transferring marginal units of benefit that would make a great deal of difference to the

employee. The fact that the corporation does not respond according to this logic is a continuous source of alienation for employees. This modality of representation provides the foundation for the Marxian and labor union worldviews that frame workers and management as inherently locked in an oppositional, zero sum relationship with each other. Alternatively, if the individuals networked together into a corporation represent themselves as members of a team, then the invited motivational consequences are a convergence of interests. In either case, workplaces usually bring together bounded sets of individuals in continuous association with each other, often facing an open-ended set of situations. Ancestrally, spending a major fraction of each day together indicated mutual membership in an amplification coalition, implicitly obliging members to provide a form of cooperative social insurance to each other. The expectations generated by different forms of organization are not just provided by explicit rules or by local cultural practices. Our evolved mechanisms, triggered by experience, assign their own system of evolved meanings to social interactions even in defiance of explicit rules or attempts at contrary cultural indoctrination.

Anti-exploitation Motivational Circuitry in *n*-person Exchange

Neither dyadic nor *n*-person exchange can evolve or be evolutionarily stable unless cheating is detected and then effectively responded to. Numerous experiments have established the first element: that the human brain does indeed contain evolved circuitry specialized for detecting cheating in both dyadic and *n*-person exchange relationships (Cosmides, 1989; Cosmides and Tooby, 1989, 1992). To function properly, however, cheater detection circuitry needs to be coupled to a component that motivates effective responses to cheating. An effective response is one that helps to make participation in exchange fitter than either cheating or nonparticipation (as well as fitter than other, more complex uncooperative strategies). The key property that these motivational circuits must have is the effect of making the average lifetime fitness of those equipped with cooperative circuits, on balance, greater than the average lifetime fitness of those with a stronger disposition to free ride. Given this design criterion, what should the outputs of this motivational component look like?

The problem of sustaining cooperation in the face of potential exploitation is severe and ubiquitous: After all, the ordinary psychology of effort minimization should make everyone a tactical free rider. In all conditions, including exchange, motivational adaptations in humans should have been selected to identify and curtail unnecessary investment (i.e. units of effort whose investment does not trigger or sustain returns). Hence, our evolved cognitive equipment should connect the motivation to expend effort in an exchange to an assessment of the sensitivity of others' responses to one's own contributions, both upwards and downwards. That is, do they increase their contributions when I increase mine? Do they downregulate their contributions when I downregulate mine? Are they appreciative if I sacrifice for the joint effort? Do they even notice if I do not?

As a result, the motivational component should have been designed by evolution to modulate one's own level of compliance to one's obligations in an exchange in response to others' levels of compliance (or at least to others' monitoring and eventual responsiveness). In dyadic exchange, I am no longer motivated to live up to my part of the exchange when my partner is not living up to hers. In *n*-person exchange, motivational levels should be a function of the distribution of contributions by others. In general, such circuitry motivates greater effort when others are investing sufficiently (and contingently), and curtailed effort in the presence of free riding (or indiscriminate investing). This set of conditional rules regulating the motivation to participate is an important first line of defense against being exploited by others (although it permits the conditional exploitation of others).

Accordingly, our evolved motivational system should be designed to recognize situations where others systematically benefit from one's own sacrifices without contributing proportionately. Such situations should be cognitively represented as distinct from other situations, and marked as intrinsically motivationally significant. The detection of such a situation (or its possibility) should then trigger motivational outputs designed to effectively redress the potentially adverse fitness ordering manifested in an exploitive situation (i.e. some property of the cooperative strategy must insure that on average, noncooperative exploitive

strategies have lower fitness, at least when they become common). Hence, procedures for social comparison of effort and benefit should be very important motivationally. Of course, the underlying logic of exchange involves delivering benefits only when the cost of providing them is less than the benefit produced—no one expects you to cut off your arm to make dinner for a friend. Given the wide range of cooperative endeavors humans engage in, this means that cooperators will usually need to represent not just behavior but also the underlying costs and benefits that are associated with behaviors. If your child is starving, I do not expect you to give me food, even if I previously fed you. We operate in a world where the inferred costs and benefits of actions are relevant to our understanding of our exchange relations.

Accordingly, to engage in *n*-party exchange behavior, the system needs to represent the accounts *own welfare* and *group welfare* in order to make decisions governing one's own allocations between these two accounts. To do this, the mind needs to compute a regulatory variable, the *welfare trade-off ratio* (WTR), that governs what trade-offs (sacrifices) the individual will be willing to engage in that have a negative impact on *own welfare* and a positive effect on *group welfare* (and *vice versa*). To protect itself against exploitation, the system also needs to be able to monitor and represent events and actions in terms of their significance with respect to others' welfare accounts (e.g. the contributions others make, the costs they incur in making contributions, the benefits they derive from the exchange, the impacts I have on them, etc.). These variables need to be tracked both for individuals, and (for some purposes) on a pooled basis. This enables the motivational system to be able to compare one's own welfare trade-off ratio to the welfare trade-off ratio that others are using, and therefore to determine whether the actor is being exploited by others (or is exploiting others). Our motivational systems switch into different modes of activation depending on the answer to questions like: *Are others contributing less* (*more*) *than I am*? (e.g. Price *et al*., 2002; Price, in press B).

The evolved aversiveness of low payoffs in interactions should be a function of comparative payoff, not just absolute payoff. In rational choice theory, any payoff is better than no payoff, but in the evolutionary competition of alternative designs, a positive absolute payoff that is a low relative payoff should (under many conditions) be selected to be perceived as worse than a zero absolute payoff that is equally bad for both participants. Widespread evidence from experimental economics suggests that this is the case, such as the high level of rejection of offers in the ultimatum game (Hoffman *et al*., 1998; Henrich and Smith, 2004). Indeed, the sucker's payoff (in prisoner's dilemma situations) should not just be aversive because it is low paying in an absolute sense, but because it is *relatively* low paying compared to the gains of the other interactant. So, algorithms in the brain characterize situations in terms of whether others are benefiting from the individual's own sacrifices in a way that is disproportionate to their contributions. For convenience, we will call this an evolved characterization of the degree of *personal fairness* or *personal exploitation* exhibited by a situation. As history and evolutionary theory both attest, evolved mechanisms in the mind do not have an automatic appetite for *global* fairness of all to all (including sacrificing personal gains to be fair to others). Our evolved mechanisms are not designed to intrinsically object to one's own unfairness to third parties (although in the right social environment such attitudes may emerge). Our evolved exchange psychology is designed instead to manifest a distaste for exploitation or unfairness by others to oneself, as well as by others to family members, to a friend or ally, to an ingroup member, or (through putting a group into the agent slot) to personally valued groups (see, Fehr and Fischbacher, 2003, for a different view).

