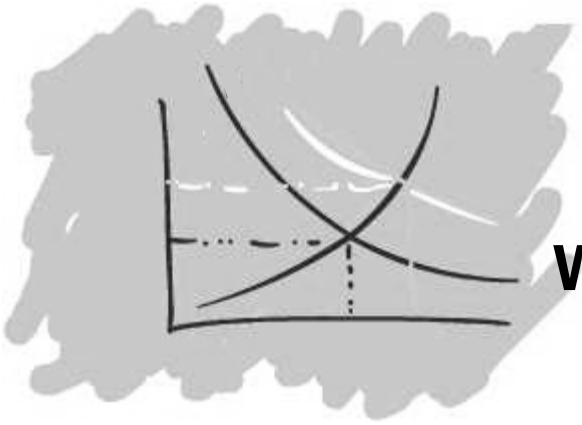
I.S.S.N: 1885-6888



ECONOMIC ANALYSIS WORKING PAPER SERIES

Emotions Enforce Fairness Norms (A Simple Model of Strong Reciprocity)

♦

Raúl López Pérez

Working Paper 11/2006

UAM
UNIVERSIDAD AUTONOMA
DE MADRID

DEPARTAMENTO DE ANÁLISIS ECONÓMICO:
TEORÍA ECONÓMICA E HISTORIA ECONÓMICA

Emotions Enforce Fairness Norms (a Simple Model of Strong Reciprocity)*

Raúl López-Pérez†

September 2005

Abstract

In experimental games, many subjects cooperate contrary to their material interest and they do that in a reciprocal manner. In addition, many subjects punish those others who behave unkindly, and previous history usually influences subjects' choices. We propose a simple game-theoretical model to account for these and other experimental phenomena, and compare it with other models of social preferences and reciprocity.

Keywords: Emotions, Fairness, Path-Dependency, Strong Reciprocity, Social Norms. JEL classification numbers: C70, C72, D63, D64, D74, Z13.

---

*I am indebted to Jordi Brandts, Marco Casari, Gary Charness, Danilo Coelho, Simon Gächter, Erin Krupka, Clara Ponsatí, Debraj Ray, Andrew Schotter, Angel Solano, Christian Traxler and seminar participants at UAB, NYU, the March 2004 Public Choice meeting and the November 2003 Urrutia Elejalde Foundation and UNED Winter Workshop on Economics and Philosophy for very helpful comments. Part of this research was conducted while visiting NYU and I would like to thank the Economics Department -specially to Andrew Schotter- for its great hospitality. I also gratefully acknowledge financial support from Spanish Ministry of Education and Science.

†International Doctorate in Economic Analysis, Departament d'Economia i d'Historia Econòmica, Universitat Autònoma de Barcelona, Spain. Mail address: Buganvilla 10, 6D, 28036 Madrid, Spain. E-mail: rlopez@idea.uab.es

1 Introduction

Experimental Economics offers abundant evidence -see Fehr and Schmidt (2002) and Camerer (2003) for excellent surveys- that contradicts the joint hypothesis that *all* agents are rational and motivated *only* by their own material interest. In a Dictator 'game' experiment, for instance, one subject is provisionally endowed with some amount of money and must decide how much of that money to transfer to another, anonymous, participant; the 'game' finishing then. Clearly, a rational and materially interested chooser would not transfer anything. Contrary to that prediction, a significant proportion of the participants give something, many times as much as half of the stake.

Why does giving occur? The dictator game is so simple that an argument based on rationality failures seems rather convoluted. Introspection points to different motivational forces that material interest. This paper investigates formally such motivations in order to offer a rational choice explanation of subjects' behavior in this and *many other* experiments. That is, this paper keeps the standard assumption of rationality and relaxes that of selfish, *homo economicus* preferences.

Importantly, the theory here proposed does not only seek to explain why people cooperate or behave generously towards others, but also why people punish others contrary to their material interest, as happens in Ultimatum game experiments. This game has the same structure as the dictator game except that now the second player (the 'responder') has a say and may accept or reject the first mover's proposal. The proposal is implemented if the responder accepts it, whereas both players get zero money if it is rejected. Clearly, the rejection of a strictly positive offer goes against material interest. However, actual responders usually reject offers of less than one quarter of the stake and even more. As the dictator's problem, the responder's problem is so simple that rejection cannot be the result of rationality failures but of emotional forces different that material joy.

Human societies are endowed with social norms that people internalize through the education process. As a result, they acquire certain emotional responses to others' and one's behavior. Self-conscious emotions like embarrassment, guilt or shame trigger when *oneself* deviates from an internalized norm. In turn, people feel aggressive emotions like

anger when *another player* violates a norm that one has hitherto respected. These two classes of painful sensations shape human preferences -other things constant, one prefers not to suffer them- and affect human behavior in two distinctive ways. First, people adjust their choices to avoid the activation of such negative emotions. Second, specific behavioral impulses appear associated with such sensations once they get triggered (Frijda, 1986). The action tendency of anger, for instance, is to punish the deviator.

As a result, these two classes of emotions crucially shape norm compliance and punishment in human societies. More precisely, people respect norms to avoid bad feelings (*internal punishment*), external sanctions (*external punishment*), or both, whereas, in addition, human punishment is many times driven by anger and, therefore, has an important component of revenge-seeking.

From our point of view, the generosity and the Pareto-damaging behavior that (some) subjects exhibit in the Dictator and Ultimatum games, respectively, may be explained by the theory sketched in the previous discussion. The argument is simple: (Some) subjects have internalized a specific norm of fairness or distributive justice which they take to the lab, and then their emotions make them act according to the norm and punish transgressors. Intuitively, these principled subjects also affect the behavior of the remaining, self-interested, agents which may find profitable to respect the norm if they risk being sanctioned otherwise. This paper provides a formalization of this theory.

Importantly, ours is a model of reciprocity in the sense that principled players respect the norm only if they expect sufficiently others to respect it as well, and they hurt transgressors. More than that, it model strong reciprocity (Gintis, 2000) because principled agents obey the norm and punish deviators *even* contrary to their material interest.

The norm we posit (the *E-norm*) praises to act so as to achieve a fair distribution of material welfare, *assuming that all agents respect the norm*. We also postulate that people see fairness as a combination of social efficiency and Rawlsian or maximin concerns for the worst off agent, and that social efficiency receives a relatively higher weight. This implies, for instance, that distribution (5, 0) is more fair than (1, 1).

We analyze the working of the model in different strategic settings like the Dictator game, Cournot game, Ultimatum game, Trust game, and the Best-Shot game, among

others, and show that it gives precise predictions and, indeed, is more consistent with experimental evidence than the conventional *homo economicus* model.

Although this may be debatable, we also believe that our model has several advantages over other models of social preferences and reciprocity, among which we may cite some. First, and contrary to some other models, it is a model of *path-dependent preferences* in that people do not only care about the material consequences of previous or present choices but also whether such choices constitute a deviation from an internalized norm. This appears to be largely consistent with experimental evidence -again, consult, Fehr and Schmidt (2002) or Camerer (2003). Second, and consistent with the extensive evidence provided by Charness and Rabin (2002), the model predicts two apparently contradictory phenomena: (i) Many subjects have both social efficiency and maximin concerns, and (ii) Many subjects engage in Pareto-damaging behavior to punish deviators.

Further, the model is very general: One may apply it to understand why people respect dressing norms, codes of etiquette, or communication norms like 'do not lie', which other models have difficulties to explain. Last, but not least, it is relatively simple and precise -i.e., it does not predict a multiplicity of *equilibrium strategy profiles* in many games.

The remainder of the paper is organized as follows. We devote the next section to survey some of the literature on social preferences. To distinguish the effects of prosocial and aggressive emotions on behavior, section 3 first describes prosocial preferences and the E-norm. In turn, section 4 applies this model to different experimental games, comparing theoretical predictions with experimental data. Section 5 adds aggressive emotions to the model and studies other experimental games. Throughout sections 4 and 5, we point out the differences between our approach and that of other models. Finally, section 6 proposes some possible extensions and concludes.

2 Other Models of Social Preferences and Reciprocity

However the pervasiveness of the *homo economicus* hypothesis, the idea that people have social emotions has an old history in Economics. Edgeworth (1881) proposed a simple model of altruism in which an individual's utility is a weighted sum of her and others'

material payoffs. This linear formulation is rich enough to express many ideas. To describe it in game theoretical terms, assume for simplicity that player i's material payoff at terminal node z coincides with her money earnings $x_i(z)$. Player i's utility at z is then

$$U_i(z) = x_i(z) + \sum_{k \neq i} \alpha_{ik}(z) \cdot x_k(z), \tag{1}$$

where $\alpha_{ik}(z) \in [-1, 1]$ for any i, k, and z. Of course, the *homo economicus* hypothesis entails $\alpha_{ik}(z) = 0$ for any i, k and z. On the opposite, it is said that player i is altruistic towards player k at z if $\alpha_{ik}(z) > 0$, and spiteful towards k if $\alpha_{ik}(z) < 0$.

