8

Groups in Mind: The Coalitional Roots of War and Morality

John Tooby and Leda Cosmides

War, Coalitions, and the Human Condition

War is older than the human species. It is found in every region of the world, among all the branches of humankind. It is found throughout human history, deeply and densely woven into its causal tapestry. It is found in all eras, and in earlier periods no less than later. There is no evidence of it having originated in one place, and spread by contact to others. War is reflected in the most fundamental features of human social life. When indigenous histories are composed, their authors invariably view wars – unlike almost all other kinds of events – as preeminently worth recording. The foundational works of human literature – the *Iliad,* the *Bhagavad-Gita,* the *Tanakh,* the *Quran*, the *Tale of the Heike* – whether oral or written, sacred or secular – reflect societies in which war was a pervasive feature.

War is found throughout prehistory (LeBlanc and Register 2003; LeBlanc 1999; Keeley 1996). Wherever in the archaeological record there is sufficient evidence to make a judgment, the traces of war are to be found. It is found across all forms of social organization – in bands, chiefdoms, and states. It was a regular part of hunter-gatherer life wherever population densities were not vanishingly low, and often even in harsh and marginal habitats. The existence of intergroup

conflict in chimpanzees suggests that our ancestors have been practicing war for at least 6 million years, and that it was a selective presence acting on the chimpanzee-hominid common ancestors and their descendants (Manson and Wrangham 1991; Wilson and Wrangham 2003; Boehm 1992). The evidence indicates that aggressive conflict among our foraging ancestors was substantial enough to have constituted a major selection pressure, especially on males (Keeley 1996; Manson and Wrangham 1991). Careful ethnographic studies of living peoples support this view (Chagnon 1983; Heider 1970). Indeed, in some ethnographically investigated small-scale societies where actual rates can be measured, a third of the adult males are reported to die violently (Keeley 1996), with rates going as high as 59 percent, reported for the Achuar (Bennett Ross 1984). Coalitions – especially male coalitions – and intergroup rivalries are a cross-culturally universal feature of human societies ranging from hunter-gatherer societies to complex, post-industrial societies. Expressions of coalitionalism include states, politics, war, racism, ethnic and religious conflict, civil war, castes, gang rivalries, male social clubs, competitive team sports, video games, and war re-enactment (Alexander 1987; Keegan 1994; Sidanius and Pratto 2001; Tiger 1969; Tooby and Cosmides 1988; Tooby, Cosmides, and Price 2006).

Our core claim is that theoretical considerations and a growing body of empirical evidence support the view that the human mind was equipped by evolution with a rich, multicomponent coalitional psychology. This psychology consists of a set of species-typical neurocomputational programs designed by natural selection to regulate within-coalition cooperation and between-coalition conflict in what, under ancestral conditions, was a fitness promoting way (Tooby and Cosmides 1988; Kurzban, Tooby, and Cosmides 2001; Price, Cosmides, and Tooby 2002; Tooby, Cosmides, and Price 2006). Ancestrally, coalitions and alliances ranged from dyads to (rarely) hundreds of individuals. Across human evolution, the fitness consequences of intergroup aggression (war), intimidation, and force-based power relations inside communities (politics), were large, especially when summed over coalitional interactions of all sizes.

These selection pressures built our coalitional psychology, which expresses itself in war, politics, group psychology, and morality. The evolutionary dynamics of war, coalitional behavior, and moral interactions are worth studying because the past world of conflict and cooperation is reflected in the present architecture of the human mind.

The Logic of Conflict

It is impossible to understand the social dynamics of collective aggression and alliance without first understanding, at least broadly, the psychological adaptations that evolved in response to the adaptive problems posed by individuals interacting with each other. It is on these foundations that subsequent adaptations for collective interactions were built.

Entropy and aggression: Aggression is the targeted infliction of disorder on one organism by another. There are two classes of benefits animals derive from aggression, and that therefore drive the evolution of control circuitry and weaponry for the targeted infliction of disorder.

The benefit of removing obstacles to fitness promotion: The first benefit occurs when the continued survival or activity of the other organism (the target) is harmful to the actor. If the target's continued survival suppresses the actor's fitness, then the actor increases its fitness by causing the death or incapacitation of the target. A typical example occurs in langurs. Langur infants whose nursing inhibits maternal ovulation are killed by the unrelated new resident male (Hrdy 1980).

Genetic relatedness and cooperative networks inside the same band and (to a lesser extent) the same tribe place restraints on the violent elimination of others whose fitness is negatively correlated with the actor. But members of other groups, outside the boundary of kinship and cooperative networks often fall into the category of fitness suppressors – for example, by virtue of occupying habitat that could benefit the aggressors, or because they threaten displacements of their own sooner or later if left unchecked. Intergroup raiding among chimpanzees (Wilson and Wrangham 2003; Boesch this volume) fits into this category. One can view neighboring groups of males locked into long-term demographic competition over productive habitat, and possibly also over the females that would be supported by it. Much raiding among small-scale human societies appears to fit into the same pattern (e.g., Chagnon 1983; Manson and Wrangham 1991; Boehm 1992). Unlike chimpanzees, however, humans also engage in more dramatic and organized wholesale slaughters and population displacements (Zimmerman 1981). History and prehistory are full of conflict-driven population displacements, and historical analysis of small face-to-face groups typically shows the same patterns (Chagnon 1983).

Aggression as bargaining power: The second class of benefit organisms accrue from aggression is bargaining power, which can be used to modify the behavior of others favorably. Obvious examples include using threats or the actuality of aggression to induce others to cede contested resources that otherwise would be monopolized by rivals; punishing others for taking actions which are fitness reducing; and deterring others from attack or exploitation. Wars among foragers commonly also have these characteristics, and power-based bargaining forms the heart of political interactions within groups.

Hate and anger as evolved computational programs: We suggest that in humans these two benefits of aggression, the elimination of fitness suppressors and bargaining, are regulated by two different motivational programs, which we will call "hate" and "anger." Hate is (1) generated by cues that the existence and presence of individuals or groups stably imposes costs substantially greater than the benefits they generate, and (2) is upregulated or downregulated by cues of relative power (formidability), and by cues signaling the degree to which one's social network is aligned in this valuation. (It is also worth investigating whether, as seems likely, there is a special emotion mode "rage" designed for combat, which orchestrates combat adaptations along with murderous motivational processes.)

We and our colleagues have proposed that anger is an evolved emotion program that evolved in the service of bargaining (Sell, Tooby, and Cosmides, 2009; Tooby, Cosmides et al. 2008). There are two bargaining tools which social organisms have available to them: (1) the threat or actuality of inflicting costs, and (2) the threat or actuality of withholding benefits. According to the recalibrational theory of anger, anger is an evolved regulatory program designed to orchestrate the deployment of these tools in order to cost-effectively bargain for better treatment and to resolve conflicts of interest in favor of the angry individual.

Welfare trade-off ratios, formidability indices, and conferral indices: We hypothesize that there are three families of computed regulatory variables that interact in the anger system to regulate decisions. The first is the *welfare trade-off ratio*, or WTR_{ij}. For a given individual *i*, the WTR regulates the weight that the actor *i* places on the welfare of a specific individual *j* compared to the weight the actor places on the self (*i*), when making decisions that have impacts on the welfare of *i* and *j* (Tooby et al. 2008; Delton et al. forthcoming; Sell et al. 2009).

The bargaining specialization outlined by the recalibrational model of anger computes the WTR it expects from specific other

to self. Its function is to elicit the maximum WTR from each specific other that it can enforce cost-effectively, given its bargaining position. This bargaining position is set by the individual's relative ability to inflict costs and to confer benefits – external variables that the cognitive architecture must internally register to regulate the individual's negotiative behavior in a fitness promoting way. Hence, the anger system uses two different families of internal variables to regulate behavior: formidability indexes, designed to track the ability of self and others to inflict costs; and conferral indexes, designed to track the ability of self and others to confer benefits. The anger system registers the formidability of self and other, and the ability to confer benefits of self and other, to set the conditions of acceptable treatment by the other.

The design and operation of the anger program: On this theory, when the anger program detects that the other party is not placing "sufficient" weight on the welfare of the actor (i.e., its WTRji is too low), anger is triggered. Indeed, experimental evidence supports the view that it is a low WTR, and not just harm per se, that triggers anger (Sell, Cosmides, and Tooby, 2009). When activated, the anger program then deploys its negotiating tactics, by the threat or actuality of inflicting costs (aggression); or where cooperation exists, by the threat or actuality of withdrawing or downregulating expected benefits. Acts or signals of anger communicate that, unless the target sufficiently increases the weight it places on the angry individual's welfare, the actor will inflict costs on, or withdraw benefits from, the target. When these anticipated or experienced fitness costs are greater for the target than the cost of placing more weight on the actor's welfare, then the target's motivational system should increase its WTR toward the actor. It will only be advantageous for the target of the anger to recalibrate its WTR_{ji} upward when the inflicted costs or withdrawn benefits would be greater than the costs of placing more weight on the welfare on the angry individual. This threshold therefore defines the conditions where anger will be effective in recalibrating the target. Because organisms are selected to pursue strategies when they are effective, this therefore also defines the conditions in which anger should be triggered in the actor. That is, the WTR that the actor considers itself "entitled to" (i.e., the level of treatment that will not provoke its anger) is a function of the actor's relative ability to inflict costs (formidability) compared to the target, or (in cooperative relationships) a function of the actor's relative ability to confer or withhold benefits. The anger system motivates

the actor to undertake actions to recalibrate the target of its anger by showing the target that it will be worse off by continuing to behave in ways that place too little weight on the actor's interests. Other things being equal, high formidability individuals are able to create incentives for low formidability individuals to assign a greater weight on their welfare.

Formidability and male combat identity: Formidability or fighting ability is the capacity to inflict costs on others (Sell, Cosmides, and Tooby 2008). Formidability is therefore a major determinant of bargaining position. Accordingly, natural selection favors the evolution of design features that enhance the ability to inflict costs – both circuitry for the effective deployment of violence, and physical structures like fangs or muscles that support successful aggression. In the human case, evidence supports the predictions from the recalibrational theory of anger that stronger men feel more entitled, anger more easily, prevail more in conflicts of interest, and more strongly approve of war as a means of settling disputes (Sell, Tooby and Cosmides, 2009).

In order to make advantageous decisions about when to persevere or defer in conflicts, humans should have evolved specializations to make accurate assessments of individual differences in formidability, and there is now strong evidence of this (Sell, Cosmides, and Tooby 2009). Moreover, humans are among the more sexually dimorphic primates particularly in upper body strength – the strength component most relevant to combat – where males are 75 percent stronger (Lassek and Gaulin 2009). Because of this, a female will almost never find herself the strongest individual in mixed sex groups, and so dispute resolution through violence or its threat tends to be a near monopoly of adult males. Across surveyed cultures and time periods, women deploy physically aggressive strategies far less often than men do (Archer 2004; Campbell 1999; Daly and Wilson 1988).

