CHAPTER 7

JUSTICE, FAIRNESS, AND STRATEGIC EMOTIONAL COMMITMENT

Timothy Ketelaar and Bryan Koenig
New Mexico State University

In this chapter we examine the role that emotions play in the psychology of justice by exploring the possibility that the emotions that underlie the human desire to "play fair" evolved, not to promote justice, but rather to promote the fitness of the individuals experiencing these moral sentiments. In short, we argue that the human capacity for fairness is a byproduct of evolved emotional commitment devices that functioned—in ancestral environments—to promote evolutionary fitness rather than fairness per se. In particular, we argue that natural selection has generated a variety of strategy types ranging from ruthlessly self-interested individuals who are emotionally predisposed to behave "unfairly" to apparently altruistic individuals who are emotionally predisposed to "play fair." These divergent preferences for fairness exist, we contend, as part of a stable population structure where individuals who routinely play fair ultimately receive the same "payoff" as individuals who routinely behave unfairly.

Advances in the Psychology of Justice and Affect, pages 129–149

OVERVIEW

We begin by articulating the relationship between the concept of fairness and the several kinds of justice that social scientists have studied. We then review three lines of empirical evidence demonstrating that substantial individual differences exist in the importance that individuals attach to fairness. Finally, we apply a game theoretic framework to demonstrate how different emotion-based strategic commitments could interact to produce a stable population structure consisting of several distinct strategy types, including both fair-minded and unfair-minded strategists. Our aim is to shed light on the psychology of justice by attempting to account for the wide range of variation in the human capacity to value fairness. We begin by recognizing that the concept of justice refers, in large part, to concerns about fairness, often expressed in terms of how one can achieve a satisfactory allocation of resources among agents (DeCremer & Tyler, 2005).

JUSTICE AS FAIRNESS

Most justice scholars would agree that the criteria that human minds routinely apply to judgments of fairness fall into three broad categories: (a) judgments about the fairness of distributions, (b) judgments about the fairness of acts of punishment and compensation, and (c) judgments about the fairness of the methods and procedures employed to generate these outcomes. These three categories of concerns have traditionally been labeled distributive, retributive, and procedural justice, respectively.

Distributive Justice

Distributive justice concerns the fairness of allocations of scarce resources, typically revolving around the concepts of equality and equity. Whereas *equality* implies that all agents receive physically identical outcomes (equivalent slices of the pie), *equity* extends the concept of fairness to incorporate more abstract ideas such as needs, tastes, and merit (see Bar-Hillel & Yaari, 1993; Deutsch, 1975, 1985). An equitable distribution of a company's reserved parking spaces, for example, may be deemed "fair" even though outcomes aren't physically equivalent (e.g., a merit-based policy in which senior employees have greater access to prime parking spaces than junior employees). By contrast, distributive justice according to the principle of equality suggests that the distribution of prize money in a bingo parlor, for example, will be deemed "fair" if each of the k winners receives $1/k$th of the pie.

Retributive Justice

Retributive justice, on the other hand, is comparable to a secular interpretation of karma. In Hindu and Buddhist philosophy, karma is the virtual account of one's morally positive and negative actions that determine whether an individual will ultimately be reborn in a less fortunate (bad karma) or more fortunate (good karma) condition. Accordingly, morally good actions build up good karma, whereas morally bad actions build up bad karma. In this sense, retributive justice is said to occur when a person's actions are appropriately rewarded and compensated (i.e., morally bad behaviors are punished rather than rewarded and vice versa, see Tyler & Boeckmann, 1997). This secular interpretation of karma suggests that retributive justice is achieved when individuals and institutions met out appropriate punishments for immoral behaviors and appropriate rewards for moral actions.

Procedural Justice

Finally, procedural justice concerns the fairness of the actual methods employed to allocate resources (Thibaut & Walker, 1975). The mere existence of the idea of procedural justice is evidence that we live in a noisy world where even the most perfect methods for fairly allocating resources (distributive justice) will sometimes be incorrectly implemented or not implemented at all. In this light, procedural justice refers to the estimation of the degree to which a particular method or procedure for allocating resources is deemed fair, even if this process—due to the noisy world—does not always generate fair distributions (distributive justice) and/or appropriate karmic rewards and punishments (retributive justice). Rawls (1971) notion of a "the veil of ignorance-where one establishes social distribution rules without knowing ones own social standing—is one possible procedure for allocating resources that can be evaluated in terms of its fairness regardless of whether or not this procedure always produces outcomes that are deemed fair. Because many forms of resource distribution are overseen by bureaucratic agencies, procedural justice is especially important in the evaluation of the activities of groups, authorities, and institutions (see De Cremer & Tyler, 2005).

THREE LINES OF EVIDENCE FOR INDIVIDUAL DIFFERENCES IN PREFERENCES FOR FAIRNESS

Although a concern for fairness appears to be an important facet of our human nature, one can still ask whether this concern ranks as more (or

less) important that other values that one might hold. Will most people forsake a fair distribution procedure, for example, simply because a respected authority figure demands that they employ an unfair procedure? In other words, what happens when an individual is confronted with a choice between being fair and serving some other core value, such as maximizing one's immediate material self-interest or obeying a respected authority? These sorts of questions beg a still larger question: Do most individuals rank concerns for fairness above all other concerns, such as material self-interest or obedience to authority?

In this section, we review empirical evidence that suggests that there is substantial variation (individual differences) in the relative importance that individuals attach to "playing fair." These individual differences conform, we argue, to a polymorphic distribution of strategy types, ranging from individuals who literally seem compelled to "play fair" to others who appear to value fairness not at all. Rather than assuming a universal human tendency "to play fair" that can be plotted as a normal distribution with random variation in how much individuals value fairness, research in political ideology, experimental economics, and social motives suggests something quite different. Across three domains—political ideology, experimental economics, and social motives—the picture that emerges, we argue, is one in which human nature is comprised of several different types of strategists, ranging from individuals who are compelled to "play fair" to individuals who are compelled to be more concerned about immediate self-interest than fairness. We then illustrate how individual differences in strategic emotional commitments could account for these divergent types of preferences.