There are several distinct motivational outputs from this system, when it recognizes a situation of personal exploitation, in addition to contribution downregulation. The balance of costs and benefits arising from the situation the individual is in determine which motivational output and behavioral response will dominate. Most fundamentally, the exploitive aspect of a situation should be experienced and represented as intrinsically unpleasant or aversive, independent of the other payoffs of the situation, so that the prospect of removing exploitation presents itself as a goal (if achievable) with a positive payoff. Not only should exploitation be aversive, but individuals who consistently exploit more than others should themselves come to be seen as aversive. Other motivated responses that defend against adverse selection favoring free riders include: punishing

free riders (see, e.g. Price *et al*., 2002); avoiding individuals who differentially undercontribute; leaving situations in which one is being exploited; refusing to initiate exchanges in which exploitation is likely to occur; refusing to continue in exchanges in which exploitation is occurring; aggressively forcing those disposed to free riding to contribute; and driving off those disposed to free riding. It is unparsimonious to assume that all of these responses were independently selected for. Rather, it seems plausible that many or all these responses are generated by a single underlying motivational element: finding personal exploitation intrinsically aversive.

Opening Cooperative Moves and Subsequent Monitoring Forms in *n*-person Exchange

Because mutual benefit through coordinated *n*-party exchange is hard to achieve and easy to undermine, joint contribution levels will often be far lower than optimum. The recurrent opportunity to capture these underrealized benefits selected for design elements that promote upward movement. Perhaps the most important of these included an initial cooperative orientation parallel to tit-for-tat's opening cooperative move (Tooby and Cosmides, 2000). This orientation involves a readiness to make the first move (at least where the number of parties is not too large). This potentially initiates new upward movements at the beginning of any new event boundary that plausibly invites new collective projects, coupled to a decision-rule to modulate downwards if the effort goes unmatched. Additionally, we recognize and respond to others' supernormative initiatives (we can be 'inspired' by others, a complement to our being cooled by others' hypocrisy or defection). We are proud of our own supernormative efforts, especially if they successfully invite others to new levels of commitment, and shamed if our efforts are discovered to be lower than others, and to have inhibited positive interactions. As a general rule, the motivational system should be more willing to make an investment if it is public; investments should preferentially be made in a continuous flow of consecutive increments (where this is not inconsistent with public delivery), so that their magnitude can be modulated contingent on others' degree of matching—a pattern that reduces the opportunity for free riding. Once a higher contribution level is made by one participant, others must match it, or their failure to do so will establish the future collective ceiling below the optimum.

So, our evolved exchange psychology contains procedures that motivate participation in joint efforts when they can be detected and assessed as beneficial. Defense against exploitation requires monitoring of others' actions and their contingencies. We believe that different structures of monitoring have selected for different modes of *n*-party exchange. These include *joint monitoring* (where everyone in a collective enterprise monitors everyone else), *leadership monitoring* (where single individuals—'leaders'—monitor and direct enforcement), *asymmetric monitoring* (where multiple individuals with disproportionate stakes in the joint effort or lower costs of enforcement regulate the effort), *role monitoring* (where specialists cultivate a socially valued identity based on monitoring), or some mixture of them. Joint monitoring was perhaps the first to evolve, but depends on the practical opportunity for small sets of individuals to directly observe each other's levels of participation. All of these organizational alternatives are anchored in an evolved psychology of social comparison that evolved in the context of *n*-party exchange.

Indeed, a standard finding of experimental collective actions (public good games), is that the majority of subjects are conditional cooperators, i.e. they cooperate more when they perceive that co-players are more willing to cooperate (Orbell and Dawes, 1991, 1993; Ledyard, 1995; Fischbacher *et al*., 2001; Lubell and Scholz, 2001), and less when they believe that co-players are free riding (Fehr and Gächter, 2000; Kurzban *et al*., 2001a). Moreover, a standard finding of anonymous public good games, in which players have no information about co-player cooperativeness at the start of the game, is that cooperative behavior follows the pattern predicted from an evolutionary perspective. Contributing is highest at the outset of the game (players open with invitations to cooperate), and then decays gradually over time, as players receive information that some co-players are free riding (Ledyard, 1995; Masclet *et al*., 2003). This gradual decay apparently occurs because higher contributors in initial rounds ratchet down their contributions as the game progresses, in order to avoid exploitation by matching the average expected co-player contribution. The average constantly dwindles due to

persistent free riding (Fischbacher *et al.*, 2001; Kurzban *et al.*, 2001a,b).

Further, as expected if cooperators are sifting for exploitation, interactants appear to monitor one another not just constantly, but accurately. Among villagers in a hunter–horticultural society, perceptions of co-villagers' engagement in general pro-village altruism correlated positively with measures of co-villagers' actual engagement in specific altruistic activities (Price, 2003). In a sugarcane cultivating workgroup in this same society, perceptions of co-worker cooperativeness (attendance record and physical effort in work sessions) correlated positively with more objective measures of this cooperativeness (Price, in press A). Moreover, participants accurately distinguished 'intentional' low contributors (those who could have contributed highly but chose not to) from 'unintentional' low contributors (those unable to contribute highly).

A psychology of conditional punishment co-evolved with a psychology of n-person exchange. The evolved motivational mechanisms regulating *n*-person exchange are designed to make each individual's actions contingent on the actions of the others in the cooperative interaction. Consequently, these mechanisms endow joint efforts with dynamical properties. That is, an *n*-person exchange is a sensitive network of intercontingent feedback loops. Depending on variables such as the opportunities for free riding and the frequency of free riders, an intercontingent cooperative structure can dynamically drive itself toward higher levels of cooperation or downwards toward lower levels of cooperation or dissolution (Ehrhart and Keser, 1999; Fehr and Gachter, 2000; Masclet *et al.*, 2003). For example, if many participants are committed and enthusiastic, then the probability that others will join and contribute is increased. In contrast, the presence of undercontributors (cheaters, free riders) in the network inhibits contributions by others. Individuals are more reluctant to join collective efforts in which they can anticipate that others will be benefiting without contributing. Indeed, the presence of individuals with a track record of greater free riding should prevent many *n*-person exchanges from coming into existence in the first place. At a minimum, the presence of free riders degrades the performance of the network at realizing potential joint gains. More seriously, sufficient penetration by free riding can trigger the termination of an otherwise mutually beneficial ongoing intercontingent effort. Consequently, unless maintained by corrective social feedback, cheating or undercontribution will spread contagiously, and contributions to joint efforts will ratchet downwards (Kurzban *et al.*, 2001a). When opportunities for monitoring and feedback are inadequate, effort minimization tends to operate unopposed, and the cooperative network erodes away.