The differential use of aggression by males and females has been a long-enduring feature of human sociality. Its enduring presence among our ancestors selected for a sexually dimorphic psychology (Daly and Wilson 1988; Tooby and Cosmides 1988) in which a central constituent of masculine identity is formidability and its deployment – individual fighting ability (which may be used in dispute resolution internal to the group) and warriorship (the characteristics responsible for successful participation in intergroup aggression). The male combat identity hypothesis is the claim that in addition to whatever cultural support there is for (or against)

a masculine identity involving formidability, there is a core of evolved adaptations and sexually dimorphic calibrations in anatomy and physiology, systems of representation, and regulatory variables in motivation and emotion that orient males to cultivate an identity that navigates the challenges and opportunities of individual and collective aggression.

Males are designed by selection to be physically stronger; to threaten or deploy aggression more readily; to have sensorimotor and motivational adaptations to combat; to participate more readily and effectively in formidability-based coalitions, and to identify with them more strongly; to respond more to the potentiality of coalitional aggression by other groups; to have a more elaborated aggression-based coalitional psychology; to be aesthetically attracted to weapons and their skilled use; to be more interested in information and observations relevant to aggression; to have an appetite to improve one's formidability and maximize one's reputation for high formidability; to exhibit greater courage in potentially lethal physical encounters; to scrutinize and police others' perceptions of their formidability, status, courage, pain thresholds, competence in emergencies, and alliances; to represent others in terms of their formidability; and to be attentive to the skills and natural aptitudes in others relevant to formidability. Male status will be more based on formidability than female status. Men have more to win and to lose in intergroup conflicts. Hence, in conditions of intergroup rivalry, men should have higher evolved welfare trade-off weightings calibrated to trade-off individual welfare for group success. Broadly speaking, males should be more competitive with respect to coalitional rivalries. We expect these dimensions of male anatomical and neurocomputational architecture to be coupled together, so that they can be upregulated or downregulated by epigenetic cues (e.g., maternal condition, testosterone, methylation), as well as by neurocomputational regulation that turns dimensions of masculine combat identity up or down based on developmental environment, social context, and personal characteristics (e.g., strength). It seems likely that male combat identity should heavily overlap with a hypothesized sexually dimorphic hunting identity, an activity in which similar skills are also deployed. Both males and females should have a "theory of group mind" parallel to "theory of mind" specializations, but in males this interpretive orientation should be more easily activated and should process information relevant to alliance-based formidability more readily.

Selection for Alliances and Coalitions

Benefits flow to individuals with higher formidabilities, and one important way an individual can increase its formidability is by coordinating his or her potential for aggression with one or more others: that is, by constituting alliances and coalitions. In conflicts, the individual not only has its own formidability, but also a coalition-derived formidability that under normal conditions will be large. In general, it is plausible to suppose that when formidabilities are not too unequal, two individuals can defeat one individual, three can defeat two, and so on. Resources and reproductive advantage flow to those who form alliances over those who do not, revolutionizing the social world. Once alliances enter the social world, and individuals are no longer social atoms, individual formidability no longer necessarily generates outcomes, and linear dominance hierarchies are no longer necessarily the overriding social dimension. The game dynamics and cognitive challenges of social interactions become far more complex. The efficiency (or inefficiency) with which individual formidability could be combined into coalitional formidability would have had a major impact on ancestral human social ecologies.

The benefits of augmenting one's own formidability with coalition-derived formidability is seemingly such a general selection pressure that it poses the puzzle of why virtually all animal species are not driven to high levels of coalitional behavior. In reality, relatively few are. Why? We think that there are a series of adaptive information processing problems that must be solved if this pathway to formidability-enhancement and mutual goal realization is to be exploited (Tooby and Cosmides 1988).

When closely analyzed, the adaptive information processing problems posed by coalition formation and its associated game dynamics are numerous and difficult to solve, restricting the evolutionary emergence of robust coalitions (Tooby and Cosmides 1988; Tooby, Cosmides, and Price 2006). The actual distribution of alliances and coalitions among animal species suggests a series of answers to this puzzle (Tooby and Cosmides 1988; Tooby, Cosmides, and Price 2006). In particular:

1. Close kinship collapses or reduces many of the game-dynamical obstacles to coalition formation because inclusive fitness effects can often outweigh costs to individual direct fitness (as among social insects or, to a lesser extent, among female matrilines in primates);
2. Cognitive sophistication in large-brained social species (e.g., mammals, especially primates) preadapts certain species for evolving cognitive specializations capable of overcoming the game-dynamic obstacles to extending coalitions beyond close kinship; robust coalitions that extend beyond close kin are particularly notable features of the social life of the more cognitively sophisticated species (e.g., chimpanzees, dolphins, humans), supporting the view that the computational complexity of coalitional behavior is a key piece of the puzzle of why so many species lack robust coalitions;
3. Coalitions are easier to evolve where there are conditions that promote fitness lotteries (such as mobbing or hunter-gatherer combat); that is, where coalition members are behind a veil of ignorance as to who will pay the costs and who will reap the benefits of a coalitional event (see Tooby and Cosmides 1988, for discussion);
4. Fast-acting weaponry or tools that operate at a distance can probabilistically decouple inflicting costs from incurring costs (e.g., in carnivores such as hyenas and lions; or in humans equipped with weapons). This is one key factor precipitating coalitional fitness lotteries. Fast-acting weaponry can shorten the interval between the time the participant detects it will incur a major injury and the time it can withdraw (the harm imminence-withdrawal interval); when this interval is routinely too short to withdraw, the cost of being in a coalition is distributed as a probability cloud that collapses unforeseeably on random participants; in such cases, the aggressing individual's fitness prospects are predicted by the average payoff minus the average cost accruing to the group – an easier threshold to cross. In contrast, in close combat (such as unarmed chimpanzees engage in), striking blows invites receiving blows, and so more aggressive individuals incur higher costs. We expect that the introduction of projectile weapons by themselves should have intensified warfare, simply because of their game theoretic effects (i.e., randomizing who pays costs, and socializing the costs of combat. Changes in technology relevant to these variables should also have an impact on social structure. Stoning, thrusting spears, throwing spears, atlatls, bows, expensive bronze weapons, inexpensive iron weapons, flintlocks – these should all have changed warfare and the associated social structure. Weaponry that reduces the advantages of individual strength and/or allows multiple individuals to cost-effectively combine their formidabilities could contribute to more egalitarian social forms. Changes in social structure over human history can be broadly linked to these technological changes. More broadly, human hunter-gatherers tend to be egalitarian to some degree. This is an outcome that can be attributed to our cognitive capacities for coalition formation – a capacity that allows the many to limit exploitation that would otherwise be perpetrated by dominating individuals (see, for example, Boehm 1992).

Indeed, the harm-imminence – withdrawal-interval is a critical variable regulating the game dynamics of alliances, as is the link between injury, post-combat formidability, and post-combat

bargaining position. The problem begins with the fact that when fighting, incurring a cost for the coalition lowers the formidability the individual can deploy to bargain for its share of the winnings. Consider elephant seals, or other species whose combat involves relatively slow attrition. Inflicting a cost in such species is closely associated with incurring damage, and at a relatively slow rate. When damage is incurred, the formidability of the animal is lowered. If the attacker is part of a dyadic alliance, the attacker's future ability to enforce its share of the winnings against its ally depends on its subsequent formidability. If attacking will decrease the attacker's formidability to a point where it cannot enforce its share of the winnings, then the individual should refrain from attacking in the first place, or withdraw when the rate of damage predicts the imminence of formidability decline. Such attacks would constitute one-shot games. In short, when the harm imminence-withdrawal interval is large, then the alliance dynamics unravel cooperation. In contrast, when the harm imminence-withdrawal interval is too short, then there is not time enough for the attacker to respond by withdrawing from the fight. The individual should withdraw at the point when the attacker ceases to share risks equally with its partner, and begins to receive unequal damage to its formidability. So, at the point of attack, each attacker faces a veil of ignorance that averages payoffs across the coalition members. As long as the net payoff is positive, the coalition should not unravel. It is also easy to see how inclusive fitness effects among related individuals could cushion and stabilize these dynamics.

Forming and Maintaining Alliances

Alliances pose a series of adaptive problems that selected for cognitive and motivational specializations for their solution: For example, individuals must be able to form and maintain alliances, recruit allies, evaluate and select allies, motivate allies to support them, influence alliances to take those actions which are beneficial to the individual, and map the alliance structure of their social world. Cognitively, individuals in a world with coalitions must be equipped with programs that detect alliances (evidence supports the view that humans have such an alliance detector; Kurzban, Tooby, and Cosmides 2000); these neurocomputational programs must be able to assign formidability estimates to alliances as well as to individuals (using a formidability-integrating function of some kind);

they must be able to integrate their estimate of an individual's formidability with their estimate of the individual's coalition-derived formidability (Ermer 2008).

Assigning formidabilities to coalitions underlies a range of human social realities: group rank, the redirection of resources from less formidable to more formidable groups, displacement from territories, and the entire panhuman suite of wars, intergroup rivalries and conflicts, group-based privileges, and power differentials. This superstructure requires a psychology of group formidability, including alliance formidability detection.

Formidability can also be altered behaviorally: It is typically too costly for humans to carry their weapons with them on all occasions, and so formidability asymmetries are magnified during ambushes and surprise attacks. Where conflicts are likely or endemic (e.g., chimpanzee or ancestral hunter-gatherer zero sum territory competition) such a payoff structure favors "first strikes," raiding, and the offensive initiation of conflicts. In short, the human entry into the cognitive niche (which carries with it the human ability to act in coalitions over long distances and extended time periods in a coordinated way) intensifies the payoffs to initiating collective aggression (Tooby and DeVore 1987; Wilson and Wrangham 2003).

The two biggest obstacles to the evolution of alliances and coalitions are the problem of free-riding (Tooby and Cosmides 1988; Price, Cosmides, and Tooby 2002), and the problem of coordination (Tooby, Cosmides, and Price 2006). A coalition can be defined as a group of individuals that coordinate their actions to achieve common goals and share the resultant benefits. From this definition, it is apparent that coalitions depend on (1) adaptations for solving problems of coordination, and (2) adaptations for solving problems arising from benefit allocation not being conditioned on contributory behavior (i.e., free-riding). If individuals do not coordinate their behavior toward some common goal or benefit, then there is no coalition. If free-riders outcompete cooperators, then coalitions cannot stably evolve.

Anti-free rider adaptations: How is the problem of free-riding solved? From a cognitive perspective, a coalition is an *n*-party exchange, in which each participant is entitled to receive benefits from the action of co-participants conditional on the participant's supplying contributions to the exchange (Tooby, Cosmides, and Price 2006). Free riders are individuals who take disproportionate benefits from coalitional projects compared to other participants without paying

proportionate costs. Evidence supports the view that humans have a cognitive specialization for detecting free riders (Delton, Cosmides, and Tooby, forthcoming).

Secondly, we and our colleagues have found evidence that humans evolved a motivational program that generates punitive sentiment toward free-riders in coalitions (Price, Cosmides, and Tooby 2002; Tooby, Cosmides, and Price 2006). Contributors to collective actions are motivated to have a negative WTR toward free riders. Another line of defense is a basic anti-exploitation orientation, including a motivational circuit to downregulate contribution as a function of how much free-riding is going on (Tooby, Cosmides, and Price 2006).