Political Ideology

Individual differences in the priorities that people assign to core values—such as fairness—are perhaps most visible in the realm of politics where opposing factions clash over issues ranging from how much influence governments should play in the education of children to whether gay marriages should be legal. As one sociologist (Hunter, 1990) notes:

AU: 1991 in REFs

> Ultimately the battle over this symbolic territory reveals a conflict over world views—over what standards our communities and nation will live by; over what we consider to be 'of enduring value' in our communities… (p. 248)

Of particular relevance to justice researchers is the tendency for societies with plurality (majority) voting systems to generate two large and distinct coalitions of *conservative* and *progressive* worldviews. The United States,

for example, has its republicans (conservative) and democratic (progressive) parties, England has its conservative and labour (progressive) parties, and Germany has its Christian Social Union (conservative) and Social Democratic party (progressive) coalitions. This apparently systematic emergence of conservative and progressive factions is relevant to justice research precisely because there is now considerable evidence that, compared to individuals who adopt a conservative world-view, individuals who adopt a more progressive world-view tend to assign strikingly different levels of importance to "fairness" relative to other values (Haidt & Joseph, 2005, Haidt & Graham, 2006, 2007).

AU: 2004 in REFs

AU: in press in REFs, no 2006 or 2007

One conclusion that clearly emerges from these studies of political world-views is that conservatives do not simply devalue the domains that progressives value. Rather, it appears that conservatives tend to view several core values as being just as important as fairness, whereas cultural progressives emphasize just two—"preventing suffering" and "being fair"—and assign less importance to other core values. In one study, Haidt and Graham (2006) asked individuals to rate their political orientation on a 7-point (1 = extremely conservative, 7 = extremely liberal) scale. Individuals were then asked: "When you decide whether something is right or wrong, to what extent are the following concerns relevant to your thinking?" Over 1,500 participants in this on-line study were then given descriptions of five core values:

AU: in press in REFs, no 2006

- Harm—whether or not someone was harmed
- Reciprocity—whether or not someone acted unfairly
- In-group—whether or not someone betrayed his or her in-group
- Hierarchy—whether or not the people involved were of the same rank
- Purity—whether or not someone did something disgusting

Individuals who rated themselves as extremely progressive (liberal) or extremely conservative, assigned dramatically different amounts of relative importance to these five core values. Of relevance to justice research is the question of whether conservatives and progressives assign the same relative weight to "fairness" when they evaluate whether something is right or wrong. As Haidt and Graham (2006) demonstrate in several studies, they clearly do not. The moral intuitions of conservatives tend to be based evenly upon all five core values. That is, conservatives rate "fairness" as being very relevant to their judgments of whether something is right or wrong, yet they also rank the other four values (harm, in-group, hierarchy, and purity) as equally important. By contrast, the moral intuitions of progressives are based primarily on just two core values: *reciprocity* and *harm*. That is, in their judgments of whether something is right or wrong, progressives tend to rate "fairness" (and concerns for preventing suffering) as much more relevant than concerns about in-group, hierarchy, and purity.

AU: in press in REFs, no 2006

AU: in press in REFs, no 2006

Haidt and Graham (2006) argue that concerns fairness (and preventing harm) constitute approximately one half of the moral psychology of progressives, whereas concerns for fairness constitute just one-fifth of the moral psychology of conservatives. In short, progressives rank "concerns for fairness" as paramount in importance when evaluating the goodness or badness of outcomes, whereas conservatives view such concerns as being no more important than respecting authorities or supporting one's ingroup. Thus, although justice concerns are an important aspect of our human nature, they are not universal in the sense that all humans rank fairness as the most important moral virtue (Haidt & Graham, 2006).

AU: in press in REFs, no 2006

Experimental Economics

Fairness and justice are not just moral values studied by political scientists; these concepts also play an important part in everyday economic decision-making. A concern for fairness is often invoked, for example, as an explanation for many intriguing findings in behavioral economics, a branch of economics that studies decision behavior in social interactions modeled as strategic games[1] (Camerer, 2003; Camerer & Fehr, 2004; von Neumann & Morgenstern, 1944). Behavioral economists have studied a wide spectrum of strategic social interactions ranging from strategy games where one participant is given a sum of money and asked to divide it with another participant who can either reject an unfair offer, in which case neither player gets anything (an ultimatum game[2]), or must accept the offer (a dictator game); to so-called public goods games where participants are allowed to make limited anonymous withdrawals from a collectively shared resource (a common pool resource game), or where participants are given an initial monetary endowment and are prompted to make simultaneous anonymous contributions of any portion of their endowment to the common group fund (a voluntary contribution game[3]).

Research in behavioral economics reveals that although many individuals behave in a self-interested manner consistent with standard economic theory, a sizeable portion of individuals appear to value fairness more than their own immediate material payoffs. Studies of ultimatum games reveal, for example, that although mean offers are less than a completely fair division of the resources (mean offers are around 40% of the pie), most offers are usually substantially more than the typical respondent's minimally acceptable offer (Camerer, 2003). Such findings suggest that many participants are as concerned about making a fair offer as they are about making an offer that won't be rejected. Similarly, in studies of dictator games, a sizeable percentage of individuals elect to give 50% of the pie to the other player, despite the fact that the recipient cannot reject any offer. Taken

together, these findings suggest that concerns for fairness can explain some of the deviations from strictly self-interested behavior observed in ultimatum and dictator games (see Camerer, 2003 for a review).