The key point is that free riding is not just an inefficiency, an injustice, or a marginal strategy. How free riding is treated is the central determinant of the survival and health of cooperative organizations. Given the existence of motivational circuitry that evolved to defend against exploitation, social penetration by free riders is a brake or even an off switch, turning cooperation down or off among individuals whose motivation to cooperate is dynamically interlinked. The social cost of free riding is not just the marginal effort that free riders do not contribute nor is it primarily the benefits they consume but do not earn: *The greatest cost that free riders inflict is the loss of all the potential gains from n-party exchanges that otherwise would have been achieved if free riding had not triggered antiexploitation motivational defenses among cooperators*. That is, by forestalling the emergence or continuation of a range of mutually beneficial *n*-person exchanges that would otherwise exist, free riders can and do inflict huge costs on cooperators out of all proportion to the parasitic benefits they derive.

One can approach the same conclusion by considering the likely sequence of steps in the evolution of the psychology of cooperation: In general, any strategy for exchange that reliably allows cheaters to outcompete cooperators cannot evolutionarily persist. But (under reasonable conditions) cheaters cannot outcompete those cooperators designed to follow a strict conditional exchange strategy involving the detection of cheaters and the refusal to enter into exchanges with them. Plausibly, the establishment of this first-order exchange strategy as part of our ancestral species-typical design was thus an early phase in the evolution of our psychology of cooperation (Cosmides and Tooby, 1989). Moreover, this conditional strategy is reasonably efficient for dyadic exchange. Individuals who abandon or refuse dyadic exchanges with cheaters are only forgoing interactions in which they are

unlikely to get benefits anyway. However, as *n* increases for *n*-party exchanges, abandonment of the exchange in the presence of a cheater becomes increasingly costly and inefficient in terms of forgone benefits. Under strictly conditional cooperation, no matter how many cooperators are present, a set of interacting cooperators can be prevented from beneficial interactions by the presence of a single cheater. Naturally, the greater the number of individuals involved, the more likely one (or more) will be a cheater, preventing the exchange (assuming an early ancestral social ecology in which individuals are equipped only with a psychology of strict exchange). Such a strategy insures that only exchanges with very small numbers of individuals will ever take place, leaving the rest as a large, unachieved residual class. Of course, despite its inability to capture gains from larger scale exchanges, a strategy of strictly conditional exchange would nevertheless have been strongly favored by selection because it did allow capturing gains from smaller scale exchanges. Indeed, for our hunter–gatherer ancestors, a substantial number of exchanges were very small-scale—dyads, triads, tetrads, and so on. Even now, in modern mass society, the great majority of daily cooperative interactions involve only a handful of individuals.

In short, a strictly conditional exchange strategy would have evolutionarily emerged and been propelled upward toward species-typicality among our ancestors because (1) it cannot be outcompeted by an exploitive strategy, (2) it is able to take advantage of any *n*-person exchange not involving a free rider, and (3) opportunities for such exchanges were ancestrally ubiquitous. This makes this strategy fitter than strategies that do not engage in exchanges, or that free ride (cheat). We think that these selection pressures established something resembling a strictly conditional psychology of exchange among our ancestors as a platform out of which a more complex psychology of cooperation subsequently evolved. Accordingly, we should not expect to see a first-order cooperative psychology whose motivated contributions are unaffected by the presence of free riders, because such mutant designs would have been exploited and outcompeted. What we predict on selectionist grounds is what we observe empirically: The presence of free riders downregulates cooperative behavior.

Because the detection of free riders by cooperators inhibits contribution to joint efforts, free riders stand in the way of reaching valuable gains through cooperation. The first line of defense for conditional cooperators—abandoning the exchange in the presence of free riders—is very costly to conditional cooperators. As a result, the presence of free riders in cooperative contexts was ancestrally (as it is now) a serious adaptive problem facing conditional cooperators. If they exist, selection would have strongly favored the subsequent evolutionary emergence of psychological mechanisms that (1) can realize gains from *n*-person exchanges even in the presence of cheaters, while simultaneously (2) preventing cheaters from outcompeting cooperators. Are there any such designs?

An evolutionarily tailored motivational system directing punitive sentiment can help to solve the adaptive problem posed by free riders (Price *et al.*, 2002). If the presence of a small number of free riders (or one) is inhibiting optimal contribution levels to an *n*-person exchange, then punishment directed at the individual will drive her off, or induce her to contribute. In either case, the situation is no longer psychologically marked for cooperators by personal exploitation, and so they are no longer inhibited from contributing effectively to the exchange. Equally, in selectionist terms, free riders are no longer outcompeting contributors. That is, it is the dynamic sensitivity of *n*-person exchanges to the presence of exploitation that selects for punishment. Rather than inefficiently removing the exploitation by depriving a number of potential cooperators of the benefits of *n*-person exchange, punishment can be efficiently targeted exactly at the source of the problem, the exploitive individuals themselves. Hence, in an ancestral social environment of potential *n*-person exchanges, the evolution of mechanisms directing punitive sentiments toward free riders was favored by natural selection. Because punitive acts and the evolved dispositions that cause them unlocked access to a wide range of previously unattained cooperative gains, the circuitry underlying this punitive psychology was advantageous. These fitness advantages increased its frequency until it became (we believe) a significant part of our species-typical social psychology of cooperation. Widespread experimental evidence indicates both that humans do have punitive sentiments toward exploiters in *n*-person exchanges, and that *n*-person exchanges can be driven to higher levels of cooperation when subjects have the ability to punish (Fehr and

Gachter, 2000). Although the strategy of avoiding cheaters can work as the sole defense against exploitation when exchange interactions are small (dyads, triads, tetrads, etc.), the larger the size of the interaction, the more efficient punitive cooperation strategies become at sustaining cooperative organization.