Adaptations for coordination: Coordination poses even greater difficulties for the formation and operation of alliances and coalitions (for discussion, see Tooby, Cosmides, and Price 2006): An individual is a unitary information processor, and can be expected to support itself. Even a dyadic alliance is not a unitary information processor, and coalitions only exist and function when their members coordinate their actions to some productive extent. Individuals have different interests, locations, loyalties, values, social relationships, formidabilities, and information. Moreover, humans are presented with an uncountably large number of alternative cooperative projects (n-party exchanges), each with a different matrix of payoffs for the participants. Indeed, coalitions are by their nature coordination games in which payoffs are a function of the number of participants who contribute to the same project productively, and their contingent interrelationships.

To coordinate on a single project, potential coalition members must mutually collapse the space of possibilities down in their minds to the same one (or a few that can be carried out simultaneously). We think that adaptations for coalitional coordination include programs implementing a theory of group mind; programs implementing a theory of interests; programs implementing a theory of human nature; programs for leadership and followership; the outrage system; theory of mind; coregistration programs for solving common knowledge problems; language; and an underlying species-typical system of situation representation which frames issues in similar ways for different individuals.

Common knowledge: One kind of difficulty of coordination can be clarified by the concept of common knowledge, which the philosopher David Lewis introduced to describe knowledge of the following

kind (Lewis 1969; see also Nozick 1963; Aumann 1976; Chwe 2003). For a group of agents, common knowledge exists of proposition x when all the agents in the group know x; they all know that they all know x, they all know that they all know that they know x, a recursion that continues ad infinitum. Cognitively speaking, this is an implausible set of representations, not least because it requires infinite storage and infinite time. Obviously, it needs to be recast in computationally realizable, adaptationist terms. But why is this important? An *n*-party exchange is an example of intercontingent behavior (Tooby, Cosmides, and Price 2006): What I do is contingent on what you will do, while what you do is simultaneously contingent on what I will do. That makes my behavior contingent on my knowing your knowledge of my behavior, which is itself contingent on your knowing my knowledge of your representation of my behavior (and so on) – mirrors reflecting each other endlessly. Mutually coordinated behavior among two or more actors cannot be achieved without some analog to the set-theoretic relationships analyzed by logicians as common knowledge. If cooperation and coalitions require common knowledge, and common knowledge requires infinite cognitive resources, how can coalitions or cooperation exist?

The first thing to recognize is that the standard common knowledge formulation rests on a flawed folk concept of "knowledge," a flawed assumption of economic rather than ecological rationality, and the flawed blank slate view of the mind. In contrast, a large number of neurocomputational adaptations involve specialized systems of valuation, and depend on internal regulatory variables that do not correspond to beliefs or communicable representations ("knowledge"), but rather to value-related settings in motivational, emotional, and other decision-making structures (Tooby, Cosmides, Sell, et al. 2008; Tooby and Cosmides 2008; for a detailed example, see Lieberman, Tooby, and Cosmides 2007). As discussed, these include such computational elements as WTRs, formidability indexes, kinship indexes, sexual value indexes, and so on. For successful behavioral coordination to occur, agents must (1) converge on a common representation of a situation, (2) converge on similar (or compatible) regulatory variable weights relevant to the situation, (3) recognize the convergence, and (4) converge implicitly or explicitly on a cooperative response. The requirement that the architectures "recognize the convergence" does not mean they represent common knowledge in the formal sense. It only means that the architecture has one or

more regulatory variables that are increased when there are cues of convergence, and that when a threshold is passed (set by a selective history of payoffs that exceed uncertainty-caused costs), implements the cooperative behavior. There need be no explicit and deliberative representation of others' knowledge states at all. Such a design acts as if it satisfied the common knowledge criterion for game play, without actually having or representing common knowledge.

Sharing the same evolved architecture provides a partial foundation for resolving the game theoretic problem of common knowledge with finite cognitive resources. For many basic aspects of jointly experienced situations, humans, by virtue of being members of the same species, already share a common architecture containing a rich and detailed set of adaptations for interpreting and responding to the world largely in the same way. The space of logical possibilities is radically pared down to manageable proportions by possession of the same situation-interpretive machinery. If humans have adaptations for different families of evolutionarily recurrent games and strategies of play, then they can count on others having the same adaptations – what might be called architectural coordination. (Because the coordination required between psychological architectures diverges from the normal meaning of knowledge, we prefer to call this necessary parallelism *mental coordination* rather than *common knowledge*.)

The more similar the information states of the two architectures, the more likely they will be to arrive at the same situation representation. Hence, the more similar the experiences of two architectures, the more coordinated they will be. We will call the process in which two or more individuals experience parallel inputs that bring about mental coordination *coregistration*. Spending time together, joint attention, being together at critical events – all obviously increase the frequency of coregistration and hence mental coordination. That is why these factors, along with the ease of mind-reading, spontaneous rapport, being *simpatico*, etc. are important facilitators of friendships and other alliances, as well as of leadership-followership relations. Indeed, the payoffs to coordination also plausibly selected for complementary adaptations that produce leadership-followership roles (Tooby, Cosmides, and Price 2006; Tooby and Cosmides, 1979). Coregistered events can also play the role of coordinating coalitions ("outrages" – coregistration of outgroup members harming ingroup members; Tooby, Cosmides, and Price 2006).

Emotions and the psychophysics of mental coordination: For cooperative action to be taken, evolved procedures must exist for inducing

or recognizing sufficient coordination in situation representation (e.g., others represent a joint threat) and regulatory variables (e.g., our formidability indices are too low to resist them). It is worth noting that specific emotions are evolved systems of internal coordination activated in response to evolutionarily recurrent situations such as danger, contamination, conflict, or pleasure (Tooby and Cosmides 2008). Their activation signals that the individual has assigned the particular situation encountered to one of a finite set of interpretations recognizable to the entire species. Emotions also organize motivational variables in predictable ways. Because of this, they are ideally suited, when signaled through facial and postural expressions of emotion, to show the individual's situation representation, and associated regulatory variable recalibrations. That is, coregistration of emotional broadcasts provides one solution to mental coordination, and its role in coordination may offer a selectionist explanation for the puzzle of why expressions of emotion are nearly automatic and quasi-involuntary.

More generally, there seems to be a psychophysics of mutual coordination and coregistration, involving (for example) joint attention and mutual gaze, especially timed when salient new information could be expected to activate emotional or evaluative responses in one's companions. The benefits of coregistration and mental coordination can explain (at least in part) an appetite for co-experiencing (watching events is more pleasurable with friends and allies), the motivation to share news with others, for emotional contagion, for gravitation in groups toward common evaluations, for aversion to dissonance in groups, for conformity, for mutual arousal to action as with mobs (payoffs shift when others are coordinated with you), and so on. The weightings occurring when information is coregistered should be more intense, because mentally coordinated weightings can be acted on with fewer costs. Issues of coordination provide a reason why coalitions should form around denser social networks whose connectivity provides faster coordination among its individual constituents. Fractal fissures in coalitional structure around which factions form should similarly track network structure.

Group interests: The problem of coordination includes but is not limited to common knowledge problems. Even if all parties had perfect mutual knowledge (which they do not), each would still be confronted with the unlimited set of alternative n-party exchanges, and the fact different payoffs and different characteristics among potential participants will typically lead each to favor different projects and

coalitional boundaries. Negotiating common projects, maintaining allies, and choosing courses of action in the context of coalitions all require the ability to predict and understand others' values, because one individual's actions affect others' interests or welfare.

Groups as agents: Another critical coalitional adaptation was the widening of the concept of agent in evolved procedures so that representations that formerly could have referred solely to human individuals become able to refer also to coalitions, alliances, communities, and other collectives (Tooby, Cosmides, and Price 2006). That is, groups can be mentally represented to be agents (a useful delusion), and so to be things to which we can attribute mental states as if they had a single mind. This delusion is a useful one, because groups sometimes do arrive at mental coordination making them similar to an agent. Common intentional states, joint action, mental coordination – and cues that increase the probability of these – should increase the perception of groups as entities – what social psychologists call *entitativity* (Ip, Chiu, and Wan 2006).

The ability to represent groups as agents allows us to construe groups as having intentions, attitudes, emotions, knowledge, and so on. Groups can have status, formidability, rank, stigma, and dominance relations, not to mention alliances, friendships and enmities. The group-as-agent construal allows individuals to represent themselves in exchange relationships with groups. Being able to represent a group as an agent allows us to apply the intuitive theory of interests specialization to groups – that is, it allows humans to think of groups as having interests, and therefore to approve or disapprove of an individual's actions. This last step is one of several keys to understanding morality. Not only can groups "have" emotions, but equally, they become interpretable as objects of our emotions and motivational programs, such as anger, gratitude, guilt, shame, welfare-trade-off representations, formidability, kin-oriented representations, and so on. That is, the whole apparatus that evolved to deal with individuals was modified so that it could be extended to groups.

Alliances, coalitions, and amplification coalitions. We define a *coalition* as a group of individuals that coordinate their actions to achieve a common goal. A coalition constitutes an *n*-party exchange, with the compliance of its participants dependent on the compliance of the others. The common goal could be anything, and therefore the coalition could be transient. However, coordination once achieved is intrinsically valuable, because it can be turned to many ends, and realize gains of many types. So coalitions among people who

repeatedly interact tend to gravitate toward becoming amplification coalitions (Tooby, Cosmides, and Price 2006). That is, an *amplification coalition* is defined as an *n*-party exchange system whose function 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 with the other members. The underlying principle is Dumas' one for all, and all for one. We use *alliance* to mean a a dyadic or small-scale amplification coalition (primarily formed out of dyadic links), whose major function is prevailing in conflicts of interest against other individuals or coalitions.

The characteristics of a given kind of social relationship (e.g., mateship, friendship, or kinship) may involve the operation of multiple adaptations reflecting distinct selection pressures. Social relationships that are coalitions are not reduced to being only coalitions. For example, we hypothesize that friendship circuits have a strong alliance/amplification dimension, sensitive to the registration of mutual support when either party is challenged by third parties. (The closeness of a friendship can be operationalized by the intensity with which one person is favored over another when there is a conflict of interest between them.) But as engagement relationships they also have strong elements of fitness interdependence, which reinforces their stability as alliances (Tooby and Cosmides 1996). And there is also the expectation that each friend's welfare will be more greatly realized in daily events by exchange, by cost-effectively combining welfare trade-offs (such as risk-pooling), and by cooperative labor.

Moreover, public signals of support (or their 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 rank, status, or formidability-justified entitlement. Because everything can be taken from a powerless individual or group, humans (especially men) have evolved specializations that motivate forming or affiliating with groups, that motivate affiliating with a coalition over no coalition, and that motivate affiliating with higher status coalitions (that will accept them) over lower status coalitions. That is, we have specializations for coalitional identity formation, and these operate even when the coalitional activities they result in have no obvious function (see, for example, Tiger 1969 on the case of fraternal organizations in developed societies; and Rofe 1984, on anxiety and the need for affiliation). Ancestral wars were intercommunity conflicts, but this system of dispute and alliance extended all the way down to the

dyadic level. Our coalitional adaptations should guide us to participate in coalitions at all fractal and nested levels, so that at whatever scale a dispute occurs, one has allies to press one's case (see Sahlins 1961; Boehm 1992 on segmentary social organization).