In the simplest formulation within this linear framework, $\alpha_{ik}(z)$ is the same constant number for any k and z. This means that the sign and intensity of our sentiments or emotions towards the others do not depend on their acts, qualities, and beliefs, or on the actual or potential distributions of material payoffs.

In an important and pioneering paper, Rabin (1993) put into question the previous formulation, providing an alternative model. Rabin pointed out that the sign of our sentiments is conditional: "[...] *the same people who are altruistic to other altruistic people are also motivated to hurt those who hurt them*."[1] Moreover, he posited that the sign of our sentiments depends on our beliefs about the others' intentions.

Roughly speaking, player B's intentions are her *expectations* about the terminal distribution of material payoffs to be reached in the game. Take then any two-player game in normal form and suppose that A *believes* that B's intentions are (x_A^*, x_B^*). B's *intentions* are kind (unkind) to A if x_A^* is larger (smaller) than the equitable payoff -i.e., the average of the maximum and minimum A's payments within the set of Pareto efficient allocations that, according to A, B believes to be reachable. In a somewhat analogous way, B is kind (unkind) to A if she expects A to get a higher (lower) payoff than what B believes to be A's equitable payoff. Now, a player's utility is the sum of her expected material payoff and a reciprocity component that is bounded above and below,[2] and which Rabin uses to model conditional altruism: A's reciprocity component is positive if A treats B kindly (unkindly) when she believes that B's intentions are kind (unkind). Further, A's reciprocity

[1] Rabin (1993, p. 1281), italics in the original.

[2] Hence, the bigger the material payoffs, the less the players' behavior reflects their concern for fairness.

component collapses to zero if x_A^* is just equal to the equitable payoff.

Rabin (1993) resort to Psychological Game Theory -Geanakoplos et al. (1989)- to model the idea that beliefs about the other player's intentions affect utility, and proposes, in line with that theory, an equilibrium concept in which players' strategies are optimal given their beliefs which, moreover, turn out to be correct.[3] Nevertheless, his solution concept is problematic in sequential games where non-optimizing behavior may be prescribed out of the equilibrium path. Dufwenberg and Kirchsteiger (2004) extend Rabin's approach to n-player extensive form games and provide a solution concept that follows the logic of subgame perfect equilibrium.

It follows from Rabin's definition of the equitable payoff, the reference point that players use to judge whether intentions are kind or not, that the whole set of allocations -including those outcomes that the other player does not intend to reach- might affect one's behavior, something that is generally compatible with experimental evidence. On the other hand, two assumptions of Rabin (1993) and Dufwenberg and Kirchsteiger (2004) are largely incompatible with experimental evidence. First, the equitable payoff is independent of the opponent's expected payoffs. Thus, if (2.1, 0) and (2, 2) are the only Pareto efficient material allocations and the second player's intentions are (2, 2) -i.e., he has unkind intentions towards the first player- then the first player might be willing to hurt the second one, if possible. A second shortcoming is that dummy players, which cannot have kind or unkind intentions, are never treated kindly (or unkindly). For instance, both models predict no giving in the dictator game.[4]

Falk and Fischbacher (forthcoming) propose another extension of Rabin (1993) to extensive form games that avoids those two problems. In it, a player's utility at a terminal node f is the sum of the material payoff at f and the reciprocity utilities that she gets at all her decision nodes that precede f -players may weight such reciprocity utilities differently, thus introducing heterogeneity. The chooser's reciprocity utility at decision node n depends on her beliefs at n about the opponents' intentions. As in Rabin (1993), kind intentions trigger *ceteris paribus* reward whereas unkind intentions trigger punish-

[3]Note well that, sensibly, beliefs are fixed exogenously and are not an object of choice.

[4]Nevertheless, appendix A of Rabin (1993) extends the main model to avoid this problem.

ment. Contrary to Rabin (1993), however, intentions are kind (unkind) at n when the other player gets a lower (higher) expected material payoff than oneself's.

Another key distinction from Rabin's model is that the *intensity* with which agent A's wishes to punish or reward B depends on the whole set of outcomes that A believes that B believes to be reachable. Roughly, A's disposition to reward B lessens if A believes that, although B has kind intentions, B could have given more to A at *any* other available alternative. Conversely, A's disposition to punish B lessens if A believes that, although B has unkind intentions, B could *not* have given more to A at *any* other alternative without putting himself in a disadvantageous position -an "unreasonable sacrifice" by player B.

Models using Psychological Game Theory are based on the interesting idea that social emotions depend on the motives we attribute to others. However, they share the drawback of being rather complex. Levine (1998) improves tractability by assuming that people are concerned about the opponent's type and not about his intentions. A typical agent A's type is completely specified by a number $a_A \in (-1, 1)$, which signals whether someone is benevolent ($a_A > 0$) or malevolent ($a_B < 0$). Given this, and using the linear framework of equation (1), $\alpha_{AB}(z)$ depends positively on a_A and a_B. For example, even if player A is benevolent, she may become spiteful ($\alpha_{AB} < 0$) towards a sufficiently malevolent player B. Since the type of each player is private information, there is a possibility for signalling, that is, players' actions may reveal how benevolent (or malevolent) they are, and their opponents care about this. In that way, non-chosen moves may be as important as the moves one actually chooses, something that, as we have already remarked, is sensible and consistent with experimental evidence. One drawback of this model is that it renders a multiplicity of equilibrium *strategy profiles* in most games.

The appendix of Charness and Rabin (2002) offers another model of reciprocity. They introduce a demerit profile $d = (d_1, ..., d_n)$, where $d_j \in [0, 1]$ for all j, and nonnegative parameters λ, δ, b, k, f where $\lambda \in [0, 1]$ and $\delta \in (0, 1)$. Player i's utility function is

$$\begin{aligned} U_i &= (1-\lambda)\cdot x_i + \lambda[\delta \cdot \min(x_i, \min_{k\neq i}\{x_k + bd_k\}) \\ &\quad +(1-\delta)(x_i + \sum_{k\neq i} \max\{1 - kd_k, 0\}\cdot x_k) - f\sum_{k\neq i} d_k \cdot x_k]. \end{aligned}$$

The key aspect of these preferences is that the greater is d_k for $k \neq i$, the less weight player i places on player k's material payoff. In fact, if f and d_k are sufficiently large then player i wishes to hurt player k.

In order to model reciprocity, Charness and Rabin endogenize each demerit d_j to make it dependent on player j's strategy. Roughly speaking, they define $g_i(s_i, s_{-i}, d)$ as a correspondence selecting those values of $\lambda \in [0, 1]$ such that s_i is a best response to s_{-i} given demerits d. Each $g_i(s_i, s_{-i}, d)$ is then compared with an exogenous 'selflessness standard' λ^* -to be interpreted as the weight that a decent person *ought to* put on social welfare. The intuition is that if $max\{g \,|g \in g_i(s_i, s_{-i}, d)\} < \lambda^*$ then other players resent player i's choice. Given all this, strategy profile s is a 'reciprocal-fairness equilibrium' (RFE) if there exists a profile d and a correspondence $g_i(s, d)$ for all i such that, for all i, s_i is a best response to s_{-i} given d, and $d_i = max[\lambda^* - g_i, 0]$ -i.e., the demerit profile must be consistent with the profile of strategies.

The model of Charness and Rabin (2002) presents several drawbacks like its complexity, the existence of many free parameters, the lack of heterogeneity in players' utility functions, and the fact that it is unclear how to compute utilities if there are multiple equilibrium demerit profiles. Charness and Rabin (2002, 851) do not see their model as "[...] being primarily useful in its current form for calibrating experimental data, but rather as providing progress in conceptualizing what we observe in experiments." In this respect, and because several of their intuitions are somehow present in our model, one may see it as a tractable continuation of their research.

All above mentioned utility models are *non-consequentialistic* or non-separable (Camerer, 2003) because a player's utility at terminal node z does *not* only depend on the distribution of material payoffs at z. Other models are consequentialistic or separable. Fehr and Schmidt (1999) and Bolton and Ockenfels (2000), for instance, model inequity aversion.

Two key hypothesis characterize Fehr and Schmidt (1999) -we use equation (1) to describe them. First, $\alpha_{AB}(z)$ is a *positive* parameter if A gets a *larger material* payoff than B, that is, if $x_A(z) > x_B(z)$. Second, $\alpha_{AB}(z)$ is a *negative* parameter if A gets a *smaller material* payoff than B -in other words, agents are envious. In addition, players are heterogeneous regarding the inequity aversion parameters, and for any individual, the

envy parameter is larger than the parameter measuring advantageous inequity aversion.