The design features and evolutionary dynamics of a psychology of punitively defended conditional cooperation. For the reasons outlined above, we think that the human mind contains an evolved, functionally specialized motivational mechanism that, when exposed to a situation of personal exploitation, generates a punitive sentiment toward the agent that is deriving an unfair advantage in an exchange. We think evidence shows that this mechanism becomes strongly activated in collective actions, and evolved as an anti-free rider device (Price *et al*., 2002, in preparation; Price, in press B; Tooby and Cosmides, 2002, forthcoming). In *n*-person exchanges, this sentiment motivates individuals to inflict a cost on free riders (exploiters) sufficient to reorder the net benefits that the interactants derive from the exchange (when this can be done cost-effectively). That is, in the absence of corrective action, free riders are better off than cooperators. After corrective (punitive) action, free riders are made worse off than contributors. The functional product of punishment is an outcome in which the most exploitive individuals either end up with a lower welfare than contributors, or end up less well off than they would have been if they had not attempted to derive benefits from the collective effort. The presence of punitive strategists (punitive cooperators) changes the payoff structure for free riders, influencing their choices (to the extent they respond to anticipated payoffs). When dispositional free riders (individuals psychologically designed to not contribute) detect that one or more punitive individuals are to be in an *n*-person exchange interaction, they should avoid the interaction—which will be low paying and hence aversive for them. When facultative (tactical) free riders detect interactions involving punitive strategists, they should avoid the interaction or should tactically contribute. At present, the weight of evidence appears to support the view that the primary function of punitive sentiment is as an anti-exploitation defense (Price *et al*., 2002; Price, in press A, B). Nevertheless, more work will be required to comprehensively rule out the hypothesis that punitive sentiment was secondarily designed by selection to serve as a system for recruiting additional labor into *n*-person exchanges.

What benefits do punitive strategists derive from their evolved motivational design and the expenditures of effort it causes them to make? Why would a cooperator equipped with punitive motivation toward exploiters be fitter than cooperators without punitive motivation? Punitive strategists switch on the productive possibilities of the groups they are in, unleashing collective efforts that would otherwise be inhibited by the presence of free riders. This happens because the presence of punitive strategists in potential exchange interactions repels free riders, causing them either to avoid such interactions or to become (facultatively, in the presence of punitive strategists) behavioral cooperators. Because free riders avoid punitive strategists, punitive strategists will far more often find themselves in groups without free riders. Punitive strategists, unlike those who are simply strictly conditional cooperators, do not have to forgo large numbers of opportunities to harvest a valuable joint gain. Even more important, the advantage of a punitive strategy applies at small scales even when it is rare. For example, it will be less often cheated even in dyads, because the cheating triggers punishment, and so the temptation to defect is lowered by the anticipation of punishment[7]. Punitive strategists can even afford to be more trusting, and can more successfully enter into one-shot dyadic exchanges (provided their interactants can anticipate that their partners will have a subsequent opportunity to punish). In sum, anti-free rider punitive strategies unlock access to a wide range of *n*-person exchange interactions that would otherwise have gone unrealized.

A number of scholars, such as Boyd and Richerson, argue that punishment emerged through cultural or biological group selection or as a product of gene–culture coevolutionary dynamics—i.e. that punishment of norm violation is altruistic in the biological sense (Boyd and Richerson, 1992; Boyd *et al*., 2003; Henrich, 2004). The central argument is that mutant punitive cooperators and nonpunitive cooperators would share the same group-wide benefits of punishment, but that only the punitive would bear the costs of inflicting punishment. So, on this view,

nonpunitive cooperators would outreproduce punitive strategists in the same group in the same way that free riders displace indiscriminate cooperators. The punishment of free riders is viewed as a collective action problem (Yamagishi, 1986), and the evolution of punishment is considered to suffer from a second-order free rider problem. Therefore, according to this line of thinking, motivational mechanisms designed to punish free riders cannot evolve without group selection or gene–culture coevolutionary dynamics (Boyd and Richerson 1992; Boyd *et al*., 2003; Henrich, 2004; see also Panchanathan and Boyd, 2004).

The objection that punishment in the context of collective actions suffers from a second-order free rider problem is an important argument, worthy of serious attention. However, the validity of this argument depends on the nature of the ancestral distribution of opportunities for cooperation, on other pre-existing components in our evolved social psychology, and on the exact nature of the computational procedures that implement the punitive psychology. We think, for example, that it is highly likely that the ancestral frequency of opportunities for exchange was a declining function of the number of individuals involved. That is, there were ubiquitous opportunities for dyads, frequent opportunities for triads, fewer for tetrads, even fewer for pentads, and so on. In normal hunter–gatherer social ecologies, twenty people rarely shared joint attention on a common project, while ten people sometimes did, while five people often did, and two people commonly did. In such a size-skewed social ecology, selection for a punitive strategy could easily proceed (1) when the benefit to punitive strategists of increased gains from participating in *n*-person exchanges sufficiently exceeds the costs of punishment, and (2) when the costs of punishment for punitive strategists are sufficiently less in their successful exchange interactions than the costs nonpunitive cooperators incur from more often finding themselves in failed exchange interactions (Tooby and Cosmides, 2002). Nonpunitive cooperators will frequently find themselves in exchange interactions that failed because there were no punitive strategists in the interaction to defend against free riders and the undercooperation they trigger. Punitive cooperators, of course, always find themselves in interactions that include a punitive cooperator, while nonpunitive cooperators only sometimes do. If all exchange interactions involved very large numbers of individuals, then mutant punitive strategists would indeed have a hard time outcompeting nonpunitive cooperators, because both would get the benefits while only the punitive would pay the costs. But because ordinary sociality consists of numerous daily interactions composed of dyads, triads, and so on, nonpunitive individuals will commonly find themselves in potential exchange interactions without the punitive individuals, even when the relative frequency of punitive individuals is high. For example, whenever a nonpunitive cooperator is in a dyad with a free rider, they will not be in a dyad with a punitive strategist. Nonpunitive cooperators will commonly be trapped in unproductive triads without a punitive strategist and with at least one free rider, until punitive strategists become very common.