The dynamics of unification and fractionation: Unless there are large benefits that can only be obtained by large-scale coalitions, the interests of smaller scale factions will undermine the cohesion or preclude the existence of larger, encompassing coalitional levels. Conversely, to form a large-scale coalition, individuals and lower level cells must surrender their agendas to the labor contributions required by the larger scale project. Negotiating phase changes to different scales of coalitional cooperation in a fitness promoting way has been a chronic adaptive problem for humans, and we appear to have circuits specialized for this function. Both strategic *factionalization* (fragmentation of a larger coalition) and strategic unification and inclusion appear to happen in response to cues of (1) the payoffs existing at different scales, and (2) cues of mental coordination, such as coregistration of collective events. The underlying theme for increasing unification is the evocation of cues of fitness interdependence (Tooby and Cosmides 1996). Coalitional identity and affiliation can shift upwards, with existing coalitional identities being shed like clothes. Computationally, this involves recalibrating WTRs upwards toward those who were previously outgroup members, as well as toward higher-level coalitions. Of course, the most reliable facilitator for higher-level coalitions is external conflict with a large competing coalition – something repeatedly found in the historical record. The prospect of large gains or huge losses through displacement, expropriation, subordination, or extermination is one of the few reliable signals that the formation of large-scale coalitions would be worthwhile.

Alliance mapping and coalitional evaluation: Social life is riddled with implicit and explicit coalitions across a range of fractal scales, and choosing courses of action requires anticipating which responses and latent coalitions might materialize in response to various actions. To accomplish this, we think there are a number of evolved inferential elements for alliance mapping (e.g., between two people, patterns of assistance and prosociality, such as sharing, close kinship, maintained proximity and approach, coresidence, positive affect, mirrored affect, empathy, etc. imply alliance and stable high WTRs; in contrast, zero-sum conflict, anger, disgust, contempt, avoidance, resource confiscation, unmirrored affect, counterempathy, unwillingness to

share, nonassistance and goal-blocking, exploitation and aggression obviously imply enmity and negative WTRs). These are combined with other evolved inferential circuits, such as generalization of social relationships to allies: e.g., positive action by one person towards a target recruits positive recalibration in the target's allies; and especially negative action towards a target recruits anger and punitive sentiment in the target's allies toward the malefactor. Such inferential and motivational elements (with proper scope limitations) can be combined recursively to deduce a social map and associated mental contents (e.g., the friend of my enemy is my enemy; the enemy of my enemy is my ally; the enemy of my friend is my enemy, etc.).

These inferential elements provide input to the alliance detector, which we have begun to map (Kurzban et al. 2001). Its ideal output should be an alliance map (perhaps resembling a social network map) that not only represents individuals in terms of the ongoing coalitions they belong to, but also: (1) the differential strength or clustering of their actual and potential alliances to others in the social system, for each likely issue; (2) factors that predict alliance value and disposition: e.g., their trustworthiness, duration of their participation in the coalition, history or summary of their level of contribution (their welfare trade-off propensity to the group account compared to self), observed WTRs, individual characteristics (like kinship) that predict interests, individual and alliance formidability, etc. The discriminative alliance system should use such characteristics to assess others' value as coalitional members or enemies, and to regulate motivations for recruitment, rejection, price setting for inclusion (e.g., subordination or required level of contribution). Coalition membership should be associated with indices that track how valuable the individual's membership is – with membership being a fuzzy set relationship. They should also track an estimated WTR propensity from the individual to the coalition. Acts of sacrifice for a coalition or individual are highly informative, as are acts of contribution and allegiance cues. Such acts should trigger categorization as coalition members, as well as changes in the index tracking how strong membership is – how "good" a member a person is.

Status. As discussed, formidability – the ability to inflict costs – is only one kind of bargaining tool. The second family of bargaining tools is the ability to withhold or confer benefits. Formidability is tracked by formidability indices, and the ability to confer or withhold benefits is tracked by implicitly registered conferral indices. Among humans, evidence indicates that both appear to be registered, both

appear to determine who prevails in conflicts of interest, and both regulate the deployment of anger as an implicit bargaining system (Sell, Tooby, and Cosmides 2009). During human evolution the ability to cooperate and to produce alienable benefits greatly expanded, so the force-based logic of animal conflict has been greatly elaborated to include a co-equal cooperative dimension. Representations of formidability and the ability to confer or withhold benefits are the two direct components of individual status (Tooby and Cosmides 1996). Secondary components of formidability and benefit control include support in each of these negotiative modalities provided by others as individuals or coalitions. A third factor is the relative support one's supporters have, the support their supporters have in turn, and so on, as devalued by the probabilistic decay of support along network links – something akin to Google's page rank algorithm. A fourth factor is the ability to mentally coordinate others in the community (leadership). A fifth factor is common knowledge or mental coordination (or discoordination) of how these representations and weightings are ecologically distributed in the relevant population of social actors. That is, do I register that everyone else registers this person as high status? A sixth factor is moral status, to be discussed later. While we hypothesize that each of these (formidability, conferral, support, leadership, coregistration of these variables, etc.) has its own proprietary representations or regulatory variables, they all need to be integrated into a single summary variable, status. A primary function of assigning a status index to an agent is to be able to assign weight to the bargaining power (and related properties) of the agent – whether that agent is an individual or coalition. That is, the function of a status index is to predict the fitness consequences that arise by engaging in conflict, cooperation, affiliation, proximity, welfare modifications, and other interactions with the agent. The status index is the summary function that evolved to track status. This index is based on an evolved status estimating system that takes the subcomponents (formidability indices, conferral indices, etc.) and integrates them into a decision-making data-structure. In general, the higher the status of the agent, the greater the WTR one expresses toward the agent.

The alliance detection system needs to assign status to individuals, clusters (potential coalitions), and mentally coordinated (actualized) coalitions, in order to usefully navigate the social world. The mental coordination of representations and status evaluations is crucial, because the reigning social reality is governed by how the population

represents status. So, humans have motivational programs whose objects are status representations in the minds of others, and their distribution and coordination in the local population. Status is zero sum, at least among non-allies, generating a status rival matrix in the population. People like status increases for themselves and their friends and allies, and status reductions in their status rivals. They respond to the prospect of alternative courses of action in part by their status consequences. They engage in status operations, designed to increase status of themselves or their allies, or reduce status in others.

Coregistration cues – that is, the mutuality of social observation –should be an important regulator of decisions. A fight that no one else observes may only change the formidability indexes in the two participants, so the winner pays a given cost to accrue greater status in the mind of a single individual. If the entire community coregistered the fight, then the same cost would purchase a recalibration of his status in the minds of everyone – and a mental coordination of his enhanced status. So the motivational intensities of the status recalibrational emotions – shame and pride – will be proportionately greater to the extent that they lead to mental coordination (common knowledge) of the changed status among a larger set of individuals. Cues of coregistration are an important regulator of status operations (see, for example, Ermer, Cosmides, and Tooby 2008). High coregistration is a lubricant of recalibration, while low coregistration produces friction in social recalibration.

Musical chairs – competitive behavior under scarcity: In addition to a territory displacement game, a power-based bargaining game, and a demographic attrition game, we think that humans ancestrally played what might be called the musical chairs game: there is one less chair than there are players, when the music stops players rush to find a chair, and the one left standing is eliminated. The abundance of habitats varied greatly and unpredictably over time. The alliance maps of local populations involve clusterings, and areas where network links are sparser. Persons who might be tolerated or welcomed as part of the community during abundance might shift to imposing fitness costs by merely existing during times of scarcity. If the social network were equally dense everywhere, then any attempted exclusion would be difficult and involve costly conflict, recruiting equal numbers of allies on both sides. To the extent that there is mental coordination on network fissures (signaled, perhaps, by small daily acts of humiliation), then a spontaneous coalition of the well

networked with high potential formidability could actualize itself to exclude marginal segments of the population. Hunger, for example, should provoke shifts in tolerance, acceptance, and themes of inclusion – not to mention, drops in oxytocin. The emergence of social dominance in complex societies that Sidanius and Pratto (2001) document may be rooted not only in adaptations for coalitional and intercommunity competition and status interactions, but also ancestral musical chairs interactions in more egalitarian societies.

Coregistration and the game of chicken: Imagine that two forager communities are locked, for example, in an ancestral, chimpanzee-like demographic war of attrition (i.e., where larger groups eventually replace groups that are slowly whittled away by raids). A male resident of the community benefits when members of the other community are killed. It weakens them as a threat, moves the territorial line so that more food is available to the male and his children, and so on (see Wilson and Wrangham 2003, on chimpanzees). Why wouldn't a sane male designed by selection to promote his fitness share in the benefits of others' actions, but shirk himself? (That is, why would not selection shape male psychology to avoid participation in offensive war?). His participation would only marginally increase these benefits, but he gets the full benefit of the costs he avoids (Olson 1965). The common participation of adult males in such situations has led some to argue that group selection is the driving force in human warfare (Bowles 2006). The risk contract of war (Tooby and Cosmides 1988) is one model of selection pressures involved in war, and models of punitive sentiment as an anti-free-riding adaptation may help to answer this question (Price, Cosmides, and Tooby 2002). However, there are other selection pressures which we think also operated. We think these act in a complementary fashion to reinforce selection for an evolutionarily stable coalitional psychology that accrues benefits from war.

Ethnographically, lethal conflict inside groups is treated very differently than homicide in intergroup conflict – within group homicide is typically punished, while attacks on enemies are typically socially valued. Death or maiming of a community member triggers factional within group conflict, the activation of allies, and potentially serious consequences. Still, conflicts of interest do occur inside communities, and may even erupt into violence. Further, humans can deploy lethal violence quickly, through the use of weapons – which are brandished and sometimes used in intracommunity conflicts (Chagnon 1983; Lee and DeVore 1976). Moreover, anger as a

bargaining emotion (as well as male combat identity) causes face offs. Such encounters can be considered a game of chicken: One or both may be injured or killed unless one defers to the other. Yet, individuals are sanctioned for killing or maiming ingroup members – an event which would otherwise cause fitness-enhancing coregistration of the formidability of the winner.

Given these facts, one strategy that might stabilize contributions to offensive war arises from the fact that killing or defeating outgroup enemies – in contrast to ingroup members – is not sanctioned, but is seen as laudable. If a group of males conduct a raid or a battle, coregistration of their mutual exploits allows the mutual assessment of their relative formidabilities within the community and between the rivals – including characteristics like courage that are hard to assess in restrained ritual combat. These then set precedents about who should defer when conflicts with the potential for escalation occur among them. The individual who is mentally coordinated as being more formidable will not defer in games of chicken, because the expectation has been set that the other will defer. This is one pathway through which coalitional contributions are good for the individual – at least for the more formidable. Research, for example, supports the prediction that stronger males are more pro-war (Sell, Tooby, and Cosmides, 2009). They have an opportunity to display (and at lower risk than weaker and less skilled individuals), from which they will derive status benefits. To stay home is not only undercontribution (free-riding), and a failure to exploit an opportunity for status enhancement ("glory," "honor"), but could potentially be interpreted as weakness and cowardice by other males in the community. Many cultural practices that seem instrumentally bizarre (such as counting coup or trophy-taking; Turney-High 1949) make sense as displays designed to coregister one's formidability with ingroup members. This dynamic is why, in so many cultures, warriorship is constitutive of masculine identity. "Glory" and "honor" – intuitive concepts that correspond to the coregistration of formidability – have been major motivations for participating in war across the historical and ethnographic record.