Studies of public goods games provide further evidence that people care about fairness in economic decision-making. A common finding in public goods games is that a substantial proportion of players abstain from the payoff-maximizing, but clearly unfair, strategy of contributing nothing to a public good (or withdrawing the maximum from a collective resource). Instead, individuals frequently make rather generous contributions to the public good or only modest withdrawals from the collective resource. Moreover, some participants routinely punish others who fail to play fair even when such acts of punishment are costly for the individual doing the punishing (Camerer, 2003; Camerer & Fehr, 2004; Fehr & Gächter, 2002). Although troubling for some economists who wonder why participants do not invariably attempt to selfishly maximize their own payoffs, such findings are hardly surprising to a layperson with an intuitive understanding of fairness and retributive justice.

Numerous behavioral economic studies (e.g., Roth 1999) have observed similar preferences for fairness among college students in several different countries, however, these studies do not tell us whether a concern for fairness extends beyond highly educated individuals living in westernized market economies. To address this question, the MacArthur Preference Network assembled a group of over a dozen scholars from economics, anthropology, evolutionary biology, and psychology to explore economic behavior in 15 small-scale societies ranging from band-level foraging societies to semi-nomadic herding communities (Henrich, Boyd, Bowles, Camerer, Fehr, & Gintis, 2004).

The results of these studies are remarkable in two respects. First, they provide evidence for considerable within culture and across cultural variability in fair behavior. In the ultimatum game, for example, the mean offers spanned from 25% of the total pie to 57% across different field sites, a considerably greater range than the typical mean offers of between 42% and 48% of the pie observed across university samples. Interestingly, although modal offers in the ultimatum game were near 50% of the pie across most of the field sites, most distributions of offers were clearly skewed toward lower offers (suggesting a proclivity for self-interested behavior) as opposed to being randomly distributed around a mean offer of 50% of the pie. Second, considerable variability has also been observed in responses to unfair offers in the ultimatum game. With students, rejections are quite common, especially when offers are below 20% of the pie (see Camerer, 2003). However, across the 15 field sights in the MacArthur Preference Network project, many groups accepted 100% of the offers below 30% of the pie while other groups routinely rejected hyper-fair

offers of more than 50% of the pie (see Henrich, et al., 2004). Findings like these, where individuals in some societies routinely accept unfair offers and where individuals in other societies routinely reject hyper-fair offers, suggest that there is a great deal of variability within and across cultures in regards to fair-minded behavior.

Social Motives Research

Even within a given cultural milieu there are important individual differences in the preferences that people report for how they prefer to distribute resources (distributive justice). The robustness of these individual differences is especially evident in research on social value orientations (Au & Kwong, 2001; Messick & McClintock, 1968; Van Lange, Otten, DeBruin, & Joireman, 1997). Social value orientation refers to stable individual differences in preferences for distributing resources and is typically assessed by asking participants to choose from one of several different payoff distributions in a series of hypothetical dictator games. This "decomposed game" methodology, in which participants are asked to select one of several different distributions of a resource, typically results in three types of social preferences referred to as: Cooperators, Individualists, and Competitors (Au & Kwong, 2001; Van Lange, Otten, De Bruin, & Joireman, 1997; Van Lange & Visser, 1999). *Cooperators* prefer outcomes that are fair, namely, outcomes that maximize joint payoffs between participants. *Individualists*, by contrast, prefer outcomes that are not fair; that is, outcomes that maximize their own payoffs regardless of what the other receives. Finally, *Competitors* tend to choose unfair outcomes that maximize the difference between their payoff and that of the other person, even if this means accepting an outcome that is lower than what they could achieve otherwise.

Numerous cross-cultural studies on social motives reveal that although the most common preference for how to distribute a resource is a fair distribution, not all individuals value fairness above self-interest. In a review of 47 studies that employed the decomposed game methodology to identify social value orientations, Au and Kwong (2001) found that the three clusters of social motives (Cooperators, Individualists, and Competitors) account for approximately 87% of the adult population (the other 13% are typically uncategorizable). Across all of the samples, the largest of these clusters (*Cooperators*) represented approximately 46% of the population. The second largest cluster—representing 25% of the population across samples—consisted of individuals who routinely prefer an unfair outcome if the outcome maximized their material gain (*Individualists*). Finally the smallest cluster—just 13% of the population—corresponded to highly non-

AU: 2002 in REFs

AU: 2002 in REFs

AU: 2002 in REFs

cooperative individuals who prefer unfair outcomes that give them the largest relative advantage over their partner (*Competitors*). The relative frequency of these three strategy types appears to be quite stable across all age groups[4] with a ratio of approximately 4:2:1 of Cooperators, Individualists and Competitors (see Van Lange, Otten, DeBruin, & Joireman, 1997). Moreover, several studies have shown that these stable preferences are predictive of actual strategic behavior in social bargaining games such as the Prisoner's Dilemma and the ultimatum game (Ketelaar & Au, 2003; Van Lange, Otten, DeBruin, & Joireman, 1997). In sum, research on social motives suggests that human nature consists of several distinct types of strategists and not just a normal distribution with random variation around a central tendency to play fair.

EMOTIONS AS STRATEGIC COMMITMENT DEVICES UNDERLYING PREFERENCES FOR FAIRNESS

Research in political ideology, experimental economics, and social motives reveals substantial variability in the relative importance that individuals assign to "fairness" compared to other values (e.g., material self-interest, obedience to authorities, etc.). We argue that this variation—in the importance that individuals attach to "fairness"—reflects individual differences in the strategies that agents employ in the indefinitely repeated series of bargaining games known as social life (Ketelaar, 2004, 2006). In this view, individuals who consistently propose fair offers in situations that resemble an ultimatum game can be treated as agents who are simply playing a different strategy than those individuals who consistently propose unfair offers. What makes this "preferences-as-strategies" view so enticing is that these individual differences—in preferences for fairness—correspond quite well to a class of decision rules known as *commitment devices* (Hirshliefer, 1987).