In short, the existence of numerous small-scale exchange interactions which fail without punitive cooperators but succeed with them can drive a strategy of punitive cooperation to high frequencies just by individual selection (in the ordinary sense[8]). Punitive cooperators outcompete both nonpunitive cooperators and free riders because they far more commonly find themselves in potential cooperative interactions without free riders and their inhibiting effects. Consequently, punitive cooperators are successful at realizing a far broader range of gains from *n*-party exchanges than are rival strategies. They are not vulnerable to the abiding weakness of a strictly conditional exchange strategy, because interactive contexts that initially include one or more free riders cannot inhibit them from harvesting valuable joint gains. The selection pressure created by free riders persists across evolutionary time because our evolved psychology of effort minimization makes a reversion toward free riding a recurrent phenomenon in the absence of corrective social feedback provided by punitive strategists. We think, therefore, that this history of selection has constructed a sophisticated cooperative psychology that includes punitive motivation against exploiters.

Of course, our evolved punitive psychology cannot be indifferent to the costs of inflicting punishment, or to the efficiency with which effort devoted to punishment affects the landscape of exploitation. Our evolved punitive psychology should be designed to be sensitive to both of these variables. Moreover, the efficiency of punishment will be affected by how many others in the exchange will participate in the punitive

enforcement of the exchange. The greater the number of individuals that collaborate in punishment, the smaller the cost of punishment for each individual, and the more efficient punishment will be. Therefore, our punitive circuits should also be designed to invite others to participate in punitive enforcement, by preceding punishment with exposures of exploitation, expressed disapproval or outrage, solicitations of social support to respond to exploitation, and monitoring for a sufficient mutuality of sentiment to make punitive action cost-effective.[9] Social life is riddled with confidential complaints about the misbehavior of third parties, as individuals sound each other out about shared views, and implicitly about various potential collective punitive enterprises. Moreover, the size of the exchange interaction will be an important variable impacting the cost-effectiveness of individual punitive action. A single blindly activated punitive strategist in a large group containing many free riders is unlikely to be effective, and therefore would be selected against. Therefore, early in our social evolution, when punitive strategists were rare, such a strategy would have only been favored to the extent that circuitry was designed to activate punishment only where individual action would be efficacious—that is, in groups that were sufficiently small. However, as punitive strategists became more common, the costs of punishing would have been shared among them, allowing the threshold of activation to be relaxed so that punishment was deployed in larger and larger groups. Still, other things being equal, punitive sentiment should be stronger (i.e. the amount of cost an individual is willing to incur in order to punish should be larger) in smaller exchange interactions than in larger exchange interactions. Equally, the greater the number of individuals who collaborate in punitive enforcement, the more readily individuals will become punitively activated, other things being equal.

In short, our evolved punitively cooperative psychology should be designed to be sensitive to a number of variables related to its function, including: the costs of punishment to the punitive strategist; the number of individuals who will share the costs of punishment; the number of free riders requiring punishment; how cost-effective punishment will be in eliminating situations of exploitation (including how effective punishment will be in reducing the excess benefits accruing to free riders); and the communicative efficiency with which coordination can be achieved in collaborative punishment.

Although our psychology of punitive cooperation evolved in the context of small-scale interactions, and derives much of its fitness advantage from the enduring pervasiveness of small-scale social interaction, the cultural emergence of rarer, larger scale cooperative contexts need not have selected against punitive cooperation designs. One reason arises from the fact that individuals are designed to be selective about who they encourage to fill the limited social niches in their lives, such as friends, cooperators, allies, and exchange partners (see, e.g. Tooby and Cosmides, 1996). Also, it is likely that the tendency to free ride or cheat in some contexts is predictive of the tendency to be exploitive in other contexts (for many reasons, such as the existence of stable individual differences in impulse control). This means that observations of how others behave in large-scale exchanges will provide some information about how they will act in small-scale exchanges. Public dereliction in large-scale exchanges should negatively influence the willingness of others to include highly delinquent individuals in advantageous, smaller scale exchanges (see Panchanathan and Boyd, 2004, for a related argument).

Asymmetric beneficiaries, the evolution of leadership, and the emergence of morality. There is an even more important dimension of *n*-person exchanges that makes the evolution of punitive strategies by individual selection nonproblematic. The economic theory of public goods and collective action is based on assuming an idealization in which the benefits that individuals receive from collective action are exactly equal. For a broad array of ancestral activities, this assumption was very unlikely to have held. Individual differences in kin arrays, status, alliances, reproductive value, strength, health, and other circumstances would commonly have created a spectrum of benefit magnitudes that differed from individual to individual for any given joint enterprise. In such a world, there would be no barrier to the evolution of a conditional punitive cooperator strategy that operated according to the following rule: Initiate and enforce an *n*-person exchange, using punitive threats against potential exploiters, whenever the cost of inflicting punishment will be less than the excess benefit the punitive strategist derives than what nonpunitive cooperators derive

(Tooby and Cosmides, 2002, forthcoming; Price *et al.*, in preperation). The punitive strategist is compensated for her excess expenditure of punitive effort by the fact that she derives a greater benefit from the interaction than the nonpunitive cooperators do. Rationally or evolutionarily, they would not be designed to resist the punitive strategist to the extent that the punitive strategist selects *n*-person exchanges that are in their interest, too. In short, by limiting the expenditure of punitive efforts only to those common enterprises that the punitive strategist has a stronger relative interest in, the punitive strategy cannot be outcompeted by a strategy of second-order free riding. Such a conditional punitive cooperator strategy reliably gains net benefits in social interaction that alternative strategies such as noncooperation, strict conditional cooperation, and dispositional free-riding do not. In sum, we conclude that a psychology of punishment in exchange became part of our species-typical design over evolutionary time, although it is expected to be differentially activated in different individuals because of different individual circumstances and different developmental trajectories.

What does the hypothesized presence of such a psychology predict? Those contributors to an *n*-person exchange who stand to gain the most from its success will be differentially involved in organizing and expending coordinative effort: punishment, encouragement, persuasion, and so on. Equally, those for whom the cost of inflicting punishment is relatively less (i.e. more powerful or formidable individuals) will also be more likely to find themselves in the envelope of conditions where punishment is cost-effective. These provide some of the kernels out of which a theory of the nature of leadership can be developed.