Morality, Valuation, and the Ability to Act in Concert

We have argued that humans have a far more elaborated evolved psychology for forming, participating in, and dealing with coalitions than other species do. Although morality seems superficially

unrelated to coalitions, we hypothesize that the evolution of adaptations for coalitions was a key trigger for the evolution of adaptations for morality (Tooby and Cosmides, 1979). That is, the evolution and elaboration of morality was midwifed by the capacity for the rapid recruitment of individuals into a coalition around a common interest that could be punitively enforced. More fully, we think that our moral psychology evolved (in part) as a natural extension of the adaptations underlying our coalitional psychology, as well as adaptations for social assortment and exclusion, in interaction with a number of other elements. As we will detail, these other elements include (1) pre-existing adaptations that evolved to solve other adaptive problems (e.g., language; negotiation); (2) a novel set of games (i.e., structured social interactions with payoffs) that were unleashed by the evolutionary expansion of coalitional and communicative adaptations; (3) novel features and adaptive problems inherent in the resulting social ecologies; and (4) adaptations that specifically emerged for successfully navigating the family of "moral" games that were endemic to this new coalitionally infused social world. We think that sketching out how these disparate elements interacted during our evolution can illuminate what the kind of thing morality is; how our species evolved a specialized moral psychology; and why, although our moral psychology is partly an outgrowth of our coalitional psychology, it is nonetheless distinct.

Negotiation and situation evaluation – the primary elements: Several kinds of pre-existing, premoral adaptations naturally interacted to produce first- order moral games. First, humans like many other animal species have adaptations for situation evaluation – values – that allow them to plan and choose more over less fitness- enhancing courses of action (Tooby, Cosmides, and Barrett 2005). Second, there exist suites of adaptations in humans that are designed to negotiate with others over the conduct of both self and others, based on valuations, alternatives, power, formidability, and status. As discussed, anger (for example) is one evolved program that implicitly organizes human bargaining, orchestrating the infliction or costs or the withdrawal of benefits in the service of prevailing (e.g., Sell, Tooby, and Cosmides, 2009). Negotiation occurs when we make our behavior conditional on others' conduct (through threatening to harm them, or to reduce or withhold benefit delivery); and vice versa. So, third, we are designed to attempt to influence others to act in conformity with our values. Fourth, others are simultaneously designed to incentivize us to act in conformity with their values.

Hence, first- order "moral games" are constituted by these complementary tugs of war, in which each agent negotiates to license the best obtainable course of action for the self, and each agent negotiates to obtain the most self-beneficial modification of the behavior of the other. Although these selection pressures operate to some extent on other species, the explosion in human instrumental behavior, and the human ability to communicate propositions with great precision vastly expanded the scope of social negotiation (far beyond messages such as "go away," or "mine" that are characteristics of other species). Whether one chooses to categorize these games of mutual influence as involving morality *per se*, it will subsequently become clear how they are foundational for phenomena that are widely categorized as moral.

The fact that first-order games of influence – even in a dyad – are treated by the mind as moral can be shown by considering the typical relationship of a parent and a young child, where the power asymmetry is large. The parent unproblematically uses the terms "right" and "wrong" to differentiate the parent's preferred courses of action for the child from those she dislikes (e.g., put away your toys when you are done; don't throw balls in the living room). Conduct in this case is moralized for the child, but not for the adult, since the child is too powerless to threaten the adult. Negotiating power is one ingredient that contributes to the formation of the moral domain. Exchange (or reciprocity), with its associated concept of "cheater" is an example of a dyadic first- order game in which the two participants have more equal power, and exercise it to modify each other's behavior advantageously.

The risk of others' coordinated action produces some components of the moral sense: The fifth element involved in the evolution of a distinctively moral psychology is an adaptive problem introduced by coalitions into the social ecology: From the perspective of any individual, there is a potentially dangerous power asymmetry. There are many others, and only one self, and others may join to form a powerful coalition (momentary or permanent) against any individual. This danger is relaxed to the extent an individual is powerful (such as a tyrant), and exacerbated to the extent an individual is powerless. If you take an action others strongly disvalue, they may combine to punish (harm) you, to your severe detriment. If you propose that others behave in ways that they strongly disvalue, they may combine to act against you. Because others' punishments or rewards are conditional on an individual's conduct, for any course of conduct being considered,

the individual needs to add to the direct payoffs (e.g., the benefits of obtaining money from the till) the contingencies of reward and punishment that will be triggered in others (e.g., retaliation for theft). In short, the existence of others, together with their ability to respond, selected for adaptations that were designed to implicitly represent the values of others, and that weight others' values cost-effectively in the individual's own decision-making process. The values of even a single other may need to be taken into account (as in dyadic exchange), but the ability of others to rapidly form coalitions greatly multiplied their power and therefore the intensity of selection for adaptations that spontaneously weighted others' values.

Adaptations that register others' values and weight them (according to predictors of the consequences) constitute one key component of what Darwin called the "moral sense or conscience" (Darwin 1871; see also Hume 1751; Hutcheson 1728; Alexander 1979) and what Freud called the superego (Freud 1923). Indeed, this system should be designed to modify the mind's native valuations by adopting others' values implicitly as one's own (to a calibrated extent). The degree of this internalization of others' values should depend on the registration of how often one is monitored, how uncertain the identification of conditions of privacy are (i.e., what the information ecology is like), how great the penalties for detected deviance are, and how great the potential power imbalance is.

The opportunity to recalibrate others' choices produced additional, complementary components in the moral sense: Ancestrally, each individual faced the risk that one or more individuals would punish her for acting in defiance of their wishes. This selected for adaptations often characterized as "conscience" or the moral sense. The sixth element is simply the reciprocal of this: Each individual has the complementary potential to join with others to enforce their values on one or more others. We expect the human mind evolved adaptations for exploiting this social opportunity – that is, adaptations for (1) leveraging one's bargaining position by recruiting others into a coalition (however transient) around common values; and (2) enforcing its values by downregulating benefits or inflicting costs on those who deviate from these values. The fact that the anger program evolved to negotiate conflicts by inflicting costs or downregulating benefits explains why anger is evoked in the negotiations characteristic of moral games. The adaptations underlying the moral sense are designed not only for cost-effectively internalizing others values, but also for causing others to internalize the individual's values. One subsystem

invites conformity to others' values, while the other unleashes morally censorious judgments, punitive sentiments or outrage designed to intimidate others into adopting one's values as their own.

Second order moral games: First- order moral games of individual mutual influence are transformed into second order moral games by the addition of coalitions. In second order games, coalitions are formed around enforcing the values that the coalition members commonly hold.

The moral domain is not a content-domain, and potentially encompasses an unlimited number of moral projects: The seventh element is a feature of the informational ecology faced by our post-coalitional ancestors. That is, agents playing second order moral games confronted a vast superset of alternative, potentially coalition-enforced values – far more than could possibly be actualized, especially given that many values and sets of values are mutually incompatible. For a morality to be actualized, this space of possibilities must be collapsed down to one. Indeed, one can get a sense of how large the set of potential moralities is by considering the immense cross-cultural range of real, documented moral projects and issues. The content that individual and local moralities are endowed with encompasses not only cross-culturally recurrent themes (e.g., don't murder an ingroup member) but extraordinarily heterogeneous and often contradictory contents (from Aztec ritual cannibalism and the National Socialist project to the psychedelic movement, Puritanism, sexual liberation, not revealing magicians' secrets, shocking the bourgeoisie, the Jainist prohibition on killing insects, and the restoration of the Caliphate). One answer to the question of why moral stances show such endless diversity is that our evaluative adaptations are designed to accept open input: All possible situations or outcomes must be able to be evaluated, in order for choice to operate with respect to encountered situations. If moral responses are derived in part from evaluative responses, and evaluation is open, it follows that morality is not a special content domain (like allocation or justice), but a posture with respect to any content that can be evaluated. Moreover, the surface contents of moralities often function merely as coalitional coordinative signals rather than as doctrines selected for their intrinsic attractiveness (e.g., the doctrine of predestination). Often moral contents are selected in order to signal the emergence of a new coalition, or to morally legitimize attacks on rivals based on pretexts arising from the surface properties of the rivals' moralities. Indeed, people often support moral projects not because they hold any intrinsic attraction but because of their

downstream effects on rivals – for example, reducing the their status or weakening their social power.

Payoff distributions, coalitional maneuvering, and situation evaluation: The eighth element is the adaptive problem agents face because different moralities make different social group members winners and losers. That is, alternative values potentially distribute different payoffs among the participants (e.g., tolerance of infidelity favors attractive men at the expense of unattractive husbands and investment-deprived wives). The fact that different moralities privilege different individuals, combined with the fact that there are an unlimited number of possible alternative moralities, creates moral games concerning which moralities should reign in the social community. That is, there will be conflict between individuals and coalitions over which values out of the potential superset should spread in the social group. Second order games of morality involve individuals and emergent coalitions endlessly jockeying to advantageously actualize their values as the standard for punishment and reward, and to displace or preclude competing value projects that do not pay off as well for them. This approach explains why morality in a community is dynamic, and changes over time; why it is historically and culturally contingent; why it is contested and debated; why in the same time and place, whenever different sets of individuals are gathered – say, coalitions at different fractal scales – subtly or grossly different moralities are evoked within each group (e.g., men in their fraternities versus when they are with their mates). Which morality will be actualized depends on the specific set of coregistering individuals. Given the heterogeneous and fractal structure of groups in a population, moralities will be evoked with respect to the anticipated circle of players who will become aware of the deployment of a jointly defended value.

These games are further complexified by the fact that there are different roles in moral games, and the characteristics of different individuals will yield higher payoffs if they adopt some roles more than others. Roles include potential targets of moral attacks, who need to defend themselves; potential initiators of moral attacks; those who monitor; those who inflict punishment; those who withhold cooperation or deliver rewards; those who ostracize; and, of course, those who support and those who oppose the moral project. The distribution of individuals positioned to embody strategies advantageously, as well as the efficiencies of their combining forces, makes a difference to the dynamics of play. For example, high formidability

individuals are better positioned to be enforcers and punishers, while leaders will be better situated to bring about mental coordination on the necessity for enforcement and punishment. Low power individuals will be more inclined to transfer information about the moral deviance of their fitness suppressors. Different adaptations should be associated with each role, although, of course, all of the different adaptations should be present in all normal individuals.; After all, individuals may often play more than one role simultaneously, and sooner or later might end up playing each role).