The idea of a *commitment device* comes from economists who point out that decision makers are often confronted with a choice between two paths: one path that leads to an immediate reward, but which is ultimately costly; and an alternative path that, although not immediately rewarding, results in a better long-term outcome (Hirshleifer, 1987; 2001). The recovering alcoholic, for example, may be faced with the choice between a tasty martini that can be enjoyed right now or a non-alcoholic beverage that, although not as pleasurable to consume, is more likely to produce a positive long-term outcome. In a similar fashion, the dieter may face a choice between eating a tasty morsel of cake versus consuming a more bland healthy food that promises a better long-term outcome. Each of these scenarios provides an example of what economists call the "commitment

problem,"[5] a decision dilemma that arises whenever immediate incentives run contrary to long-term self-interest (Frank, 1988; Hirshleifer, 1987; Schelling, 1960). In this context, a commitment device can be seen as any mechanism that compels an agent to stay on the path toward positive long-term outcomes despite the spurious attractiveness of tempting alternative paths that promise immediate rewards.

According to economist Jack Hirshliefer (2001), commitment devices come in two flavors: *pre-emptive commitments* and *reactive commitments. Pre-emptive commitments* consist of irrevocable first-moves whereby the agent—by performing this action—transforms the situation into one in which she is literally compelled to act in her own long-term self-interest. The recovering alcoholic, for example, may consume the medication *Antabuse* as a *pre-emptive* commitment device to overcome the spurious attractiveness of a drink.[6] In the context of social interactions, a pre-emptive commitment is an irrevocable first-move whereby one agent—by performing this action—transforms the subsequent social situation into a strategy game where a second party would be foolish not to take this information (about his opponent's first move) into account when deciding on his own course of action.

Although the strategic logic of a pre-emptive commitment may sound convoluted, consider the simple example of a military commander who realizes that he is now confronted by an enemy army that is considerably larger than his own. The commander is in a situation that effectively has just three possible courses of action: (a) battle to the death against overwhelming odds, (b) humiliating surrender without doing battle, or (c) humble retreat without doing battle. According to Sun Tzu (1963), author of *The Art of War,* the wise commander will realize that he can turn the situation to his advantage by instructing his soldiers to burn their bridges behind them. By engaging in the seemingly irrational act of destroying all possible means of retreat, the commander effectively transforms this dire scenario into a strategy game with just two options: (a) humiliating surrender or (b) battle to the death. Because soldiers trained in the "Art of War" will prefer the honor of battling to their death to the dishonor of a humiliating surrender, the act of burning his own bridges serves as a pre-emptive commitment that the enemy general would be foolish to ignore. The enemy general, who would much prefer his opponent to surrender than to fight to the death, will be himself compelled, or so argues Sun Tzu (1963), to avoid doing battle until the situation is more propitious. Although many forms of pre-emptive commitment are much less romantic and involve fewer participants than this rather contrived military example, the underlying logic is the same—a pre-emptive commitment can transform a social situation into a strategy game where the second party is now compelled to act in the self-interest of the first party who has made an irrevocable first move.

The second type of strategic commitment device is a *reactive* commitment. In the context of social interactions, a reactive commitment corresponds to a convincing pledge, made by one agent, to respond in a contingent manner to the first move of another agent (Hirshliefer, 2001). Reactive commitments come in two forms: *promises* and *threats* (Hirshleifer, 2001). A *threat* is a convincing pledge to inflict a cost upon the other party if the other party makes a particular first move. More formally, a threat is any communicative act of the form "IF X THEN Y" where X refers to some specific action that the recipient of the signal can choose to perform and Y is some costly (to the recipient of the signal) action that the sender of the signal will contingently bestow upon the recipient if the recipient indeed performs action X. Similarly, a *promise* is a convincing pledge to bestow a benefit upon the other party if the other party makes a particular first move. Thus, a promise is any communicative act of the form "IF X THEN Y" where X refers to some specific action that the recipient of the signal can choose to perform and Y is some beneficial (to the recipient) action that the sender of the signal will contingently bestow upon the recipient if the recipient indeed performs action X.

For both threats and promises the contingent nature of the signaler's intentions must be perceived as credible in order for the pledge (threat or promise) to serve as an effective incentive or deterrent for the second agent's behavior (Hirshleifer, 2001). So how can one determine whether a particular threat or promise is a strategic commitment that the signaler is literally compelled to follow through on, or is simply cheap talk? We turn now to an exploration of the claim that some emotions operate as strategic devices that guarantee pre-emptive and reactive commitments (Hirshleifer, 1987; Frank, 1988).

Emotions as Pre-emptive and Reactive Commitments: The Case of Blushing and Gloating

Certain pro-social emotions (e.g., embarrassment, shame, etc.) appear to be excellent candidates for *pre-emptive commitment devices* that signal to others that a particular individual is compelled to play a particular first move in certain social situations. Consider the following example in which an emotional signal might convey strategic information that would allow you to predict whether a person is likely to behave in an unfair manner in a resource distribution task. You are walking near the university when you notice a close friend who is sitting in the driver's seat of a car that appears to be illegally parked in a handicapped zone. A parking enforcement officer is handing your friend a ticket. Your friend, noticing that you are observing the scene, displays either: (a) a blank (neutral) expression or

(b) a red-faced blush accompanied by an embarrassment display. If you are like most people, you will interpret your friend's emotional display (or lack thereof) as a strategic signal. Previous research shows that in ambiguous situations such as this, where it is not clear whether your friend's transgression (parking illegally) was premeditated, you are more likely to interpret your friends behavior as an intentional act—that he intended to violate the law—if he blushes than if he does not blush (De Jong, Peters, De Cremer & Vranken, 2002). Conversely, you are more likely to infer that the person who did not blush in the parking ticket scenario is somehow less guilty of intentionally violating this social norm than the person who did blush in this situation. What makes an embarrassed blush such a good candidate for signal of strategic commitment is that is is unfakeable—it is simply not possible for an agent to voluntarily produce (or inhibit) a blush**.**