Analogously, Bolton and Ockenfels (2000) posit two basic assumptions -we use again equation (1). Roughly speaking, $\alpha_{AB}(z)$ is positive (negative) if A's material payoff is above (below) the *average material* payoff at z. Note well that these two conditions hold independently of how big B's material payoff is. Bolton and Ockenfels (2000) also assume that individuals are heterogeneous.

To finish, Andreoni and Miller (2002) and Charness and Rabin (2002) also offer consequentialistic models to explain evidence coming from some experiments. Charness and Rabin (2002), for example, hypothesize that $\alpha_{AB}(z)$ is positive and, moreover, $\alpha_{AB}(z) > \alpha_{AC}(z)$ for any other player C if B happens to be the worst off agent. That is, players are altruists with Rawlsian maximin concerns.[5]

3 A Model of Normative Preferences

Material Games and Norms. Consider any extensive form game of perfect recall. Let $N = \{1, ..., n\}$ denote the set of players, and $u(z) = \{u_1(z), ..., u_n(z)\}$ the vector of players' payoffs at terminal node z. Players are rational, that is, each one seeks to maximize her own payoff given her beliefs about other players' strategy.

In addition, let $x(z) = \{x_1(z), ..., x_n(z)\}$ denote the vector of *material* payoffs at z. That is, $x_i(z)$ represents the cardinal utility that player i gets from consumption, money, and effort exerted along the history of z. Nevertheless, in lab games -our main concern here- it seems safe to simplify and assume that subjects' material welfare just coincides with earned money. Throughout the paper, hence, the terms 'monetary payment' and 'material payoff' are synonyms.

Material payoffs and payoffs are not the same thing -i.e., generally $x_i(z) \neq u_i(z)$ for any player i and node z. However, the researcher may initially have information only about the *material game*, that is, the researcher may know $x_i(z)$ for any i and z but not $u_i(z)$. We propose in what follows a theory on how to derive any $u_i(z)$ from the data

[5] When mentioning Charness and Rabin (2002) in what follows, and unless otherwise noted, we refer to their model of quasi-maximin preferences and not to the previously described reciprocity model.

contained in the material game.

Definition 1 *A norm Ψ is a nonempty correspondence $\Psi : h \rightarrow A(h)$ that applies on any information set h of any material game. Action $a \in A(h)$ is said to be consistent with norm Ψ if $a \in \Psi(h)$. Otherwise, a is a deviation from Ψ.*

One may interpret a norm as a prescription indicating how one *ought* to behave at any conceivable situation at which one may be called to move. To put it like that, a norm orders the available actions at any information set: Some are commendable and others are not. We provide below an example of an specific norm: The E-norm.

Preferences. To simplify matters, assume that the E-norm is the only norm in the society and that there exist two types of agents: Cold and warm. Cold people ignore the E-norm and just care for their material payoff. Therefore, the utility of any such player at node z is given by

$$u_i(z) = x_i(z).$$

On the contrary, warm people have internalized the E-norm and suffer a cost when violating it, to be interpreted as a painful emotion. Furthermore, the intensity of the emotion depends *inversely* on the number of transgressors. Thus, a warm deviator feels happier if every player deviates than if she is the only deviator. One can interpret these assumptions as modelling the effects of shame on preferences. In effect, in López-Pérez (2005) we provide psychological evidence and argue that a deviation from an internalized norm triggers shame and that shame intensity is strongly correlated with inferiority feelings -e.g., on how one's actions compare with others'.

To formalize this, let $R(z)$ designate the set of players that respected the norm in the history of z. Namely, $R(z)$ includes all players who made choices consistent with the norm *or* no choice at all in the history of z. Further, let $r(z)$ denote the cardinality of set $R(z)$. Given all this, a typical warm player's utility at z takes the following form:

$$u_i(z) = \begin{cases} x_i(z) & \text{if } i \in R(z). \\ x_i(z) - \gamma \cdot r(z) & \text{if } i \notin R(z);\ (\gamma > 0). \end{cases}$$

Parameter γ measures how intensely warm types have internalized the norm. The larger it is, the more pain a warm deviator feels ceteris paribus. Importantly, γ is independent of the particular deviation oneself made in the past. Although this is indeed an extreme simplification, we show throughout the paper that it is enough to replicate many *qualitative* experimental results.

The E-norm. Let h denote a typical information set, $A(h)$ denote its corresponding set of available actions, t_0 denote a typical *initial* decision node, that is, any node immediately succeeding Nature's moves -i.e., random shocks- and $X(t_0)$ denote the set of all $x(z)$ that succeed decision node t_0.

Definition 2 *Given function*

$$F_{\varepsilon\delta} = \varepsilon \cdot \sum_{i \in N} x_i - \delta(\max_{i \in N} x_i - \min_{i \in N} x_i), \tag{2}$$

*vector x^* is an (ε, δ)-**fairmax distribution** of a material game if $x^* \in \arg\max_{x \in X(t_0)} F_{\varepsilon\delta}$ for at least one node t_0. A path connecting node t_0 and one of its (ε, δ)-fairmax distributions is an (ε, δ)-**fairmax path** of the material game.*

Assuming $\varepsilon, \delta > 0$, function $F_{\varepsilon\delta}$ depends *positively* on the *social efficiency* of x -measured as the sum of monetary payoffs- and *negatively* on the degree of *inequality* embodied in x. In what follows, and given two real numbers a and b, F_{ab} designates function $F_{\varepsilon\delta}$ when $\varepsilon = a$, and $\delta = b$.

Unless otherwise noted, we normalize the efficiency parameter ε to one and keep δ strictly positive but smaller than one. Assumption $1 > \delta$ indicates that social efficiency is relatively more important than equality. To simplify the exposition, we refer in what follows to a $(1, \delta)$-fairmax distribution and a $(1, \delta)$-fairmax path as a 'fairmax distribution' and a 'fairmax path', respectively.

To apply the E-norm to any material game, start by finding all its fairmax paths.[6] Once this task has been completed, two cases are distinguished:

[6]Infinite material games may have no fairmax distribution. Suppose, for example, that t_0 is such that $X(t_0)$ consists of all vectors (x_1, x_2) such that $x_1 + x_2 = 1$, *except* $x = (1/2, 1/2)$. It is trivial then that no distribution maximizes function $F_{1\delta}$ over $X(t)$ when $\delta \neq 0$. All the material games we consider in the applications have at least one fairmax distribution. For completeness, however, one may assume that

(i) If information set h has *at least one* node on a fairmax path, the E-norm selects all actions of $A(h)$ that belong to a fairmax path.

(ii) Otherwise, the E-norm selects the whole set $A(h)$.

In other words, unless one is certain that a deviation from a fairmax path has happened, the E-norm commends any action pointing, from any node on a fairmax path, towards one of the available fairmax distributions. On the contrary, if someone knows that a deviation has occurred then the E-norm allows any available move. This latter feature is indeed extreme but it is enough to get our results and simplifies much the analysis. In López-Pérez (2005), alternative, more sophisticated norms are described.

To illustrate how to apply the E-norm, consider the *material* game at Figure 1. Note first that, since there are no random shocks, this material game has just one initial decision node -the upper one. In addition, there is clearly only one fairmax distribution, that is, (5, 5) and two terminal nodes with that associated distribution. Consequently, there are two fairmax paths. One of them consists just of action *r* whereas the other consists of actions *l* and R. This implies that the E-norm selects both actions *l* and *r* at the initial decision node, action R at player 2's information set and both actions *l'* and *r'* at the lower node -here the norm selects all available actions because this node does not belong to any fairmax path.

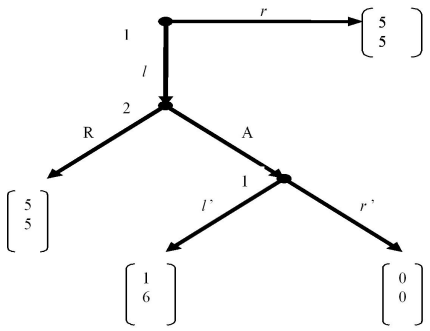


Figure 1: Two Fairmax Paths

For another example, take a two-player *material* game in which players choose simultaneously between two actions and material payoffs have the same structure as utility

the E-norm allows any move at any h of a material game with no fairmax distribution. For an alternative, consult López-Pérez (2005).

payoffs in the Prisoner's Dilemma game. Thus, if both players choose the cooperative move then distribution (b, b) ensues, each player gets a material payoff of m if they defect, and a lone cooperator (defector) gets zero (t), where $t > b > m > 0$. Trivially, (b, b) is the unique fairmax distribution and, consequently, mutual cooperation conforms the unique fairmax path. Thus, the E-norm selects the cooperative move at each player's information set. The reader is encouraged to think of other examples.