Moral communication to potential allies and targets: The ninth element shaping our moral adaptations involves the requirements that moral games place on moral communication. For an agent to successfully recruit others in support of her preferred value project, the value must be communicated first to potential recruits (the proponents of the value) – and (if there is enough support) then to those whose behavior is to be modified. That is, if it is to have the desired effect on modifying behavior, then it must also be communicated to its targets (e.g., potential robbers; partisans of equality for a dominated group, potentially unfaithful wives). Morality is "public" within a certain circle of players, however small. The issue of joint awareness within a circle of players is one factor that makes the moral game different from just any individual or small-scale bargaining or wielding of social influence. Approval and disapproval are exploratory probes that allow potential proponents to assess how widespread support will be for the proposed value. They also warn violators that they are being categorized as transgressing. Moral communication also often involves deception. Leaders, for example, may play a moral game in which they acquire moral power by conflating themselves and their discretionary actions with widely supported value projects (e.g., wrapping themselves in the flag). To attack them is to be seen as opposed to the moral project they are emblematic of.

The ladder effect and the pull toward depersonalization: The tenth element is what might be called the ladder effect in second order moral games – that is, what is the effect of larger and larger audiences on the content of successful moral projects. If individuals' interests often conflict, then the expression of a value in its personal form – this allocation injures me – will have only limited and small-scale appeal as a value around which to recruit the support of others. Only personal allies will support the cause of the injured individual, and so individualized disputes will have moral resonance only in small groups. To gain wider support, the value an individual launches can be recast

in broader terms so that it simultaneously works in the interest of enough others to become the winning coalition-propelled value of its kind. As the moral representation moves from individual to individual and coalition to coalition, it will "evolve" – be progressively modified – by the set of players who engage it. Equally important, it will evolve (morph) because of how those who engage the value anticipate it will be received by a broader set of players.

To climb the ladder of increasingly wide support, a moral project cannot be seen as the instrument of simple parochial self-interest. This means that moral issues, as they encompass more players, increasingly take the form of rules, with "I support individual *i*" in an initial dispute evolving toward a rule, "for all *x*'s in condition *c* must do/must not do action *a*." As the value spreads more widely within a circle, and is perceived as applying to more individuals, the different players will take sides depending on how they represent the proposal as potentially affecting them and their family and local allies. The contagious moral result therefore often sums up to outcomes not wholly divorced from average population utility (as distorted by social power, such as male privilege or the divine right of kings). However, this does not mean the evolved function of morality is maximizing group utility.

Morality, common knowledge, and mental coordination: The eleventh element is the effect on moral games of mental coordination about (or common knowledge of) different candidate values. The payoffs to behavioral conformity to or deviation from a given value depends on (1) the distribution of supporters, neutrals, and opponents, and (2) the aggregate effects of how each person represents what positions everyone else will publicly take and behaviorally support. There is much less risk to defying a rule that only some people support. In contrast, a winning rule is one that is coregistered as being supported by everyone (or by a winning combination of power holders, at least). To the extent individuals can spread the representation to others that everyone is mentally coordinated on the proposition that some value is moral (supported), then this flips the incentives on dissent. Dissenters risk widespread withdrawal of support or community wide punishment, motivating them to conform. In the highest level of the moral game, what individuals and groups are perpetually jockeying to do is to actualize mental coordination about which values will reign within the moral community. The ideal outcome is to forge a value project that is beneficial to its proponents, which then wins enough support that everyone publicly endorses it, and that finally is coregistered by everyone as universally endorsed. For an individual, that is winning the jackpot in second order moral games.

The motivational bias toward moral realism: The twelfth element is moral *antirelativism* – the preference to spread and enforce the belief that morality is objective and has an intrinsic reality. This follows from the role that mental coordination plays in empowering moral values. The winning outcome in social negotiation is to get everyone to adopt your position as their own, so that they conform to it, effectively enforce it, and carry the costs of enforcement. Mental coordination is defeated to the extent that it is publicly recognized that there are differences of position. This is usually recast not as moral relativism, but as individual mistakes in perceiving what the moral position "really" or "truly" is. It is almost definitional of morality that people intuitively represent morality to be intrinsically good, and support its being seen as real and objective. We argue that this is an evolved circuit.

There are a set of games that humans have played so intensively over the course of our evolution (such as dyadic exchange and collective action) that we have reasoning and motivational adaptations that were specialized by natural selection for these particular games (Cosmides and Tooby 2005; Price, Cosmides, and Tooby 2002). These games have what might be thought of as their own proprietary moral concepts (e.g., cheater, free-rider) and moral sentiments (e.g., punitive sentiment toward free-riders). When situations fall into the domains of these evolved games, our moral stance tends to be organized by these dominating evolved interpretations and motivational agendas. Yet, the scope of possible moral contents is so large, we cannot have evolved responses to all of them, and so many of our moral responses must come from other sources. Obviously, explicit representations of self-interest play a large role in our attraction to, or resistance to candidate moralities, and help to fill this gap. There are many moral phenomena, however, which do not fit either pattern. For example, it is in each male's fitness interest for reproductively competing males to adopt a homosexual orientation. Yet across many cultures, homosexuality is intuitively viewed as immoral. One might similarly ask why people in so many societies are morally concerned with third party incest that has no impact on them. These and other cases can be explained by considering that selection would have favored a tendency to endow a modest moral realism to one's personal evaluative reactions to others' behaviors, when reframed as if it were a first- person experience. On this view,

what happens when the mind's evaluative systems implicitly assess a represented situation positively (e.g., if I were to experience that, I would like it)? When this reaction is not overruled by more powerful and specific features of our moral psychology, by conformity, or by self-interested strategic thinking (e.g., if this person gets away with theft, then that will injure me), then this first-person situation assessment migrates toward being positively moralized (or less negatively moralized). If the mind evaluates the situation as negative (if I were to experience it myself), then our moral psychology migrates toward viewing it as immoral. Thus, heterosexuals imagining themselves engaging in homosexual acts tend to have a disgust reaction, underpinning a negative moralization. Indeed, our data shows that subjects' moral objections to sibling incest by third parties track their preferences for their own personal sexual behavior toward siblings – as activated by the evolved anti-incest system (Lieberman, Tooby, and Cosmides 2003). While there is no logical relationship between the two, this moral circuit would be favored by selection because individuals would usually have benefited by having their preferences moralized. On balance, such an adaptation would make the local moral consensus more favorable to realizing the individual's preferences, even though sometimes such outputs are functionless or fitness reducing (e.g., opposition to homosexual behavior).

N-party exchange, hypocrisy, and private versus public behavior: The thirteenth element is that moral proposals in second order games are treated by our moral adaptations as a variant of n-party exchange or collective action. Just like for any other collective action, support from others for a moral proposal is purchased by the proponent's conformity with it – their contribution is their following the moral rule, and (to a lesser extent) their enforcing the moral rule. Evidence for this can be seen by the fact that our moral psychology includes an anti-hypocrisy circuit. This feature deflates both support for a proposal and the willingness to adhere to it when its proponents are discovered to be not following it themselves (Tooby, Cosmides, and Price 2006). There is no logical reason why others' adherence or abandonment makes a moral precept more or less worth following. It makes perfect sense, however, if morality is an n-party exchange, and our adherence to a costly moral rule is equivalent to the sucker's payoff when others have abandoned the rule. Because the negative reaction to hypocrisy closely parallels the negative reaction to free-riding, individuals prefer not to be seen as hypocrites. Obviously, there are benefits to evading the negative consequences of detected hypocrisy, as well as other advantages of evading others' responses to other kinds of moral defection undertaken to obtain gains. This dynamic leads to evolved programs for distinguishing one's own private behavior from public behavior, registering when acts are private, and modifying behavior in private to be more self-interested. As a consequence, moral games (among equally powerful players) favor individuals to be privately non-Kantian (do not act as you would if everyone were to do it) but publicly Kantian (publicly acting in conformity with those values that you are motivated to spread).

The fact that second order moral games are intuitively treated as n-party exchanges is one of the things that distinguishes morality from negotiation, politics, trade, aggression, or other ordinary paths to influencing others' behavior. This implicit framing explains why acting out of self-interest is intuitively contrasted with moral behavior (i.e., to fulfilling one's part in the implicit exchange). Moreover, it explains why, if other people feel no stake in a dispute, it is not seen a moral issue. In contrast, if people feel actions in a dispute would set a precedent for future actions by others, then the issue intuitively becomes moralized. Precedents implicitly define the terms of the n-party exchange that members coregister as being obligated to follow. These neurocognitive circuits provide the intuitive seeds for the common law practices of using precedent, and (when trying to escape precedent) distinguishing cases.

Partisanship and impartiality: Together, the combination of the ladder effect, morality as n-party exchange, the recognition of hypocrisy as cheating, the requirements of mental coordination for morality, and the moral realism bias explain the complex game dynamics revolving around representations of moral impartiality. The collective action of moral support for a value project is sustained or unraveled to the extent there are cheaters or free riders – people who benefited or would benefit from the moral rule applying to others, but depart from it when it costs them. People recognize the advantages individuals derive from benefits flowing to their kin, friends, and allies. This means that those who are not intrinsically drawn to being impartial moral actors when it injures them (i.e., everyone) are tempted to be cheaters. Partiality is (sometimes) restrained by the recognition that acting on this temptation leads to anti-free-riding punitive sentiments or the costs to the cheating individual of others' ceasing to adhere to the moral rule. Moreover, the proportion of a group's members that are the targeted beneficiaries of an act of partiality will be smaller the larger the group, while

the number of individuals with punitive sentiments will be proportionately larger. If the group size (moral community) becomes large enough, impartiality can become a publicly endorsed (if secretly unpracticed) ideal. In contrast, the perception that a value is being flexibly deployed to serve someone else's self-interest undermines recruitment of support to that value.

As moral projects climb the ladder to broader audiences (being recast and potentially applied to increasingly broad sets of individuals), any given individual will be bombarded with increasing numbers of candidate moral rules. Although these moral rules were not initially formulated with him or her in mind, the fortunes of life may make them applicable. Each individual can help to propel, or help to extinguish a rule or value, and should modulate their support based on their evaluative response. The evaluative response in each individual should be guided by which role (beneficiary or target) the individual anticipates as most likely to apply to him or her (and to those whose welfare matters to that individual). To the extent the individual is uncertain which role she might end up in, she should balance the two. This should lead to some tendency for rules to mutate toward greater "fairness" when large numbers of players are involved who are uncertain about their future role in the rule. If for most individuals, what one gives up if the rule is adopted (the probability that the rule applies to you as target, plus its cost if it does) is less than the benefit derived from having the rule to apply to everyone else, then the community will tend to move toward making the rule general and mentally coordinated. In principle, an individual could benefit by retaining the option to rob or commit murder without community opposition, but benefits even more giving up that option in order to have all others in the community prevented from doing so (Hobbes 1651). In this case, the individual will publicly support the rule, even if privately she may not always follow it. Thus, moral rules forbidding murder (of ingroup members), robbery (of ingroup members), and the like, will often become mentally coordinated, reigning moral rules within human communities, without having been favored by group selection, uncompensated self-restraint, or individual dedication to group welfare.