As discussed, leadership spontaneously emerges from a differential ability to solve the coordination problems involved in *n*-person exchanges, including the threat of dissolution through individuals choosing to undercontribute. Consequently, one element of leadership is the ability to effectively deploy punishment to prevent potential participants from defecting on a joint cooperative venture. To the extent that an individual is more formidable (i.e. for whatever reason, can inflict necessary punishment at lower costs), she is better positioned to become a leader. To the extent an individual is a more effective communicator (and so can communicate the benefits and requirements of a coordination), she is better positioned to become a leader. To the extent that an individual derives an asymmetrically greater benefit from a given collective effort, she will be more motivated to expend coordinative effort leading the group to achieve the group gain.[10] Equally, a leader needs the knowledge and intelligence to make those coordinative decisions that lead to success in a joint enterprise. That is, she must be cognitively equipped to determine what causal steps are needed to successfully orchestrate the intended joint outcome. Gains from *n*-person enterprises can be achieved or lost depending on whether the plan corresponds to reality. Finally, a leader needs to choose, out of the available array of alternative plans of action, those coordinated efforts that (1) benefit her, while also (2) sufficiently benefiting a large enough subset of the potential participants that they (3) support the joint project strongly enough for it to succeed. Social groups are confronted with an indefinitely large array of potential joint enterprises, each with its own matrix of payoffs. A leader must be able to compute how different proposals will differentially impact the perceived welfare of potential participants. A leader who proposes enterprises that differentially benefit herself at the expense of others will not be supported by others, while an individual who sacrifices her own interests too strongly for the enterprise will not be motivated to expend effort to act as leader. To be successful, a leader must be able to anticipate, perceive and represent the perspective-specific values of the interactants, in order to converge on those collective enterprises with sufficient political support. Persuasion involves the ability to discover and represent the values of others, and to awaken a sense of how personally valuable the anticipated outcome of the enterprise will be to the audience. Of course, to the extent that an individual naturally and inextricably benefits from the same set of outcomes that most other group members do (although perhaps to a greater degree), she is better positioned to become a leader. Of course, leadership is not an all or nothing matter: All participants actively or passively play a role in directing coordinative effort, instigating or inhibiting action, and otherwise exercising social influence on the direction that the collective enterprises we participate in will take.

Finally, it is worth briefly considering how these proposals concerning the evolved psychology of

n-person exchange provides a theory of the transmission of moral precepts in organizations and communities. We think that the willingness to follow many classes of moral rules within communities and organizations is implicitly treated by the mind as a form of *n*-person exchange. Logically speaking, there need be no relationship between whether some individuals follow a rule and whether others do, if the goal is to be ethical. Yet, we implicitly treat many rules as if following them were somehow conditional on others' conduct. Many people feel sentiments according to the following social exchange logic: I will give up the benefits of violating this moral rule if others in my social world do. If I followed the rule, and you did not, I have been cheated by you. The more others cheat on a rule I follow, the more exploited I feel, and the more tempted I am to discontinue following the rule when it is costly to do so. Whether or not we act on it, we feel a link between our motivation to follow moral rules and others' adherence to them. The hypothesis that morality is treated psychologically as a collective exchange also provides an explanation for our otherwise irrational response to hypocrisy. Although the value of an argument is logically unrelated to who says it, individuals act as if their willingness to adopt a moral proposal depends on whether the proposer follows it herself. Nothing destroys a leader's credibility faster than the discovery that she has been a hypocrite. It is as if a cheater proposes a social exchange: Our psychology is designed to avoid entering into exploitive relationships, and adopting a moral rule from one who is not following it maps into our exchange psychology as an instance of allowing oneself to be cheated. This hypocrisy circuit makes no sense logically, but makes evolutionary psychological sense to the extent that morality in organizations and communities is one expression of our exchange psychology.

CONCLUSIONS

The picture of *n*-person exchange psychology that we have sketched here is of course highly simplified both computationally and game theoretically. For example, we have not been able to discuss a number of game theoretic complexities involving the strategy sets against which punitive cooperator strategies prosper, the likely effects of frequency-dependent dynamics on the evolution of the mechanisms, and the necessary additional cognitive abilities and motivational circuits humans must have for conditional cooperation, punishment, leadership, and morality to emerge. Nevertheless, we think that this treatment can provide some insights into how humans are psychologically designed behave in organizations, and what variables will govern the dynamics, success, and failure of various organizational forms.

Acknowledgements

Thanks to Pat Barclay, Will Brown, Oliver Curry, Ed Hagen, Lin Ostrom, Jade Price, Steve Rothstein, Don Symons, Masanori Takezawa, Robert Trivers, members of the Behavioral Science Group at the Santa Fe Institute, and members of the Center for Evolutionary Psychology Research Seminar. Funding was provided by the Indiana University Workshop in Political Theory and Policy Analysis, and by a grant to the Santa Fe Institute from the James S. McDonnell Foundation-21st Century Collaborative Award for the Study of Complex Systems.

NOTES

1. It is better not to think of economic rationality as an explanation for behavior—that is, as a set of causes of anything. That is, theories of economic rationality are high level *post hoc* descriptive idealizations of outputs, rather than a model of a real process that produces or explains behavior. As economics, organizational behavior, and psychology mature and integrate, the ground level theories will be detailed descriptions of the information-processing architectures of the mechanisms in the human brain, and how they interact in sets in structured environments. Using rationality as an account of choice is parallel to using animism as an account for the ability of animals to move—admirably descriptive, while failing to track or connect with the science of actual causation.
2. One can view the phenomenon of increasing conservatism with increasing scale either negatively or positively, as either cognitive inertia or the coordinative facilitation of pre-existing coordinative structures. On the plus side, far more large-scale cooperation is realized than otherwise would exist if humans had to build exchange structures *de novo* for each individual opportunity for an interaction to realize a gain in trade. On the downside, the fact that the revision of pre-existing practices is increasingly retarded the larger the interaction structure means that larger organizations increasingly grandfather in obsolete and nonfunctional features. Conversely, the smaller the interaction set, the more dynamic it will

be, and the more rapidly it will be tailored to meet changing conditions. The larger the structure, the more processes of cultural transmission will be important, while the smaller the structure, the more game-theory like interactive responses to existing conditions will predominate. This principle explains why small, young companies more effectively exploit new economic opportunities, despite the capitalization advantages large companies enjoy.