To finish, it is worthy to mention the ideas that are buried in the E-norm. First, agents see distributive justice or fairness as positively depending on social efficiency and equality. Second, if the mover at h knows that every previous mover complied then the E-norm suggests any action pointing towards a *feasible* fairmax distribution. Now, a player at h may be uncertain about previous play. In that case, the norm praises to put one's faith on any previous mover, and play *as if* one knew that every previous mover respected the E-norm. Finally, if the mover at h knows that at least one deviation has taken place then any action becomes commendable.

Players' Information. Equilibrium Concept. To explain some experimental results, it is necessary to assume that each player's type is private knowledge. That is, prior to the start of any game Nature draws players independently from a population with a binomial distribution over the set of types. Let μ denote the *objective* probability of being a warm agent.

Assuming that μ is common knowledge, we may use Perfect Bayesian Equilibrium (PBE) as a solution concept. A PBE consists of a probability assessment (beliefs) over the nodes of each player's information sets and a strategy profile. Assessments reflect what the player moving at the corresponding information set believes has happened before reaching it. They must be, to the extent possible, consistent with Bayesian updating on the hypothesis that the equilibrium strategies have been used to date. In addition, any player's strategy in a PBE must be sequentially rational. That is, everybody must choose optimally at any of her information sets given her beliefs at that set and the fact that future play will be governed by the equilibrium strategies.

In our model, importantly, beliefs do not play any practical role at information sets out of a fairmax path. That is, once a "bad" action has taken place and that becomes common

knowledge, beliefs about the type of the opponent are unimportant to explain behavior. Because of this, we will not mention such beliefs when describing a PBE. Further, this fact considerably reduces the number of *equilibrium strategy profiles* in many games, making the behavioral predictions of the model very precise.

Though most results assume common priors, it is convenient sometimes to relax this and posit heterogeneous priors.[7] Let μ_i denote the belief player i has about μ so that beliefs may be heterogeneous -i.e., $\mu_i \neq \mu_j$ for some $i \neq j$- and mistaken -i.e., $\mu_i \neq \mu$. To simplify matters, we also assume that all players *believe* (maybe incorrectly) that priors are homogeneous and correct. More formally, if player i believes μ_i then i also holds the belief that μ_i is common knowledge. McKelvey and Palfrey (1992) called a hypothesis of this sort an *Egocentric model*. This assumption is tractable and convenient because it does *not* require us to define a new solution concept. In effect, as player i believes that all players have common priors μ_i, we still predict that she will play according to a PBE of the game with common priors μ_i. To obtain behavioral predictions for any game, therefore, we will first find its PBEs as if priors were common and then we will discuss informally how belief heterogeneity affects players' behavior.

4 Explaining Experimental Evidence (I)

In this section we use the model to explain experimental evidence coming from a number of games. In addition, we will provide some tentative answers to three questions: (1) How well does the E-norm approximate the actual moral standards that some subjects take to the lab? (2) What are the factors that explain norm compliance in one-shot games if deviations cannot be punished? and (3) Do rates of norm compliance at h when (a) the opponent has made *no* choice up to h differ from those when (b) the opponent has been *active and compliant* up to h?

[7] Some experimental evidence supports this point of view. Palfrey and Rosenthal (1991), McKelvey and Palfrey (1992), Offerman et *al* (1996) and Sonnemans et al (1999) are some examples. Offerman et *al* (1996, pp. 838-839) remark that to explain their results "Not only is an altruism component needed in the utility function [...], but also an equilibrium concept which relaxes the assumption of accurate expectations".

4.1 *On Individual Decision Problems: Efficiency and Equality Matter*

It is convenient to begin by applying our model in the simplest scenario: An individual decision problem with externalities. In general terms, we predict that cold agents choose the allocation that maximizes monetary earnings whereas warm agents choose the same allocation that cold ones if their guilt parameter γ is sufficiently low and the fairmax distribution *that gives them a highest monetary payoff* otherwise.[8] To understand this latter result, the reader should recall that passive players -i.e., those who make *no* choice in the game- belong to set $R(z)$ for any z. This implies that any deviation carries an internal punishment.

Let us begin by considering the so-called Dictator game: One subject, endowed with a sum of money M, must decide how much of that money to transfer to another subject. The unique fairmax distribution is equal sharing. Hence, cold agents give nothing whereas warm ones give half of the cake if γ is larger than $\frac{M}{2}$ and nothing if γ is smaller than $\frac{M}{2}$ -they are indifferent if γ equals $\frac{M}{2}$. In other words, warm people follow their principles if that is not too costly.

Experimental results on the Dictator game -see Camerer (2003), pp.57-58, for an extensive survey- are somewhat sensitive to the degree of anonymity enjoyed by subjects when choosing, and the wording of instructions. Nonetheless, one may reasonably contend that (i) the average offer is around $0.25M$, (ii) an average of 35-40% of the participants give nothing, and (iii) there are virtually no offers above 50% of the stake. Result (iii) is clearly replicated by our model, whereas results (i) and (ii) are consistent if we assume that μ and γ take appropriate values.[9]

Our predictions depend heavily on the values of the efficiency parameter ε and the inequality parameter δ- function (2). To illustrate this, assume for a moment $\varepsilon = 1$ and $\delta = 0$. Since any monetary allocation in the dictator game is $(1, 0)$-fairmax, the model

[8] Hence, marginal changes in parameter γ may produce radical switches in warm agents' behavior. This unrealistic feature disappears if we assume that the intensity of the internal punishment conveniently depends on the particular deviation a warm agent does. We have investigated this issue in López-Pérez (2005).

[9] However, our model fails to provide an accurate picture of the actual distribution of offers, which are usually scattered along the interval $[0,\ M/2]$ and not concentrated on the extremes. See the previous footnote to this respect.

then forecasts that *no* type of player gives any money. If, on the contrary, one assumes $\varepsilon = 0$ and $\delta = 1$, equal sharing is the only $(0, 1)$-fairmax distribution and predictions coincide with those when $\varepsilon = 1$ and $0 < \delta < 1$.

Therefore, the standard dictator game does not discriminate between a formulation based on function F_{01} and the one we use throughout the paper, based on $F_{1\delta}$ for $0 < \delta < 1$. In contrast, the ingenious design of Andreoni and Miller (2002) allows for that. In their dictator game experiments, transfers of money were multiplied by a factor that differed from session to session and was common knowledge. In one session, for instance, the factor was equal to 3 so that a transfer of x units of the dictator's initial endowment translated in final earnings of $3x$ for the receiver. In this session and also when the factor was equal to 2, a significant number of dictators made transfers such that they ended up with less money than the receiver. This is consistent with our specification based on $F_{1\delta}$ for $0 < \delta < 1$ but not with one based on F_{01}. Thus, agents seem to be concerned with *both* social efficiency and equality, assigning a larger weight to the first variable.

More experimental data supports this conjecture. In Study 2, Decision 1 of Charness and Grosskopf (2001), subjects had to choose between (self, other) allocations of pesetas (625, 625) and (600, 1200).[10] Trivially, cold agents should choose the former allocation whereas warm ones choose the latter, efficient one *if* γ and δ are high and small enough, respectively, and the former, egalitarian one otherwise. Charness and Grosskopf (2001) report that only 33.3% of the subjects (N=108) chose the egalitarian allocation.

Additionally, in their Study 2, Decision 3, the *same* subjects received 600 pesetas and had to choose any payoff for another participant between 300 and 1200 pesetas. 74.1% of the subjects chose 1200 pesetas -i.e., the only fairmax distribution. Only 10.2% of them chose opponent's earnings equal to 600 pesetas -i.e., the egalitarian distribution.

Let us now compare our predictions with those from other models. For instance, Fehr and Schmidt (1999), Bolton and Ockenfels (2000) and Falk and Fischbacher (forthcoming) assume inequity averse players with no efficiency concerns. It should be intuitively

[10] At the exchange rate of the moment, one US dollar was around 150 pesetas. Each participant took decisions in three different problems but was paid for only one of those problems, chosen at random at the end of the session.

clear why those models are inconsistent with the evidence cited previously. The same occurs with Levine (1998)'s model of altruism and spitefulness, at least if we take the distribution of types that Levine posits. Rabin (1993) and Dufwenberg and Kirchsteiger (2004) are also inconsistent with the above mentioned data because they predict that no agent sacrifices her own material payoff to reward a dummy player -this is true at least for the simplest version of Rabin (1993). On the contrary, Charness and Rabin (2002) is largely consistent, thus being closest to our model.[11]

Charness and Rabin (2002) also report abundant experimental evidence that contradicts a utility model based *exclusively* on (linear) inequity aversion and material interest. In game Berk23 of Charness and Rabin (2002), for instance, subjects choose between (self, other) allocations of US dollars (2, 8) and (0, 0). Our model predicts that *all* agents choose the first allocation and this is exactly the actual result. On the contrary, Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) predict that a significant proportion of subjects choose (0, 0). Levine (1998) also predict that.