From a game-theoretic perspective, your friend's emotional display (blushing or not) in response to being publicly observed violating an important social norm is his first move in the subsequent social bargaining game[7] that he is playing with you. By displaying an embarrassed blush, one could argue that your friend is signaling to you that he is compelled to act in his own self-interest in certain contexts, namely when he believes that the costs are low enough (e.g., when he perceives that there is no one around to enforce the rules). Thus, as an observer of this signal, you could use this information to predict that the blushing individual will be more inclined, compared to the non-blushing individual, to distribute money in a selfish manner in an anonymous dictator game (where, by design, there is essentially no cost for making an unfair offer). Before we explain how an individualist could actually benefit from signaling this apparently "less-than-pro-social" disposition to others (e.g., by blushing), we consider one more example of a strategic emotional commitment device.

Although some emotions convey pre-emptive commitments to behave in a certain manner, other emotions (e.g., anger, gratitude, etc.) are excellent candidates for *reactive commitment devices* that signal that an individual is compelled to carry through with a threat or promise. Consider the following example in which an emotional signal can convey strategic information that allows you to predict whether this person is likely to engage in costly[8] acts of punishment aimed towards norm violators (retributive justice). Imagine again that you are observing your friend in the parking ticket scenario described above. However, this time you notice that the parking enforcement officer is displaying a true smile of pleasure while giving the ticket to your friend. Moreover, you observe that a bystander, a fellow student, is displaying a similarly contemptuous smile of enjoyment as she walks by the scene. These enjoyment smiles could be interpreted as strategic emotional signals. Although there is considerably less research on the strategic meaning of smiles that occur when a person is punishing another

(a form of gloating) or whilst observing another being person being punished (a form of *schadenfreude*[9]), we propose that smiles of enjoyment that occur in such contexts are excellent candidates for strategic signals that communicate the emotional commitments of the agents displaying them.

Consistent with the idea that smiles displayed while administering punishment (gloating) or while merely observing punishment (schadenfreude) could be signaling an individual's predisposition to enjoy punishing norm violators, a recent study in neuroeconomics reveals that a brain region associated with pleasure (i.e., a sub-cortical region of the brain referred to as the striatum) shows heightened activation when participants punish an individual who cheated them in a social bargaining game (DeQuervain, Fischbacher, Treyer, Schellhammer, Schnyder, Buck, & Fehr, 2004; Knutson, 2004). From a game-theoretical perspective, flashing an enjoyment smile while punishing could signal to others that the agent doing this signaling is literally compelled to enjoy punishing norm violators. What makes a smile a good candidate for a strategic commitment display is the fact that a true smile of enjoyment—a so-called Duchenne smile—is difficult to fake. We propose that a Duchenne[10] smile—emitted whilst punishing a norm violator—is likely to be interpreted as a genuine signal of a commitment to punish, or at least as a genuine indicator that the agent truly enjoys administering acts of retributive justice. If it is only those individuals who are willing or able to incur the costs of administering retributive justice upon norm violators that display this gloating Duchenne smile, then these smiles (or their absence) could effectively signal one's predisposition (or lack thereof) toward retributive punishment.

Emotional Commitments as Frequency Dependent Strategy Types

The idea that individual differences in certain emotions can give rise to an assortment of different strategy types can be illustrated by considering an analogy borrowed from the world of chess-playing computers. Looking ahead just 10 moves in a typical chess match can often involve constructing a game tree containing over 40 billion sequences of moves. For this reason, most computer chess programs employ a simpler strategy in which a "move evaluation function" is employed to sort through a limited game tree of possible future moves to locate a "best" next move (Shannon, 1950; Newborn, 1996). The use of a "move evaluation function" often results, interestingly, in instances where one computer chess program will view a particular chess move as "good" while a different computer program, with a different "move evaluation function," will view the same move as "bad." This happens simply because neither computer chess program is evaluat-

AU: not in REFs

ing these moves in terms of their *immediate* payoffs. Instead, each computer program is computing the likely *future* payoffs associated with a given chess move and each computer may use a somewhat different move evaluation function to do this. By simply giving more or less weight to a particular variable (e.g., the defensive value of a position) two different chess computers can evaluate the very same move in quite different ways (Newborn, 1996). It is in precisely this sense that individual differences in emotions might be analogous to the different move evaluation functions employed by competing chess computers (Ketelaar & Todd, 2001). Indeed, the typical view of emotional commitment devices (Frank, 1988; Hirshleifer, 1987; Ketelaar & Au, 2003) argues that emotions (such as guilt) function to move the estimated *future* cost of particular action (such as bad move in the social world) into one's *current* situation, in the form of a powerful and overwhelming emotional state is evoked merely by contemplating that particular action (Ketelaar & Todd, 2001). It is in this sense that one might argue that the "emotional wiring" of a punisher strategist (i.e., one who routinely punishes norm violators) may literally compel the agent to feel good whilst administering punitive acts of retributive justice. Thus, it is as if the "move evaluation function" of the punisher strategist automatically computes that punishing is a "good" next move; whereas for other agents (non-punishers) their "move evaluation function" computes that punishing is not a "good" move. So why is it the case that different human agents appear to have very different beliefs about what constitutes a "good" next move in the series of repeated chess moves that one might refer to as social life? We argue that evolutionary game theory provides a possible answer to this question: frequency dependent selection of strategy types.