3. By parallel logic, it should be easier to grow a coordinated *n*-person exchange structure from a smaller to a larger size than to create it *de novo* at a large size.
4. This is true even of riots, which usually outgrow the initial focus or pretext and become a wide-ranging opportunity to realize those joint values of the rioters that are usually individually deterred by a lack of sufficiently coordinated collective effort.
5. Socially marginal individuals should be more eager to covertly feed disputes among others.
6. Cults, for example, place a strong emphasis on cutting ties to families and other outsiders (e.g. Lalich, 2004), and basic training isolates new recruits from family members for the first several months of training.
7. We consider anger to be the expression of an evolved emotion program (a functionally structured neurocomputational system) whose design features and subcomponents evolved to advantageously regulate thinking, motivation, and behavior in the context of resolving conflicts of interest in favor of the angry individual. Two negotiating tools regulated by this system are the threat of inflicting costs (aggression) and the threat of withdrawing benefits (the downregulation of cooperation). Humans have a system that, in each individual, recognizes the welfare trade-off ratio expressed in the actions another party takes with respect to oneself. Anger is conceptualized as a mechanism whose functional product is the recalibration in the mind of another of this other person's welfare trade-off ratio with respect to oneself (or other valued agent). That is, the goal of the system is to change the targeted persons' disposition to make welfare trade-offs so that they more strongly favor the angered individual (or their group) in the present and the future. As in animal contests, the target of anger may relinquish a contested resource, or may simply in the future be more careful to help or to avoid harming the angered individual. In cooperative relationships, where there is the expectation that the cooperative partner will spontaneously take the welfare of the individual into account, the primary threat from the angered person that potentially induces recalibration in the targeted individual is the signaled possibility of the withdrawal of future help and cooperation if the welfare trade-off ratio is not modified. In the absence of cooperation, the primary threat is the infliction of damage. Concepts that are anchored in the internal psychological variable *welfare trade-off ratio* include respect, consideration, deference, status, rank, and so on. Anger is an ancient system, and provides much of the evolved infrastructure out of which the *punitive sentiment* system evolved. Indeed, to the extent that our evolved psychologies are designed to recognize that free riding inhibits *n*-person exchanges and the gains they generate (while simultaneously benefiting the free rider) the anger system would process this as a situation in which the free rider is trading off the welfare of others for her welfare. To the extent that the exploiter is inflicting unacceptably high costs in pursuit of insufficient personal gains, this should trigger anger at exploitation. That is, anger and punitive sentiment are closely related functional systems, using much of the same infrastructure.
8. *Individual selection* and *group selection* have come to acquire such a broad variety of meanings that it is difficult to speak briefly about them with any clarity. David Sloan Wilson, for example, uses such a broad definition of group selection that dyadic reciprocation and kin selection both become examples of 'group selection' because they involve interactions in 'groups' (Wilson, 1980). We prefer to restrict the term *group selection* to those cases where the higher level ('group' derived) component of the fitness of the organism derives from an organism's stable membership in a single group for a majority of its life history, rather than from its transient participation in thousands of different 'groups' or (as we would put it) interactions. In our view of the evolution of *n*-person exchange, different individuals (embodying different designs) accumulate fitness differences because of how they act across large numbers of different social interactions. That is, a person might engage in 20 dyadic interactions in a day, and 8 triadic interactions, and 1 twenty-person interaction. Confusion arises by labeling every social interaction a group.
9. A strategy is defined computationally by the detailed specification of a behavior control program, including how it represents conditions in order to regulate behavior. The nature of the implementation matters: Consider, for example, a program design for a punitive strategist that sums both punitive effort and cooperative effort together into one unitary account as its definition of the contribution it considers either self or others makes to the collective enterprise. This particular design will not suffer from a second-order free rider problem. The punitive strategist will contribute less nonpunitive effort to make up for the additional punitive effort it is contributing. Alternatively, it could insist that nonpunitive cooperators make a larger nonpunitive contribution to make up for the additional punitive effort it is expending. Indeed, the second-order free rider problem only exists to the extent that participants do not compute punishment as itself a contribution. But punishment is in reality a contribution, and punitive strategists (as well as rational cooperators, if any) will view it as such. In any case, a mutant design that considers punishment a cooperative contribution cannot be exploited by second-order free riders. It may prosper, or not, depending on the

details of how conditional cooperation is implemented in nonpunitive cooperators.

10. It is a commonplace of democratic politics that many proposals that would modestly benefit large majorities of the population are not acted on, in favor of other proposals that strongly benefit far smaller fractions of the population.

REFERENCES

Axelrod R, Hamilton WD. 1981. The evolution of cooperation. *Science* **211**: 1390–1396.

Baron-Cohen S, Leslie AM, Frith U. 1985. Does the autistic have a 'theory of mind'? *Cognition* **21**: 37–46.

Boyd R, Gintis H, Bowles S, Richerson PJ. 2003. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences* (*USA*) **100**: 3531–3535.

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

Brase GL. 1998. Human reasoning about social groups: evidence of evolved mechanisms for a coalitional psychology. *Dissertation Abstracts International*: *Section B*: *The Sciences and Engineering* **58**(9-B): 5150 (UMI# 9809613).

Brewer MB. 1991. The social self: on being the same and different at the same time. *Personality and Social Psychology Bulletin* **17**: 475–482.

Cosmides L. 1989. The logic of social exchange: has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition* **31**: 187–276.

Cosmides L, Tooby J. 1989. Evolutionary psychology and the generation of culture, Part II. Case study: a computational theory of social exchange. *Ethology and Sociobiology* **10**: 51–97.

Cosmides L, Tooby J. 1992. Cognitive adaptations for social exchange. In *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, Barkow JH, Cosmides L, Tooby J (eds). Oxford University Press: New York; 163–228.

Cosmides L, Tooby J. 2000. Consider the source: the evolution of adaptations for decoupling and metarepresentation. In *Metarepresentations*: *A Multidisciplinary Perspective*, Sperber D (ed.). Oxford University Press: New York; 53–115.

Cosmides L, Tooby J. 2005. Neurocognitive adaptations designed for social exchange. In *The Handbook of Evolutionary Psychology*, Buss DM (ed.). Wiley: New York.

Darley JM, Latane B. 1968. Bystander intervention in emergencies: diffusion of responsibility. *Journal of Personality and Social Psychology* **8**: 377–383.