In game Barc2 of Charness and Rabin (2002), subjects choose between (self, other) allocations of pesetas (400, 400) and (375, 750). Warm players choose the second allocation if γ is high enough whereas cold agents choose the first allocation. On the opposite, inequity aversion models predict that *all* agents choose the first allocation. It turned out that 50% of the agents chose the first allocation. In the same line, Charness and Rabin (2002) show that 69% of the participants choose (Self, Other) allocations of dollars of (4, 7.5,) over (4, 4).

To sum up, the E-norm that we postulate seems a rather accurate approximation to the distributive concerns that a significant proportion of the subjects bring to the laboratory. Furthermore and since most of the remaining subjects seem to act selfishly, our model is also consistent with their behavior.

[11] Defining a fairmax distribution as an allocation maximizing function $Q(x) = \sum_{i \in N} x_i + \tau \min_{i \in N}\{x_i\}$, where $0 < \tau$, seems a more natural extension of Charness and Rabin's approach. However, such formulation gives similar results to ours, at least in two-player games.

4.2 *Norm compliance without punishment threats*

People respect norms in nonrepeated interactions even if transgressions cannot be punished and if compliance is contrary to material interest. This section explores what factors affect norm compliance in such settings. Intuitively, a warm player will obey the E-norm -and, by extension, any other norm- if two conditions hold.

First, she must have internalized the E-norm with enough intensity. In effect, warm agents suffer a psychological cost if they deviate from the E-norm. Nevertheless, if the expected material benefits of deviating are sufficiently high to overcome the expected pang -which depends, among other factors, on parameter γ- she may succumb to the temptation and deviate. Consequently norm compliance requires sufficiently *strong convictions*.

Second, she must believe that sufficiently many other players will comply as well. This follows from the fact that the psychological cost of deviating depends directly on the number of norm followers, and means that norm compliance follows a *reciprocal* logic. Let us also remark that expectations that the others will comply subtly depend on one's expectations about the other players' *types*. Believing that player B is cold suffices to infer that B will indeed deviate. On the contrary, believing that B is warm is *not* enough to sustain the belief that B will comply. We will come back to this point later.

To illustrate all those points, consider a two-player material game in which players 1 and 2 must choose simultaneously positive numbers q_1 and q_2, respectively. As a result, player i gets a monetary reward $x_i = Kq_i - q_iq_j - q_i^2$, where K is a positive number and $i, j \in \{1, 2\}, i \neq j$. As the reader may have noticed, this is a Cournot duopoly game in which firms' marginal costs take a common, constant value c and demand is linear, that is, $p = M - Q$, where M is a constant ($M > c$), p denotes the price, and Q the sum of quantities produced by each firm. In this setting, $K = M - c$.

Standard optimization techniques show that the sum of monetary rewards is maximized when $q_1 + q_2 = \frac{K}{2}$. Moreover, each player gets the same material payoff if $q_1 = q_2$. Therefore the unique fairmax distribution of this game is implemented when both players choose $q_F = \frac{K}{4}$, and that is what the E-norm commends.

When do players respect the E-norm? A first point to make in this respect is that no player will do that if she does not expect the opponent to comply as well. Basically, this

occurs because producing q_F is never in the firms' material interest -a standard textbook result shows that *if the firm is selfish* then q_F is a strictly dominated strategy. Consequently, cold firms will never produce q_F, and warm ones will only do if they expect the opponent to produce q_F as well.

Proposition 1 *A strategy profile in which both cold and warm players choose $q_{NC} = \frac{K}{3}$ is a PBE for any prior μ. This is the only PBE strategy profile in which both types deviate from $q_F = \frac{K}{4}$.*

Proof. Cold agents always seek to maximize material reward $x_i = Kq_i - q_iq_j - q_i^2$. The same is true for warm agents if the opponent deviates from the norm ($q_j \neq \frac{K}{4}$). Fixing q_j, one may show by standard optimization techniques that maximization of x_i requires $q_i = \frac{K-q_j}{2}$. Now, given the symmetry of the problem, both players should make the same choice at equilibrium. Hence, we have $q_1 = q_2 = \frac{K}{3}$. ■

Production level q_{NC} corresponds to the textbook *Nash-Cournot* prediction when both firms are self-interested -i.e., cold. Further, note that this equilibrium exists for any μ and, in particular, for $\mu = 1$, that is, in a *complete* information game played by two warm agents. Since warm types follow norms reciprocally, mutual *distrust* -i.e., the mutual expectation that the opponent will not comply- destroys any respect for the norm. Assuming then that warm types trust each others, does an equilibrium exist?

Proposition 2 *A strategy profile in which warm players choose $q_F = \frac{K}{4}$ and cold ones $q_C = \theta q_F$, where $\theta = \frac{4-\mu}{3-\mu}$, is a PBE strategy profile if $4(3-\mu)\sqrt{\mu\gamma} \geq K$.*

Proof. We show first that cold agents play a best-response. Expected utility of playing q_i is given by

$$\mu(Kq_i - q_iq_F - q_i^2) + (1-\mu)(Kq_i - q_iq_C - q_i^2) \tag{3}$$

or, alternatively, $q_i\pi - q_i^2$ where $\pi = K - \mu q_F - (1-\mu)q_C$. Differentiating $q_i\pi - q_i^2$ with respect to q_i and equating that to zero, we get as a necessary (and sufficient) condition for maximum that $q_i = \frac{\pi}{2} = q_C$.

To prove that warm agents play a best response as well, we first compute their expected utility of playing q_F. For that, and since they follow the norm and feel no remorse, it

suffices to substitute q_F for q_i at expression (3) so that one gets that expected payoff equals $q_F\pi - q_F^2 = q_F^2(2\theta - 1)$.

Suppose now that a warm agent deviates from q_F. Her expected utility is then $q_i\pi - q_i^2 - \mu\gamma$ and the best she can do is producing $q_i = \frac{\pi}{2}$, hence getting an expected payoff of $\frac{\pi^2}{4} - \mu\gamma = \theta^2 q_F^2 - \mu\gamma$. It follows that playing q_F is optimal if

$$q_F^2(2\theta - 1) \geq \theta^2 q_F^2 - \mu\gamma. \tag{4}$$

And some algebra proves inequality (4) to be equivalent to $4(3-\mu)\sqrt{\mu\gamma} \geq K$. ■

As a first remark, note that this equilibrium exists only if parameter γ is large enough. Second, we may now introduce the Egocentric model to discuss informally how heterogeneous priors affect norm compliance. Intuitively, only those warm players who have a large enough prior will follow the norm. More precisely, if i is a warm firm with priors μ_i, she might produce q_F only if $4(3-\mu_i)\sqrt{\mu_i G} \geq K$. Finally, heterogeneous priors also affect cold firms' choices. To see that, observe that because θ depends positively on μ, a cold agent i increases her choice $q_C = \theta q_F$ as her prior μ_i increases. In that way, we generate some behavioral heterogeneity from belief heterogeneity.

Experimental evidence on the Cournot game is summarized in Holt (1995). Although results are far from conclusive, they show that a significant number of participants in one-shot games attempt tacitly to collude, choosing output levels close to the joint-income maximizing level q_F, whereas remaining subjects make quantity choices around the Nash-Cournot equilibrium. Interestingly, if repetition (with rematching) is allowed, cooperation tend to vanish with time and most output decisions shift back to the Cournot level. Apparently, participants who have initially large priors tend to update their beliefs and thus move to the noncooperative equilibrium.

4.3 *Active and Passive Players*

Most modern models of reciprocity -Rabin (1993), Levine (1998), Dufwenberg and Kirschteiger (2004), and Falk and Fischbacher (forthcoming)- assume that people are, in average, more generous and willing to sacrifice own material payoff towards those that exhibited kind behavior in the past than towards passive players that did not perform any action -the reci-

procity model of Charness an Rabin (2002) is an exception because if a player does not misbehave then her demerit is zero, as a dummy player's. The E-norm, on the contrary, only allows "unkind behavior" if it is certain that the opponent deviated before. This implies that dummy players are as equally legitimated to receive a kind treatment as active norm compliers.