Moral cascades and moral warfare: Cognitive adaptations for moralizing disputes, attacking others' transgressions, and for launching contagious coalitional value projects are products of a selective history of offensive moral warfare. The ultimate prize in the moral game is one that causes a moral cascade that leads an advantageous value

project or attack to increase in frequency until it becomes accepted by the community as the reigning or equilibrium understanding: that is, by crafting and launching a representational bundle whose contagious runaway adoption by others amplifies the individual's preferences into a mentally coordinated valuation that recruits others and reaches high frequency in the local population. It recruits allies, and disadvantages opponents or rivals, for example by mobilizing punishment against rivals. While complete victories in moral warfare are rare, it is common for individuals and groups to better their position by playing second order moral games, both offensively and defensively. We expect motivational and cognitive adaptations in our moral psychology evolved to play these games well. To create a well-crafted project, the individual must, for example, have a good intuitive map of the local moral and social ecology. To be successful, the set of launched representations must be packaged to appeal to others (and to publicly disarm opposition). It must be apparently de-personalized, so that the issue does not begin or end just with the launcher's self-interest. To be successful in larger populations, it needs to attract supporters who have no particular interest in the launcher, the launcher's situation, or in the dispute, except as it sets a precedent that might be useful to them. Ideally, the project should be crafted so that others can see that is in their interests as well, and can foresee how it will apply. To reward the launcher, it must be self-interested, yet not appear self-interested. The constraints on a moral project are twofold: Finding a convergence of interest between launcher and potential allies; and finding a convergence of interest among a large and powerful enough set of potential allies that they successfully advance the project to the necessary level of social support. Moral offensives often focus on behavior of the target. Allies are more easily recruited if there are cues that make it easier to arrive at mental coordination on the wrongness of the behavior of the target (as well as the "facts"). For example, sins of omission are considered much less wrong than sins of commission, even though a utilitarian logician would defensibly equate them. Computationally, however, everyone is always omitting to take an infinite number of actions. With few exceptions, attaining a meeting of the minds that one omission out of this amorphous set is uniquely immoral is far harder to cognitively arrange than is the joint identification of a mutually objectionable act of commission. In contrast to sins of omission, sins of commission are finite, and knowably specific to the person or persons who commit them. Moral outrage, if it evolved to be useful in coordinating

joint action should show intensities that correspond to how easily a type of action can be used to mobilize joint assault. Hence, because of the lower payoffs, people should be far less outraged by omissions than commissions. Reciprocally, from the perpetrator's point of view, acts of omission are more masked from moral retaliation – so humans should feel far less guilty about them. Equally, outrage elicited from an individual should be stronger when the underlying value is known to be shared by others. For example, someone who shows very low WTRs toward others poses an exploitive threat to others, and intentional behavior is more revelatory of underlying WTR dispositions than are actions with unintended consequences. This predicts that unintended negative consequences will trigger less outrage than intended consequences. This prediction is also supported by evidence (see, for example, Sedlak 1979; Hauser 2006).

The game of fault and blame: One particularly important moral game is the attribution of responsibility. We have special terms for social causality: fault, blame, credit, and so on. These terms have two components, one explanatory and one evaluative: First, the explanatory attribution is the claim that the causal source of the event originates in the person or group identified as responsible, and causation is not traced to other agents beyond them. Second, the evaluative implication is that the agent's n-party social exchange account and status is decreased or increased as a result (if you were at fault, you owe something; and you are worth less as a social partner, either because of bad character – bad intentions – or incompetence). Third, analyses of blame and credit are strongly affected by the benefits or costs that the actor and those affected experience as a consequence of the act that "caused" the outcome. These are the elements that allow the actor's WTR toward others to be computed, a third evaluative implication (see below). Others are morally blameworthy if they exhibit unusually low WTRs toward others.

Because human agents are complexly computational, there are usually an uncountably large number of possible paths and interactions involving many persons that could have prevented a negative outcome, or could have led to a better positive outcome. There is usually a large range of possible choices of individuals or groups who, if they had acted differently, would have changed the outcome. Any of these people could be said to have caused the outcome – allowing choice about which thread in the tapestry of causation to isolate as culpable. Positive and especially negative events provoke complex representational contests over just who is to blame (or to credit).

Disasters (such as the Black Plague, the 9–11 attacks, the flooding of New Orleans, the Great Depression) present a special opportunity to bring together a punitive coalition and turn it loose on one's status rivals: To play the attribution game, one manufactures representations that promote mental coordination on the interpretation that one's status rivals are at fault – are indeed fitness suppressors of the community at large. This serves as a triggering coordinative signal for those who want to take action against the blamed – that is, it triggers outrage (Tooby, Cosmides, and Price 2006). The blameworthy are picked by an emerging power-based consensus. If an attack on rivals is framed as punishment for a moral transgression that injured the encompassing community, then it disarms defenders of the target of the attack. By standing up for the transgressor, they can be framed as defending moral transgression and injuring the community.

Welfare trade-offs, disgust, and the sorting game: Ejecting cheaters from exchange relationships is a process of disaffiliation. This is one component of a more general adaptive problem: filling one's finite association niches with individuals that have high association value, and removing individuals with low or negative association value. There are differential payoffs to playing this discriminative association game well or poorly (Tooby and Cosmides 1996). Depending on who the actor's associates are, the actor may be helped or exploited, may incur positive or negative externalities, may accrue changes in status, and so on. In dyadic games, affiliation is under direct individual control, but in social networks, your associates' choices influence the actualization of your own preferences. You may be forced to associate with your friend's friends, if you are to maintain your relationship with your friend. This makes conflicts over network affiliation or group membership a second order moral game – and an n-party exchange. The stakes of competitive exclusion and differential affiliation are high, with displacement and ostracism at one extreme and valued centrality at the other.

Several families of pre-existing adaptations were co-opted into adaptations for the discriminative association game. Disgust is the emotion program that evolved to recognize, evaluate, and motivate avoidance of harmful things, and the output of adaptations for discriminative association is motivated avoidance. That is why actions (and characteristics) can lead persons to be seen as disgusting.

Adaptations for computing and responding to WTRs are a second family of programs involved in discriminative association games. Humans are subject to kin selection, reciprocity, retaliation,

bargaining, and other selection pressures which make it important to place weight (under certain conditions) on the welfare of the other as well as the self. Experimental evidence supports the prediction that humans evolved architectures that compute specific WTRs for each social relationship – WTRs that determine which self-benefiting choices, and which other-benefiting choices the actor will make (Tooby et al. 2008; Delton et al. forthcoming). Humans have complementary adaptations which use observations of others behavior to infer the WTRs that others are using. Low WTRs from others to self trigger anger and disgust; WTRs from the self to another that are too low trigger guilt and (if public) shame; and so on (Tooby et al. 2008; Sell, Tooby and Cosmides 2009).

To fill one's finite association niches with others' whose WTRs toward you are low is a fitness mistake – you will rarely be helped, and often exploited. That is, WTR machinery delivers information that should guide behavior in the discriminative association game. The expression by others of a high WTR toward you (and those you value) elicits affection and affiliation; the expression of low WTRs toward you (and those you value) elicits disgust (if the person can be ejected or avoided) or anger (if they cannot). Since it is advantageous to be seen as harboring a high WTR toward others, individuals dissemble by behaving more favorably when they are being observed, or when the cost is low. Someone's WTR-based association value is therefore set by the minimum WTR they are caught expressing – it reveals the true but hidden magnitude of how little they value others – a representation that should be sticky or "staining." Someone who is capable of severely damaging affiliates in pursuit of trivial personal gains is a threat to be avoided. Actions toward ingroup members that express unusually low WTRs elicit the shared value project of avoidance and exclusion. But such acts are only useful for discriminative association if they are intentional, and WTR concepts help us define what intentionality in this domain means. According to the WTR theory of intentionality, intentionality is computed as a tool to help the judger infer a maximum upper bound on the WTR that the actor is using toward the other person. If I choose to benefit myself by 1 unit of welfare, at a cost to you of 50 units, and I know these costs and benefits, then others can infer that my maximum WTR is lower than .02, a very low WTR. This theory predicts that I will be seen as intentionally depriving you of 50 units, just to get myself 1 unit. In contrast, if I choose an option that benefits me by 1 unit, and benefits you by 1 unit (over an option

that benefits neither one of us), we expect I will not be judged to be intentionally helping you, because no (positive) WTR toward you from me can be inferred – I might have done it just to get myself 1 unit, regardless of how this impacted you. Moral credit should show the reverse effect – if the act benefits the self as well as others, it should be seen as less praiseworthy, and my helping you less intentional. Experiments on the so-called *side effect* are consistent with this view – that is, foreseen but negative side effects on third parties of self-interested choices are seen as intentional, but foreseen but positive side effects are not (Knobe 2003). According to WTR theory, one function of intentional attributions is moral inferences about WTRs. In short, blame and credit are partly functions of imputed WTRs, and mental coordination that targeted persons are blameworthy can be mobilized in the discriminative association game to trigger their elimination from the social unit.

Playing moral games defensively: For humans, it is important not only to have adaptations to play moral games offensively, but also defensively. Offensive moral warfare involves attempting to influence others to modify their behavior so that it conforms to your preferences – a process that gains special force when allies are recruited. One set of preferences may involve social exclusion of targets. However, at least as important as offensive play is the complementary process: Individuals have to conduct themselves to minimize retaliation or exclusion from others for any moral missteps. Incautious actions expose individuals to severe or even fatal sanctions by empowered moral coalitions. Individuals can stigmatize themselves, be found at fault for a negative outcome, or be identified as a violator of a moral understanding, whether explicit (e.g., a law) or implicit. A course of action is in the defensive moral domain if it might draw an negative evaluative response, exclusion and/or a punishment from a sufficiently empowered and mentally coordinated number of social actors in the community. The goal of defensive adaptations is to guide behavior in the actor so that it minimizes others' pretexts for moral attack, and maintains others' valuations of the self. Moreover, they seem better designed to win in the assortment game, for example by advertising adherence to restraints that increase the attractiveness of the individual as a close social partner (e.g., would you choose to live with someone who would kill you to save 10 others?). The moral sense operating in the self systematically departs from maximizing general social utility, nor do moral judgments of others' behavior map on to utility maximization (Tooby and Cosmides, 1979).

From the point of view of defensive adaptations, something is a moral issue if others deem it to be one. Blame, condemnation, ostracism, and punishment are endemic to small-scale communities, and very much to be feared. How can they be avoided? First, as already discussed, our psychological architecture should be equipped with a moral sense designed to assimilate and weight others' values, in proportion to their social power and the likelihood of surveillance. (The moral sense will also integrate outputs from the individual's altruistic adaptations, since both put weight on others' welfare.) Second, the computational problems inherent in launching a moral attack are formidable. Defensive moral psychology is designed to make those computational problems more intractable, in order to offer less traction to potential antagonists. Hence, many evolved features of our defensive moral psychology are counterparts to the cognitive coordination problems faced by those on the moral offensive. One obvious application of this is the fact that our evolved moral psychology treats sins of omission as less culpable than sins of commission, making them safer to commit and easier to get away with. When in doubt, do nothing. Similarly, humans are inclined to hide self-interested actions under other publicly valued guises; to voice public support for reigning values; to see where general sentiment lies before publicly committing oneself, and so on. If accused of transgressions we are designed to plead lack of intentionality, shift blame, attack accusers as hypocrites, and intimidate or retaliate against those who are mobilizing the moral attack.