AU: not in REFs

AU: not in REFs

Evolutionary game theory provides an account of how several distinct strategy types (punishers, free riders, cooperators, etc.) can be maintained in the same population as a polymorphic equilibrium where (a) each strategy type receives the same average payoff as every other strategy type and (b) each strategy type experiences lower average payoffs if it becomes more or less frequent in the population (Lomborg, 1996; Maynard Smith, 1982). This evolutionary selection process, in which the success of each strategy type depends upon the relative frequency of other strategy types in the same population, is known as *frequency dependent selection* (Maynard Smith, 1982). Computer simulation studies show that frequency dependent selection can be a potent force in shaping the population structure that emerges over evolutionary time and can lead to some very interesting population dynamics (Lomborg, 1996). Political scientist Bjorn Lomborg (1996), for example, conducted a series of computer simulations of a repeated Prisoner's Dilemma game based upon Axelrod's (1984) original computer tournament. Rather than starting with a diverse set of strategies (as was the case in Axelrod's well-known studies), Lomborg's simulations

began with all agents adopting the very same self-interested strategy: defection. Lomborg then modeled a noisy world in which the initial population was allowed to evolve over hundreds of thousands of generations via a combination of mutation (modeled by small random changes in strategies) as well as "cultural-evolutionary" processes such as innovation and imitation. In this complex and noisy environment, Lomborg (1996) observed something rather remarkable: in a majority (60%) of the simulations a frequency dependent population structure consisting of three distinct strategy types consistently emerged. The most common strategy type to evolve (43% of the agents in Lomborg's simulated population) consisted of agents who invariably cooperated and essentially never punished non-cooperators. The next most common strategy type to emerge (18% of the population) consisted of agents that, although quite cooperative, appeared to function mainly as punishers who routinely imposed sanctions on the relatively smaller number of ruthlessly non-cooperative agents that made up approximately 10% of the population. Similarities between Lomborg's "evolved" population structure and the known distributions of social motives in the human population (see Au & Kwong, 2002; Ketelaar, 2004) are intriguing. In each case, nearly twice as many agents were predisposed to continuously play fair (cooperators) than were predisposed to behave in a strictly selfish manner.

Given that frequency dependent selection can generate a stable population consisting of both fair-minded and unfair-minded individuals, one can still ask why these individuals would ever signal their strategic dispositions to one another. Such a scenario (mutual signaling) seems ripe for exploitation. In particular, some agents could use these signals to their own advantage, as when an individualist strategy type decides to exploit another agent who is emitting the signal of a perpetual cooperator, or when an individualist decides to wait until another agent—who has signaled that they are a punisher type—leaves the scene before exploiting her next victim. In such cases, it seems that signals that advertise one's strategy type are unlikely to evolve because they appear to benefit the audience more than the signaler!

Why Would an Agent Signal That They Do (or do not) "Play Fair?"

Costly signaling theory suggests that the propensity to signal one's strategy type will depend upon the relative costs and benefits of signaling (Searcy & Nowicki, 2005). An agent should not invariably ignore a particular class of signal simply because these signals are sometimes dishonest. Instead, an agent is expected to pay attention to (rather than ignore) a signal if the average *benefit* of paying attention to that signal exceeds the aver-

age *cost* of paying attention. Thus, if the long-term benefits of paying attention to a signal when it is honest (e.g., when a cooperator signals that they are a cooperator) exceed the long-term costs that one pays for occasionally being deceived, then agents should pay attention to that signal. Similarly, an honest signaler should not stop signaling to others simply because the signal is sometimes exploited. After all, a signal that allows a cooperator-type to successfully coordinate in collective actions with other cooperator-types could be very beneficial to both signaler and recipient (see `Skyrms, 1996), possibly outweighing the costs of occasional exploitation. In short, an agent is expected to display a particular signal (rather than repress it) if the long-term benefits of displaying the signal exceeds, on average, the long-term costs associated with displaying the signal. We argue that most human populations resemble a meta-stable population structure in which fair-minded and unfair-minded individuals co-exist in a meta-stable equilibrium where individuals predisposed to advertise their preferences for fairness achieve approximately the same payoffs (benefits) as individuals who are predisposed to advertise their preferences for unfair, but self-benefiting, outcomes. How can this be?

A key assumption that we make in the current chapter is that most human populations correspond to a relatively stable distribution of strategy types where the long-term benefits of signaling exceed the long-term costs of occasionally being exploited. This may especially be the case, we contend, when signals correspond to advertisements for strategic commitments. Perhaps the only individuals whom we expect to never signal would be those ruthless non-cooperative agents, such as competitor types, who routinely maximize their immediate self-interest, never adjusting their strategy as a function of the type of strategist with whom they are interacting. Consistent with this view, Lomborg's (1996) simulations suggest that so long as the population has a critical amount of punishers and if punishers interact frequently enough with cooperators, a population structure that includes a small number of ruthlessly non-cooperative strategists can be relatively stable over evolutionary time. This is the case, game-theoretically speaking, because certain population structures (e.g., a ratio of 4:2:1 of cooperators, punishers, and non-cooperators) may generate a meta-stable dynamic equilibrium in which the small number of non-signaling ruthless non-cooperators cannot achieve a higher payoff by increasing their representation in the population. This is the case we propose, because both the fair-minded always signaling cooperators and the never-signaling ruthless non-cooperators are in frequency dependent population structure where each strategy type cannot do better if it becomes more or less prevalent in the population.[11]

CONCLUSIONS

Is our uniquely human sense of justice an evolutionary adaptation designed to promote fairness? Or, instead, is our capacity for justice simply a byproduct of an evolved psychology designed to promote something else? In this chapter we argued that the human capacity for fairness is a byproduct of evolved emotional commitment devices that functioned—in ancestral environments—to promote evolutionary fitness rather than fairness per se. We reviewed evidence from political ideology, experimental economics, and social motives demonstrating individual differences in the importance that people attach to fairness. We argued that individual differences in the propensity to experience certain emotions might explain this variability in how much individuals value fairness. Finally, we used a game theoretic framework to demonstrate how different emotion-based strategic commitments could interact to produce a stable population structure consisting of both fair-minded cooperators and less than fair-minded noncooperative strategists.