To illustrate the differences between our model and other reciprocity models, consider the mini-trust *material game* represented at Figure 2. The first mover (the 'investor') chooses either not to trust (move D) or to trust (move T) the second player (the 'trustee'). In the first case, the investor gets x monetary units and the trustee gets 0 units. Alternatively, the investor may trust and give the trustee the chance to repay trust (move R) or not (move A). If trust is repaid, both earn r $(> x)$ monetary units. If trust is not repaid, the investor gets the 'sucker' payoff s $(< x)$ and the trustee earns the highest payment t. To sum up, we have $s < x < r < t$. In most experiments, values are chosen so that (r, r) is the unique fairmax distribution for any $\delta < 1$. We assume that in what follows.

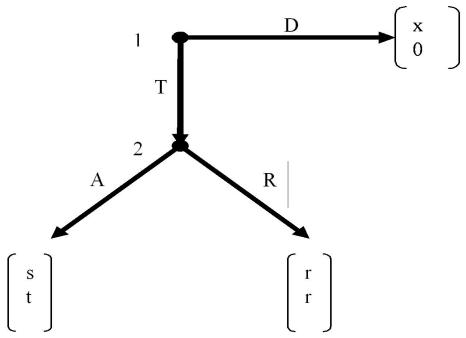


Figure 2: Mini-Trust Material Game.

Consider now two variations of this mini-trust game. In the *intentions* treatment player 1 is active and she effectively chooses her move whereas in the *random* treatment player 1 is passive and her move is decided by Nature -e.g., with the flip of a coin. Note that the unique fairmax path of the random treatment simply consists of action R -recall that a fairmax path always starts at an initial decision node and the only such node in the random treatment is the trustee's node- whereas the only fairmax path of the intentions treatment is formed by moves T and R. Suppose then that player 2 is asked to move, will he behave

differently in *each treatment*? The answer is negative.

Proposition 3 *In equilibrium and independently of the treatment, cold trustees choose A whereas warm ones choose R if γ is high enough and A otherwise.*

Proof. Cold movers go for the highest material payoff so that they play A. Since the E-norm commends trustees to move R in both treatments, warm trustees comply if the utility of playing R is larger than that of playing A, that is, $r > t - \gamma$. ■

Let us compare with other models. Consequentialistic models as the standard one (*homo economicus*), Fehr and Schmidt (1999), Bolton and Ockenfels (2000) and Charness and Rabin (2002) predict no behavioral difference between both treatments. On the contrary and as we stated above, models of intention-based reciprocity predict a significant decay in repay in the random treatment. Levine (1998) also predicts some decay because trusting in the intentions treatment signals benevolence, which is rewarded, whereas the random treatment does not allow this kind of type-selection.

Experimental evidence seems to be consistent with our model. Dufwenberg and Gneezy (2000) report data from an experimental Lost Wallet game which is very similar to our mini-trust game. The main differences are that player 2 faces a continuum of choices (more precisely, he plays a dictator game with stake size $2r$) if player 1 trusts him and that x is larger than r in some treatments (but $x < 2r$), changes that are inconsequential for our model. They compare second movers' choices with data from a pure dictator game with stake size $2r$ and do *not* reject the hypothesis that both sets of data come from the same distribution. Since a pure dictator game can be seen as a particular case of our random treatment, this is plainly consistent with proposition 3. Charness and Rabin (2002), Offerman (2002), and Cox and Deck (2005) report similar results.[12] Finally, and contrary again to reciprocity models -except that of Charness and Rabin (2002)- but consistent with ours, Dufwenberg and Gneezy (2000) also show that second movers' payback is uncorrelated with player 1's outside option x.

[12] Still, there is not a conclusive answer in this regard because some experimental papers report opposite results. Fehr and Schmidt (2002) and Camerer (2003, pp. 89-90) provide useful discussions and more references on positive reciprocity.

5 Anger and Punishment

We assume that only warm agents -i.e., those who have internalized the E-norm- display anger. This hypothesis is somewhat speculative, although there is some supporting evidence coming from Burnham (1999). In this experiment, subjects played a constrained ultimatum game where the only two offers were either \$5 or \$25 out of \$40. Moreover, subjects' testosterone levels were measured using saliva samples. Now, it is well known that high levels of testosterone are correlated with aggressive behavior. Therefore it is not surprising that subjects with high testosterone levels were relatively more likely to reject the \$5 offer. But Burnham (1999) shows something more: Subjects with high levels of testosterone were also relatively more likely to make an offer of \$25. This kind of correlations are consistent with our model.

Hence, one only needs to introduce some changes in warm agents' utility function, which is now given by

$$U_i(z) = \begin{cases} x_i(z) & \text{if } R(z) \equiv N \\ x_i(z) - \gamma \cdot r(z) & \text{if } i \notin R(z);\ (\gamma > 0) \\ x_i(z) - \alpha \max\limits_{j \notin R(z)} x_j(z) & \text{if } R(z) \subset N, i \in R(z);\ (1 \geq \alpha > 0). \end{cases}$$

Since anger goes associated with a desire to punish the deviator, we model it as history-dependent spite. Clearly, parameter α measures aggressiveness. Further, and for simplicity, anger intensity does *not* depend on the particular deviation that the deviator made, and angry agents focus at the best off deviator. Although all this seems a bit unrealistic, that does not impede the model of explaining much qualitative evidence. We maintain other assumptions of the model of prosocial behavior.[13]

[13] One may wonder whether previous results, obtained without the anger assumption, still hold. With a small caveat, the answer is positive for two reasons. First, if a deviation from the E-norm occurs in any of the games we studied then the action that maximizes the material payoff of the nondeviator also minimizes the deviator's material payoff. Hence, an angry player would make the same choices as a selfish one -note that this is no longer true for the games that we study in what follows. Second, if no deviation has taken place there is no place for anger, except as an expected emotion. Now, if someone expects the opponent to deviate then she will be less willing to respect the norm, because she expects to be angry, which is painful. To respect the norm, therefore, warm people require a larger parameter γ or a larger prior μ_i. Except for this quantitative differences, previous results still hold.

5.1 *Explaining Experimental Evidence (II): The Ultimatum Game*

In this sequential material game player 1 (the 'proposer') is provisionally allocated $M > 0$ monetary units and has to propose how to divide that money between her and player 2 (the 'responder'). Given any proposal of sharing $(x_1, M - x_1)$, the responder can either accept or reject. If he accepts, he gets $M - x_1$ and the proposer gets x_1. If he rejects, both get nothing.

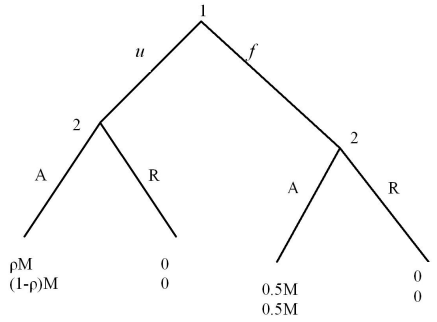


Figure 3: Mini-ultimatum material game.

Having a continuum of offers does not change our predictions. Thus, we have represented at Figure 3 a reduced version of the ultimatum material game in which player 1 has only two choices available: 'Unfair' (*u*) and 'fair' (*f*). The former choice consists of an offer of $(1-\rho)M$ monetary units to player 2, where ρ is a number in the interval $[0, 1]$. In turn, choice *f* consists of a proposal of equal sharing and it is thus consistent with the E-norm. Player 2 can either accept (A) or reject (R) -note that if player 1 offers *u* then the E-norm allows player 2 to choose both A and R, whereas if player 1 chooses *f* the E-norm only allows acceptance.

Proposition 4 *For any common priors μ and almost any ρ, the mini-ultimatum game has a unique PBE strategy profile. A warm responder always accepts the equal sharing whereas he accepts an offer of $(1-\rho)M$ if $\rho < \frac{1}{1+\alpha}$ and rejects if $\rho > \frac{1}{1+\alpha}$. A cold responder accepts any offer if $\rho > 0$. A warm proposer's choice depends on the values of γ and ρ:*

$\rho < \frac{1}{2}$ *then she offers the equal sharing.*

$\frac{1}{2} < \rho < \frac{1}{1+\alpha}$ *then she offers u if* $\gamma < \frac{M(2\rho-1)}{2}$ *and f otherwise.*

$\frac{1}{1+\alpha} < \rho$ *then she offers u if* $\gamma < \frac{M[(1-\mu)2\rho-1]}{2}$ *and f otherwise.*

Finally, a cold proposer offers u if ρ *and* μ *are large and small enough, respectively, and the equal sharing otherwise.*

Proof. Clearly, the unique fairmax distribution of this game is equal sharing. Therefore, any type of player 2 accepts offer *f*. Offering *u* makes a warm responder angry so that he accepts *u* only if

$$(1-\rho)M - \alpha\rho M > 0.$$

Trivially, a cold responder accepts *u* for any $\rho > 0$.