Conclusion

The set of evolved programs that enable and drive warfare and politics strongly overlap with the set of evolved programs that drive human morality. The mapping of these evolved programs and their embedded circuit logics is only in its infancy, and we have only sketched out some of the known or predicted features of our coalitional and moral psychologies. However, progress in this enterprise holds out the possibility of gradually throwing light on some of the darkest areas of human life.

References

Alexander, R. D. 1979. *Darwinism and human affairs*. Seattle: University of Washington Press.

Andrea, J., and A. J. Sedlak. 1979. Developmental differences in understanding plans and evaluating actors. *Child Development* 50(2):536–60.
Aumann, Robert. 1976. Agreeing to disagree. *Annals of Statistics* 4(6):1236–9.
Aumann Robert and Adam Brandenburger. 1995. Epistemic conditions for Nash equilibrium. *Econometrica* 63(5):1161–80.
Baumeister, R. F., and K. L. Sommer. 1997. What do men want? Gender differences and the two spheres of belongingness. *Psychological Bulletin* 122:38–43.
Boehm, C. 1992. Segmentary 'warfare' and the management of conflict: Comparison of East African chimpanzees and patrilinieal-patrilocal humans. In *Coalitions and Alliances in Humans and Other Animals*, edited by A. H. Harcourt and F. M. B. de Waal, 137–73. Oxford: Oxford University Press.
Boehm, C. 1999. *Hierarchy in the forest: The evolution of egalitarian behavior*. Cambridge, MA: Harvard University Press.
Bowles, S. 2006. Group competition, reproductive leveling, and the evolution of human altruism. *Science*, 314:1569–72.
Chagnon, N. 1983. *The Yanomamo*. New York: Holt, Rinehart & Winston.
Cosmides, L. and J. Tooby. 2000. Consider the source: The evolution of adaptations for decoupling and metarepresentation. In *Metarepresentations: A multidisciplinary perspective*, edited by D. Sperber, 53–115. New York: Oxford University Press.
Cosmides, L. and J. Tooby. 2005. Neurocognitive adaptations designed for social exchange. In *The Handbook of Evolutionary Psychology*, edited by D. M. Buss, 584–627. Hoboken, NJ: Wiley.
Chwe, M. S. 2003. *Rational ritual: Culture, coordination, and common knowledge*. Princeton: Princeton University Press.
Darwin, C. 1871. *The descent of man, and selection in relation to sex*. London: John Murray.
Delton, A., L. Cosmides, and J. Tooby. (forthcoming). *The psychosemantics of free-riding*.
Delton, A., J. Tooby, L. Cosmides, D. Sznycer, and J. Lim. (forthcoming). *Welfare trade-off ratios*.
Ermer, E. 2007. Coalitional support and the regulation of welfare tradeoff ratios. Dissertation, University of California, Santa Barbara.
Ermer, E., L. Cosmides, and J. Tooby. 2008. Relative status regulates risky decision-making about resources in men: Evidence for the co-evolution of motivation and cognition. *Evolution and Human Behavior* 29:106–18.
Freud, Sigmund. 1927[1923]. *Das Ich und das Es, Internationaler Psychoanalytischer*, translated as *The Ego and the Id* by Joan Riviere. London: Hogarth Press and Institute of Psycho-analysis.
Haidt J., C. McCauley, and P. Rozin. 1994. Individual differences in sensitivity to disgust: A scale sampling seven domains of disgust elicitors. *Personality and Individual Differences* 16(5):701–13.

Harner, M. 1972. The *Jivaro: People of the sacred waterfalls*. Berkeley: University of California Press.
Hauser, M. 2006. *Moral minds*. New York: Harper Collins.
Heider, Karl 1970. *The Dugum Dani*. Chicago: Aldine Publishing Company.
Henrich, J. and F. Gil-White. 2001. The evolution of prestige. *Evolution and Human Behavior* 22(3):165–96.
Hobbes, T. 1651. *Leviathan, The Matter, forme and power of a common wealth ecclesiasticall and civil*. London: Andrew Crooke and William Cooke Publishers.
Hrdy, S. B. 1980. *The langurs of Abu*. Cambridge, MA: Harvard University Press.
Hume, D. 1751. *An enquiry concerning the principles of morals*. London: A. Millar.
Hutcheson, F. 1728. *An essay on the nature and conduct of the passions and affections, with illustrations upon the moral sense*. Dublin: J. Smith and W. Bruce.
Ip, G. W., C. Y. Chiu, and C. Wan. 2006. Birds of a feather and birds flocking together: physical versus behavioral cues may lead to trait- versus goal-based group perception. Journal of Personality and Social Psychology 90(3):368–81.
Keegan, J. 1994. *A history of warfare*. New York: Random House.
Keeley, L. H. 1996. *War before civilization: The myth of the peaceful savage*. New York: Oxford University Press.
Knobe, J. 2003. Intentional action and side effects in ordinary language. *Analysis* 63:190–3.
Kuran, T. 1998. *Private truths, public lies: the social consequences of preference falsification*. Cambridge, MA: Harvard University Press.
Kurzban, R., J. Tooby, and L. Cosmides. 2001. Can race be erased? Coalitional computation and social categorization. *PNAS* 98(26):15387–92.
LeBlanc, S. A. 1999. *Prehistoric warfare in the American Southwest*. Salt Lake City: University of Utah Press.
LeBlanc, S. A. and K. Register. 2003. *Constant battles: The myth of the peaceful, noble savage*. New York: St. Martins Press.
Lee, R. and I. DeVore. 1976. *Kalahari hunter-gatherers*. Cambridge, MA: Harvard University Press.
Lieberman, D., J. Tooby, and L. Cosmides. 2007. The architecture of human kin detection. *Nature* 445(7129):727–31.
Lieberman, D., J. Tooby, and L. Cosmides. 2003. Does morality have a biological basis? An empirical test of the factors governing moral sentiments relating to incest. *Proceedings of the Royal Society London (Biological Sciences)* 270(1517):819–826.
Lewis, David. 1969. *Convention: A philosophical study*. Oxford: Blackburn.
Manson, J. H., and R. W. Wrangham. 1991. Intergroup aggression in chimpanzees and humans. *Current Anthropology* 32(4):369.

Maynard Smith, J. 1964. Group selection and kin selection. *Nature* 201:1145–7.
Maynard Smith, J. 1977. Parental investment: A prospective analysis. *Animal Behaviour* 25:1–9
Olson, M. 1965. *The logic of collective action: Public goods and the theory of groups*. Cambridge, MA: Harvard University Press.
Pemberton, M. B., C., A. Insko, and J. Schopler. 1996. Memory for and experience of differential competitive behavior of individuals and groups. *Journal of Personality and Social Psychology* 71:954–66.
Price, M. E., L. Cosmides, and J. Tooby. 2002. Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior* 23:203–31.
Rofe, Y. 1984. Stress and affiliation: A utility theory. *Psychological Review* 91:235–50.
Sahlins, M. 1961. The segmentary lineage: An organization of predatory expansion. *American Anthropologist* 63:322–45.
Sell, A., L. Cosmides, J. Tooby, D. Sznycer, C. von Rueden, and M. Gurven. 2009. Human adaptations for the visual assessment of strength and fighting ability from the body and face. (Plus electronic supplemental material). *Proceedings of the Royal Society B (Biological Sciences)* 276(1656):575–84.
Sell, A., J. Tooby, and L. Cosmides. 2009 Anger, strength, and the logic of formidability. *Proceedings of the National Academy of Sciences*. Published in online Early Edition, 3 August 2009.
Sherif, M. 1966. *In common predicament: Social psychology of intergroup conflict and cooperation*. Boston: Houghton-Mifflin.
Sidanius, J., and F. Pratto. 2001. *Social dominance*. Cambridge: Cambridge University Press.
Smirnov, O., H. Arrow, D. Kennett, and J. Orbell. 2007. Ancestral war and the evolutionary origins of "heroism." *The Journal of Politics* 69:927–40. Cambridge: Cambridge University Press.
Tajfel, H., and J. C. Turner. 1979. An integrative theory of intergroup conflict. In *The Social Psychology of Intergroup Relations*, edited by W. G. Austin and S. Worchel, 33–47. Monterey: Brooks Cole.
Tiger, L. 1969. *Men in groups*. Toronto: Thomas Nelson and Sons Ltd.
Tooby J. and I. DeVore. 1987. The reconstruction of hominid behavioral evolution through strategic modeling. In *Primate Models of Hominid Behavior*, edited by W. Kinzey, 183–237. New York: SUNY Press.
Tooby, J., and L. Cosmides. 1979. Adaptationist approaches to moral phenomena. *Institute for Evolutionary Studies Technical Report* 79–1.
Tooby, J., and L. Cosmides. 1988. The evolution of war and its cognitive foundations. *Institute for Evolutionary Studies Technical Report* 88–1.
Tooby, J., and L. Cosmides. 1996. Friendship and the banker's paradox: Other pathways to the evolution of adaptations for altruism. In *Proceedings of the British Academy*, 88:119–43.

Tooby, J., and L. Cosmides. 2000. Cognitive adaptations for kin-based coalitions: human kinship systems at the intersection between collective action and kin selection. *Current Anthropology* 41(5):803–4.

Tooby, J., and L. Cosmides. 2008. The evolutionary psychology of the emotions and their relationship to internal regulatory variables. In *Handbook of Emotions*, edited by M. Lewis and J. M. Haviland-Jones, 3rd edition, 91–115. New York: Guilford.

Tooby, J., L. Cosmides, and M. Price. 2006. Cognitive adaptations for n-person exchange: The evolutionary roots of organizational behavior. *Managerial and Decision Economics* 27:103–29.

Tooby, J., L. Cosmides, A. Sell, D. Lieberman, and Sznycer, D. 2008. Internal regulatory variables and the design of human motivation: A computational and evolutionary approach. In *Handbook of approach and avoidance motivation*, edited by A. J. Elliot, 251–71. Mahwah, NJ: Lawrence Erlbaum Associates.

Tooby, J., L. Cosmides, and H. C. Barrett. 2005. Resolving the debate on innate ideas: Learnability constraints and the evolved interpenetration of motivational and conceptual functions. In *The Innate Mind: Structure and Content*, edited by P. Carruthers, S. Laurence, and S. Stich, 305–37. New York: Oxford University Press.

Turney-High, H. 1949. *Primitive war.* Columbia: University of South Carolina Press.

Woodfine, P. 1998. *Britannia's glories: The Walpole ministry and the 1739 War with Spain.* Woodbridge, UK: Royal Historical Society/Boydell Press.

Wilson, M. L., M. D. Hauser, and R. W. Wrangham. 2001. Does participation in intergroup conflict depend on numerical assessment, range location, or rank for wild chimpanzees? *Animal Behaviour* 61:1203–16.

Wilson, M. L., and R. W. Wrangham. 2003. Intergroup relations in chimpanzees. *Annual Review of Anthropology* 32:363–92.

Wrangham, R. W. and D. Peterson. 1996. *Demonic males: Apes and the origins of human violence.* New York: Houghton Mifflin.

Zimmerman, L. 1981. *The Crow Creek Site massacre: A preliminary report.* US Army Corps of Engineers, Omaha District.