Are there distinct emotional commitment devices for each type of justice? In other words, do we possess one set of emotions to insure that we are treated fairly when scarce resources are allocated (distributive justice emotions), another set of emotions insure that we act fairly when we allocate rewards and punishments to others (retributive justice emotions), and yet another set of emotions to insure that the methods and procedures that we employ to allocate resources are themselves fair and just (procedural justice emotions)? We think not. Instead, we propose that the emotions that underlie the psychology of justice evolved not to promote justice, but rather to promote the fitness of the individuals who possessed these emotions. This is a subtle, but important, distinction. In the same sense that one can argue that "selfish genes" can give rise to cooperative (i.e., nonselfish) individuals (Dawkins, 1976; Sober & Wilson, 1998), we argue that certain emotions which evolved because they promoted the genetic self-interests of the individuals who possessed them, can actually give rise to a multitude of different strategy types ranging from the ruthlessly self-interested individuals to apparently altruistic individuals who are emotionally predisposed to "play fair." This sort of argument should not seem extraordinary to justice researchers familiar with Adam Smiths notion of the "invisible hand." Natural selection for self-interested moral sentiments may have generated, we argue, a paradoxical outcome akin to Adam Smith's invisible hand whereby the human capacity for fairness is based upon emotional commitment devices that exist, not because they function to promote fairness, but rather because they have functioned—in ancestral environments—to promote the evolutionary self-interests of the individuals who possessed them.

NOTES

1. A *game* corresponds to any social interaction in which participants' outcomes are dependent upon not only their own behavior (what they elect to do), but also the behavior of their interaction partner (what their opponent elects to do).
2. An ultimatum game refers to a two-person scenario in which one player (the proposer) is given a divisible resource and is asked to offer some portion of that resource to the second player (Guth, Schmittberger, & Schwarz, 1982). The second player (the responder) is informed of the offer and can "take it" or "leave it." If the responder accepts the offer, the resource is divided as proposed. If the respondent rejects the offer, neither party receives anything. In a dictator game the second player is not given a choice of accepting or rejecting the first players offer.
3. In a common pool resource (CPR) game multiple players are allowed to make limited anonymous withdrawals from a collectively shared resource. Whatever portion of the shared resource remains is then replenished at some rate (i.e., increased by some multiplier) and distributed equally among the players. In voluntary contribution (VC) games, players begin with an initial monetary endowment and are then prompted to make simultaneous anonymous contributions of any portion of their endowment to a common group fund. Whatever money is in the common fund after all players have had the opportunity to contribute is then doubled (or increased by some multiplier) and distributed equally among the players.
4. There is a slight tendency for the percentage of cooperative strategists in the adult population to increase slightly with age (see Van Lange, Otten, DeBruin, & Joireman, 1997).
5. The core of such "commitment problems" centers on the fact that the psychological reward mechanism produces a representation of one's circumstances that displays the short-term benefits right now (see Frank, 1988 pp. 76–80).
6. *Antabuse* is a drug used to support the treatment of chronic alcoholism by producing an acute sensitivity to alcohol. After consuming alcohol a patient who has previously consumed antabuse will experience the effects of a severe hangover (nausea and vomiting) for a period of 30 minutes up to several hours.
7. We do not mean to imply that an agent will necessarily be consciously aware of the strategy that they are playing. Conscious awareness of the ultimate motives behind a behavior are not a requirement for the behavior to be deemed "strategic."
8. Retributive justice (punishment) is theoretically interesting because game theorists have argued that the act of punishing a non-cooperator can itself be considered a second-order public goods game in the sense that when there is an actual cost associated with administering punishment (i.e., opportunity costs, strategic costs, etc.) there is an inherent conflict of interest between the desire to deter low contributors (non-cooperators) and the desire to free-ride on the costly acts of deterrence administered by other group members (Yamagishi, 1986).

9. *Schadenfreude* is a German term that refers to the experience of pleasure at another's misfortune.
10. The true smile, or "Duchenne smile," is named after the French anatomist Duchenne de Boulogne who was among the first to observe that authentic smiles of enjoyment –as opposed to insincere smiles or authentic displays of embarrassment—are indicated by contractions of two muscles: (1) the orbicularis oculi which surrounds the eye and (2) the zygomaticus major which turns up the corners of the mouth (Ekman, Davidson, & Friesen, 1990).
11. Ruthless non-cooperative strategies benefit more from interacting with cooperator typess than they do from interacting with other ruthless non-cooperators, thus if ruthless non-cooperators increase drastically in number, they effectively lower their own average payoff per interaction. Similarly, if the number of ruthless non-cooperators becomes too low, the benefit of being a free-riding cooperator type (who never punishes) increase relative to the benefits of being a punisher type and the number of cooperator (never punishing) types increases. This in turn, creates an environment where ruthless non-cooperators can thrive, therefore setting into motion a dynamic process whereby the benefits of being a non-punishing cooperator begin to decrease relative to the benefits of being a punisher type. In this manner, the relative proportion of cooperators, punishers, and non-cooperators could wax and wane in a meta-stable, frequency dependent population structure.

REFERENCES

Au. W. T., & Kwong, J. Y. Y. (2002). Measurements and effects of social value orientation in social dilemmas: A review. In R. D. Suleieman, D. V. Budescu, I. Fischer, & D. M. Messick (Eds.), *Contemporary psychological research on social dilemmas* (pp. 71–98). London: Cambridge University Press.

Axelrod, R. (1984). *The evolution of cooperation.* New York, NY: Basic Books.