Consider now a cold proposer. If $\frac{1}{1+\alpha} > \rho > \frac{1}{2}$, offer *u* is always accepted and gives more money to the proposer than offer *f*. Thus, she offers *u*. For analogous reasons, she offers *f* if $\rho \leq \frac{1}{2}$. Suppose then that $\rho > \frac{1}{1+\alpha}$ so that offer *u* is not accepted by a warm responder. In that case cold proposers offer *u* if $(1-\mu)\rho M > 0.5M$, that is, if $\frac{2\rho-1}{2\rho} > \mu$, and *f* otherwise.

To finish, assume that player 1 is warm. The 50-50 offer is clearly optimal if $\rho \leq \frac{1}{2}$. Finally, offering *u* gives $\rho M - \gamma$ units of utility if $\frac{1}{1+\alpha} > \rho > \frac{1}{2}$ and $(1-\mu)\rho M - \gamma$ units of expected utility if $\rho > \frac{1}{1+\alpha}$. Simple algebraic manipulations prove that a warm proposer's strategy is optimal.[14] ■

The mini-ultimatum game, in its simplicity, shows many of the implications of our model regarding punishment. First, warm people punish -i.e., reject an offer- because they feel angry at violators of the E-norm. Second, angry responders trade off their desire for revenge and their material interest. Note that rejecting an offer costs $(1-\rho)M$, that is, the amount of the offer. As ρ decreases, the cost of punishment increases, and that explains why warm responders do not reject very large *unfair* offers. The threshold depends crucially on the aggressiveness parameter α.

[14] This proposition holds for almost any ρ. More than one equilibrium exists if $\rho = \frac{1}{1+\alpha}$ ($\rho = 0$) because warm (cold) responders are then indifferent between accepting or rejecting the unfair offer. The interested reader may easily find those equilibria.

The previous points are consistent with empirical evidence. Table 1 shows data reported in Slonim and Roth (1998) from one ultimatum game in which the stake size was 1500 Slovak Crowns (Sk), valued almost 48.5$ at the exchange rate of that moment. For instance, 32.4% of all offers were in the offer range [40- 45) -i.e., each of them was larger or equal than 40% of the stake and smaller than 45% of the stake- and 4.9% of these offers were rejected. Consistent with our prediction, low offers are frequently rejected, and the probability of rejection tends to decrease as the offer increases. This result has been replicated in many other ultimatum game experiments.[15]

With respect to the proposer's behavior, our model predicts that she will never offer more than half of the cake, which is basically consistent with experimental evidence. Moreover, the precise offer a proposer makes depends on her type, parameters α and γ, the size of the cake M, and initial priors μ. Let us consider each one separately.

The proposer's type and parameter γ largely influence her degree of norm compliance. To make this point clear, assume for a moment that warm players may differ on the degree of internalization of the E-norm so that each one is characterized by a particular γ_i. In that case, warm proposers with a sufficiently large γ_i would choose equal sharing -note that one does not need to change proposition 4 to get this result. On the other hand, cold and weak-willed warm proposers -i.e., those with a small γ_i- tend to choose meaner offers, if available.

Nevertheless, if the amount of money M at play is large enough, even a large parameter γ_i might not be enough to offset the material benefits of deviating from the fair sharing. In other words, the larger the size of the stake, the meaner (in *percentage*) the average offer. In any case, this should not be overemphasized: Experimental evidence shows that changes in stakes have small effect on proposals.[16] Maybe this is due to a strong internalization of values (high γ).

Notice that, interestingly, our model predicts a positive correlation between the offers

[15]See Camerer [2003] or Güth [1995] for evidence on this. Note that one could easily introduce more heterogeneity regarding anger parameter α. If conveniently modelled, this idea could indeed explain why offers are scattered.

[16]Notice incidentally that actual responders' thresholds change also modestly with stake changes. Evidence on those two points is surveyed in Camerer (2003, pp. 60-62).

TABLE 1

SUMMARY OF SLONIM AND ROTH (1998)

Percentages of offers and rejections by range of offers

Offer ranges	Offers	Rejections
> 50	7.2	0
= 50	30.8	1.3
(45-50)	6	0
(40-45)	32.4	4.9
(35-40)	5.2	0
(30-35)	7.2	11.1
(25-30)	3.2	37.5
< 25	8	60

that *a same agent* would make in the dictator and the ultimatum games, specially if the stake is not big. Although we are not aware of any within-subjects experiment testing this, we can at least compare ultimatum and dictator game data coming from between-subjects designs. The two most important results are that offers are less concentrated in the dictator game than in the ultimatum game, and that average offer is smaller in the dictator game. Our model is consistent with those facts and explains them because of the impossibility to punish deviators in the dictator game.

As we saw before, parameter α determines warm responders' acceptance threshold. The larger α is, the larger such threshold is. Well informed proposers who pretend to deviate from the E-norm should take this into account and, consequently, adjust their offers to their beliefs about α and μ. In fact, it is a robust experimental fact that there are almost no offers below $0.2M$. This seems to indicate that deviant proposers expect a high proportion of aggressive responders.

To illustrate the previous point more clearly, consider the model with heterogeneous and egocentric beliefs. Intuitively, we predict that cold and warm proposers make *large* offers whenever her priors μ_i surpass a certain level, which is different for cold and warm types. Now, in almost any study the vast majority of offers is in the interval $40\% - 50\%$. A

possible interpretation is that subjects come to the lab with rather large priors μ_i. In fact, the available evidence indicates that proposers tend to *overestimate* the actual proportion of people that reject unfair offers -i.e., parameter μ.[17]

5.2 *The Mini-Best-Shot Game*

In this game, player 1 moves either 'left' (*l*) or 'right' (r). Player 2 observes her move and then either 'accepts' (A) or 'rejects' (R). Figure 4 shows its *material game* tree, in which $\rho M > (1-\rho)M$ that is, $\rho > \frac{1}{2}$. A remarkable feature of this material game is that there is no Pareto efficient allocation that gives both players equal monetary payoffs. As a result, this material game has two fairmax paths: One leads to allocation $[\rho M, (1-\rho)M]$, the other one to allocation $[(1-\rho)M, \rho M]$. This largely drives our predictions.

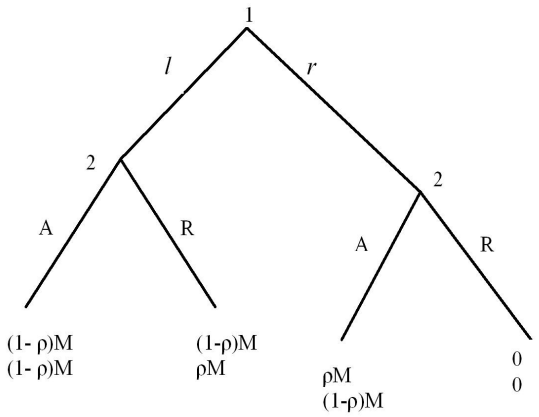


Figure 4: Mini-Best-Shot Material Game

Proposition 5 *The mini-best-shot game has a unique PBE strategy profile. Independently of their types, player 1 chooses 'right', and Player 2 rejects 'left' and accepts 'right'.*

Proof. Player 2 accepts 'right' and rejects 'left' because that is consistent with the E-norm and maximizes material payoff. For the same reasons, player 1 offers 'right'. ■

[17] See Camerer (2003, p. 56) for a brief discussion of this point and some references. A natural question is whether more experienced subjects would adjust their strategies. The answer is positive: There is some evidence that offers slightly fall over time if repetition -with rematching- is allowed. Camerer (2003, pp. 59-60) also discusses this problem.

The mini-best-shot game shows how the model works in games with more than one fairmax path. The player that chooses at a decision node in which two or more fairmax paths diverge has a strategic advantage: She can choose the fairmax path that favours her most without making the opponents angry. Models of inequity aversion, on the contrary, predict that some of the responders will not accept 'right' because the ensuing distribution is disadvantageous for the responder.

Prasnikar and Roth (1992) study a best-shot game with a richer strategy space than the one we analyze here. Their results, however, are still consistent with our predictions. This is specially true concerning proposers. On the other hand, some responders prefer $(0, 0)$ than $[\rho M, (1-\rho)M]$ -something that our model cannot explain. Does this indicate that they are inequity averse? We will deal more in detail with this issue in the following subsection. In the meanwhile, however, it is convenient to note that Prasnikar and Roth (1992) also study behavior in a comparable ultimatum game, and they show that the rate of rejection of offer $[\rho M, (1-\rho)M]$ is significatively *larger* in the ultimatum game.