Dumas A. 1844/2001. *The Three Musketeers*. Random House: New York.

Ehrhart K-M, Keser C. 1999. *Mobility and Cooperation: On the Run*. Scientific Series. CIRANO: Montreal.

Fehr E, Fischbacher U. 2003. The nature of human altruism. *Nature* **425**: 785–791.

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

Fischbacher U, Gächter S, Fehr E. 2001. Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters* **71**: 397–404.

Fiske, AP. 1991. *Structures of Social Life: The Four Elementary Forms of Human Relations*. Free Press: New York.

Henrich J. 2004. Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior and Organization* **53**: 3–35.

Henrich J, Smith N. 2004. Comparative experimental evidence from Machiguenga, Mapuche, Huinca and American populations. In *Foundations of Human Sociality*, Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, Gintis H (eds). Oxford University Press: Oxford; 125–167.

Hoffman E, McCabe KA, Smith VL. 1998. Behavioral foundations of reciprocity: experimental economics and evolutionary psychology. *Economic Inquiry* **36**: 335–352.

Kurzban R, McCabe K, Smith VL, Wilson BJ. 2001a. Incremental commitment and reciprocity in a real time public goods game. *Personality and Social Psychology Bulletin* **27**: 1662–1673.

Kurzban R, Tooby J, Cosmides L. 2001b. Can race be erased?: Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences* **98**(26), 15387–15392 (December 18, 2001; Epub 2001 Dec 11. PMID: 11742078).

Lalich J. 2004. *Bounded Choice*: *True Believers and Charismatic Cults*. University of California Press: Berkeley.

Ledyard JO. 1995. Public goods: a survey of experimental research. In *The Handbook of Experimental Economics*, Kagel JH, Roth AE (eds). Princeton University Press: Princeton; 111–194.

Lubell M, Scholz JT. 2001. Cooperation, reciprocity and the collective-action heuristic. *American Journal of Political Science* **45**: 160–178.

Madison J. 1788/1982. Federalist paper no. 10. In *The Federalist Papers* (2nd ver), Hamilton A, Jay J, Madison J (eds). Bantam Classics; Reissue edition: New York.

Masclet D, Noussair C, Tucker S, Villeval M. 2003. Monetary and nonmonetary punishment in the voluntary contributions mechanism. *American Economic Review* **93**: 366–380.

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

Olson M. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press: Cambridge.

Orbell J, Dawes R. 1991. A 'cognitive miser' theory of cooperators' advantage. *American Political Science Review* **85**: 515–528.

Orbell J, Dawes R. 1993. Social welfare, cooperators' advantage, and the option of not playing the game. *American Sociological Review* **58**: 787–800.

Ostrom E. 1990. *Governing the Commons*: *The Evolution of Institutions for Collective Action*. Cambridge University Press: New York.

Ostrom E. 1998. A behavioral approach to the rational choice theory of collective action. Presidential Address, American Political Science Association, 1997. *American Political Science Review* **92**: 1–22.

Panchanathan K, Boyd R. 2004. Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* **432**: 499–502.

Price ME. 2003. Pro-community altruism and social status in a Shuar village. *Human Nature* **14**: 191–208.

Price ME. in press A. Monitoring, reputation and 'greenbeard' cooperation in a Shuar work team. *Journal of Organizational Behavior*.

Price ME. in press B. Punitive sentiment among the Shuar and in industrialized societies: cross-cultural similarities. *Evolution and Human Behavior*.

Price ME, Cosmides L, Tooby J. 2002. Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior* **23**: 203–231.

Price ME, Tooby J, Cosmides L. in preparation. The evolutionary psychology of work teams.

Sidanius J, Pratto F. 1999. *Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression*. Cambridge University Press: New York.

Stone V, Cosmides L, Tooby J, Kroll N, Knight R. 2002. Selective impairment of reasoning about social exchange in a patient with bilateral limbic system damage. *Proceedings of the National Academy of Sciences* **99**(17):11531–11536.

Tiger L. 2004. *Men in Groups* (3rd edn). Transaction Publishers: New York.

Tooby J, Cosmides L. 1992a. The psychological foundations of culture. In *The Adapted Mind*: *Evolutionary Psychology and the Generation of Culture*, Barkow JH, Cosmides L, Tooby J (eds). Oxford University Press: Oxford; 19–136.

Tooby J, Cosmides L. 1992b. Ecological rationality and the multimodular mind: grounding normative theories in adaptive problems. *Center for Evolutionary Psychology Technical Report 92-1* (Reprinted in: Tooby J, Cosmides L. (in press) *Evolutionary Psychology*: *Foundational Papers*. With a foreword by Steven Pinker. MIT Press: Cambridge).

Tooby J, Cosmides L. 1996. Friendship and the Banker's Paradox: Other pathways to the evolution of adaptations for altruism. In *Evolution of Social Behaviour Patterns in Primates and Man*, Runciman WG, Maynard Smith J, Dunbar RIM (eds). Oxford University Press: Oxford; 119–143.

Tooby J, Cosmides L. 2000. Cognitive adaptations for kin-based coalitions: human kinship systems at the intersection between collective action and kin selection. *Current Anthropology* **41**: 803–804.

Tooby J, Cosmides L. 2002. The evolution of collective action: an adaptationist dissection. *Human Behavior and Evolution Society Meetings*, Rutgers University, Newark, NJ, 19–23 June.

Tooby J, Cosmides L. 2005. The functional architecture of human motivation. Paper presented at the *Human Behavior and Evolution Society Meetings*, Austin, TX.

Tooby J, Cosmides L. forthcoming. Human coalitional psychology. *Oxford Handbook of Evolutionary Psychology*. Dunbar RIM, Barrett L (eds). Oxford University Press: New York.

Tooby J, Cosmides L, Barrett HC. 2005. Resolving the debate on innate ideas: learnability constraints and the evolved interpenetration of motivational and conceptual functions. In *The Innate Mind: Structure and Contents*, Carruthers P, Laurence S, Stich S (eds). Oxford University Press: Oxford.

Trivers RL. 1971. The evolution of reciprocal altruism. *The Quarterly Review of Biology* **46**: 35–57.

Wilson DS. 1980. *The Selection of Populations and Communities*. Addison-Wesley: Menlo Park.

Williams GC. 1966. *Adaptation and Natural Selection*. Princeton University Press: Princeton.

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