Bar-Hillel, M., & Yaari, M. (1993). Judgments of distributive justice. In B. A. Mellers and J. Baron (Eds.), *Psychological perspectives on justice: Theory and applications* (pp. 56–84). New York: Cambridge University Press.

Camerer, C. F. (2003). *Behavioral game theory: Experiments on strategic interaction.* Princeton, NJ: Princeton University Press.

Camerer, C. F., & Fehr, E. (2004). Measuring social norms and preferences using experimental games: A guide for social scientists. In J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, & H. Gintis. (Eds) *Foundations of Human Sociality: Experimental and Ethnographic Evidence from 15 Small-scale Societies* (pp. 55–95). Oxford University Press.

Dawkins, R. (1976). *The selfish gene.* Oxford: Oxford University Press.

De Cremer, D., & Tyler, T. R. (2005). Managing group behavior: The interplay between procedural justice, sense of self, and cooperation. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 37, pp. 151–218). New York: Academic Press.

De Jong. P. J., Peters, M., De Cremer, D., & Vranken, C. (2002). Blushing after a moral transgression in a prisoner's dilemma game: appeasing or revealing? *European Journal of Social Psychology, 32*, 627–644.

Deutsch, M. (1975). Equity, equality, and need: what determines which value will be used as the basis of distributive justice, *Journal of Social Issues, 31*(3), 137–150

Deutsch, M. (1985). *Distributive justice: A social-psychological persepctive.* New Haven, CT: Yale University Press.

DeQuervain, J. F., Fischbacker, Y. U., Treyer, V. Schellhammer, Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment, *Science, 305,* 1254–1258.

Ekman, P. Davidison, R. J., & Freisen, W.V. (1990). The Duchenne smile: emotional expression and brain physiology II, *Journal of Personality and Social Psychology, 58*, 342–353.

Fehr, E., & Gaechter, S. (2002). Altruistic punishment in humans. *Nature, 10,* 137–140.

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

Guth, W., Schmittberger, R., & Schwarz, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization, 3*, 367–388.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, (Fall), 55–66.

Haidt, J., & Graham, J. (in press). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research.*

Henrich, J., Boyd, R., Bowles, S., Camerer, C. Fehr, E., & Gintis. H. (2004) *Foundations of human sociality: Experimental and ethnographic evidence from 15 small-scale societies.* Oxford University Press.

Hirshleifer, J. (1987.) On the emotions as guarantors of threats and promises. In J. Dupré (Ed.), *The latest on the best: Essays on evolution and optimality* (pp. 307–326). Boston: Bradford Books-MIT Press.

Hirshleifer, J. (2001). Game-theoretic interpretations of commitmen. In R. Nesse (Ed). *Evolution and the capacity for commitment* (pp. 77–94). New York: Russell Sage Foundation.

Hunter, J. D. (1991). *Culture wars: The struggle to define America.* New York: Basic Books.

Ketelaar, T. (2004). Ancestral emotions, current decisions: Using evolutionary game theory to explore the role of emotions in decision-making. In C. Crawford, & C. Salmon (Eds), *Evolutionary psychology, public policy and personal decisions* (pp. 145–168). Mahwah: Erlbaum.

Ketelaar, T. (2006). The role of moral sentiments in economic decision making. In D. de Cremer, M. Zeelenberg, & K. Murnighan (Eds.), *Social psychology and economics* (pp. 97–116). Mahwah, NJ: Erlbaum.

Ketelaar, T., & Au, W. T. (2003). The effects of guilty feelings on the behavior of uncooperative individuals in repeated social bargaining games: An Affect-as-information interpretation of the role of emotion in social interaction. *Cognition & Emotion, 17,* 429–453.

Knutson, B. (2004). Sweet revenge? *Science, 305,* 1246–1247.

Lomborg, B. (1996). Nucleus and shield: The evolution of social structure in the iterated Prisoner's Dilemma. *American Sociological Review, 61*, 278–307.

Maynard Smith, J. (1982). *Evolution and the theory of games*. Cambridge: Cambridge University Press.

Messick, D. M., & McClintock, C. G. (1968). Motivational bases of choice in experiments. *Journal of Experimental Social Psychology, 4*, 1–25.

Newborn, M. 1996. *Kasparov versus Deep Blue: Computer chess comes of age*, New York: Springer-Verlag, Inc.

Rawls, J. A. (1971). *Theory of justice*. Cambridge, MA: Harvard University Press.

Roth, A. (1999). The redesign of the matching market for American physicians: Some engineering aspects of economic design. *American Economic Review, 89*, 748–778.

Schelling, T. C. (1960). *The strategy of conflict*. Cambridge, MA: Harvard University Press.

Searcy, W. A., & Nowicki, S. (2005). *The evolution of animal communication: reliability and deception in signaling systems*. Princeton: Princeton University Press.

Skyrms, B. (1996). *Evolution of the social contract*. Cambridge, MA: Cambridge University Press.

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

Sun Tzu (1963) *The Art of War* (S. B. Griffith, Trans.). Oxford: Oxford University Press.

Thibaut, J., & Walker, L. (1975). *Procedural justice: A psychological analysis*. Hillsdale, NJ: Erlbaum.

Tyler, T. R., & Boeckmann, R. J. (1997). Three strikes and you are out, but why? The psychology of public support for punishing rule breakers. *Law & Society Review,. 31*, 237–266.

Van Lange, P. A. M., & Visser, K. (1999). Locomotion in social dilemmas: How people adapt to cooperative, tit-for-tat, and noncooperative partners. *Journal of Personality and Social Psychology, 77*, 762–773.

Van Lange, P. A. M., Otten, W., DeBruin, E. M. N., & Joireman, J. A. (1997). Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. *Journal of Personality and Social Psychology, 73,* 733–746.

Von Neumann, J., & Morgenstern, O. (1944). *The theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

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