The divergence in results cannot be explained by inequity aversion models because they assume consequentialistic preferences -i.e., the only thing that agents care about is how material resources are distributed, not how this distribution is achieved. Our model, on the contrary, explains the divergence because offer $[\rho M, (1-\rho)M]$ constitutes a deviation from the E-norm in the ultimatum game, where the equal sharing is feasible, but not in the best-shot game. Consequently, such offer makes the second mover angry in the ultimatum but not in the best-shot game. We will also consider more in detail this point in what follows.

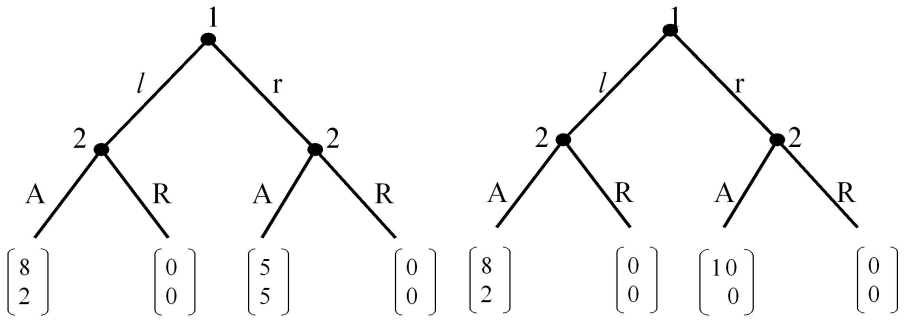
5.3 *Path-Dependent Preferences: The Choice Set Matters*

Warm player's utility function is *path-dependent* because its functional form changes with previous history, more precisely, it changes when someone deviates from an internalized norm. Hence, our model differs radically from consequentialistic models in which utility only depends on the material outcome of an interaction: One's feelings in two different games may differ *even* if the material outcome coincides.

Path-dependency is crucial to understand violence and conflict. One crucial idea in

this respect is that *the choice set matters*: An action with equal material consequences may be perfectly right in one setting but not in another in which, due to a larger choice set, the norms at work commend different behavior. To illustrate this, imagine one firm and a trade union setting wages: A wage increase that is *fair* during a recession may be completely insulting during a period of expansion -see Kahneman et al (1986) for evidence on this. As a result, workers' reactions in each case -e.g., the probability of going to strike- may differ.

As another illustration, consider the mini-ultimatum *material games* represented at Figures 5a and 5b. Player 1 can either offer 'left' (l) or 'right' (r) and player 2 can accept (A) or reject (R) any offer. In both games, 'left' consists of an offer to give eight and two monetary units to player 1 and 2, respectively. In game (5/5), 'right' is an offer to share equally ten monetary units while in game (10/0) 'right' consists of a demand of the whole cake for player 1.



(a) (5/5) Material Game. (b) (10/0) Material Game.

Figure 5:

Consider first player 2's behavior if he is offered 'left'. Falk, Fehr and Fischbacher (2003) find that 44.4% of the (8, 2) offers were rejected in game (5/5) while only 8.9% of those were rejected in game (10/0).[18] Furthermore, proposers were able to anticipate the different rates of rejection. Around 30% of the proposers offer (8,2) in game (5/5) while almost 100% of the proposers offer (8, 2) in game (10/0).

[18]Brandts and Solà (2001) report similar results. See Fehr and Schmidt (2002) for more evidence and references on the topic.

In (5/5) game, our model predicts that warm responders reject left -assuming α is high enough- and accept right. The reason is simple: Offering (8, 2) in this game is an unfair move because there exists a more fair offer -the equal sharing. Thus, offering (8, 2) activates anger and provokes rejection. Cold responders, on the other hand, accept any offer. Therefore, a cold *proposer*'s move depends on her initial expectation μ_i that the opponent is warm. She offers (8, 2) if μ_i is small enough and (5, 5) otherwise. Finally, warm proposers offer the equal split if γ is large enough.

Contrary to (5/5) game, the fairmax distribution of game (10/0) is (8, 2). Hence, offer (8, 2) is always accepted and this explains why offer (8, 2) is rejected at different rates in each game.[19] On the other hand, an offer of (10, 0) is unfair and very cheap to punish so that it is always rejected by warm responders, whereas cold ones are indifferent between accepting or rejecting it.

To sum up, the whole set of alternatives is important because people determine what is fair by looking at that set. Consequently, an action may be fair in one context but not in another one in which a more fair move is feasible. Since anger activates by unfair moves, we thus have that anger activation and punishment also depend on the initial set of alternatives.

5.4 *Path-Dependence: Responsibility Matters*

In our model, *responsibility* becomes an important variable because sanctions are directed only towards violators. Suppose, for instance, that vegetable production in a certain region has been minimal because of low irrigation, and think of two possible scenarios:

[19] As we noted before, offer (8, 2) is rejected by a small minority of responders in game (10/0) so that our model is inconsistent with that result. May this phenomenon be due to inequity aversion? To analyze this point with a bit of detail, consider an *individual decision problem* in which the chooser must decide between two (self, other) material allocations: (2, 8) and (0, 0). Note that, from a consequentialistic point of view, this is exactly the same problem that a responder faces in the (10/0) game if the fair offer is made. However, and contrary to pure inequity aversion models, some experimental evidence shows that the rate of choice of allocation (0, 0) differs significatively in both cases -basically, *no* subject chooses (0, 0) in the non-strategic setting; see Charness and Rabin (2002) on this. All this seems to indicate that inequity aversion is not the force motivating rejection of the (8, 2) offer in game (10/0).

In one, the cause of low irrigation was a heavy drought whereas in the other it was the incompetence of the agency in charge of the irrigation ditches. Although distributional consequences -low agricultural incomes- may be identical in both cases, farmers are likely to anger at the agency in the latter scenario, thus generating conflict, but not in the former one. In general, unfair outcomes may be the result of Nature moves or third parties' choices. Economic crisis in little countries, for example, may be caused by policy choices made by big countries or international institutions. Citizens' response in this case is likely to be different that if the crisis is caused by bad economic policy at the domestic level.

It is possible to test in the lab whether responsibility matters or not. Suppose, for instance, that a computer generates randomly the proposer's offer in an ultimatum game. Since the proposer is not *responsible* of any deviation, the model predicts that no responder rejects (punishes) a randomly chosen offer x. Thus lower rejections rates are predicted in the computer treatment than in the typical, intentional treatment.

Blount (1995) was the first to provide experimental evidence on this regard.[20] Our model is consistent with Blount's experimental data. Indeed, there is a significant and substantial reduction in the acceptance thresholds of responders in the computer treatment.[21] Blount also studied rejection rates in case a third party chooses the proposer's offer. Interestingly, acceptance thresholds in this condition did not differ significatively from those in the usual condition. Although this seems inconsistent with our model, we believe that it can be easily accommodated. When a third party chooses a sharing that favours either the proposer or the responder, that third party violates the norm of fairness. The responder may then 'punish' the third party by rejecting that unfair sharing -of course, rejection is more likely in case the split favours the proposer because then it is relatively cheaper.

The concept of responsibility is rather alien to consequentialistic models. Inequity aversion models as Fehr and Schmidt (1999) or Bolton and Ockenfels (2000), for instance,

[20]Blount's results are problematic because, among other reasons, subjects were deceived in one of the treatments. Further research shows, however, that Blount's qualitative results are not an artifact of the experimental design. See Fehr and Schmidt (2002) for a discussion.

[21]Nevertheless, there exists a very small proportion of actual responders that reject low offers in the random treatment. Again, see Fehr and Schmidt (2002) for a survey on this and related issues.

predict that if punishment is cheap enough, a relatively well-off agent will get punished independently that he/she is responsible of any transgression. As a result, they predict no change in rejection rates between the computer treatment and the typical, intentional treatment. On the contrary, models of intentions and type-based reciprocity predict some change.

5.5 *Punishment is not a Means to Reduce Inequity*

Models of altruistic motives like Charness and Rabin (2002) are unable to explain why people punish. On the contrary, inequity aversion models -Fehr and Schmidt (1999) and Bolton and Ockenfels (2000)- do provide a rationale: Punishment is a means to reduce disadvantageous *material* payoff inequality. For two-player games, this idea has a series of implications which we will contrast in what follows with ours.

First, inequity averse agents never punish an opponent that gets a *lower* payment than oneself. On the opposite, we predict that a warm agent will punish any transgressor (including disadvantaged ones) if punishment is cheap enough. As an illustration, consider the *material game* represented in Figure 6. Observe that, if given the choice, player 2 may punish the first mover by choosing R. Note also that player 2 gets a larger payoff than player 1 at any terminal node. Consequently, inequity aversion models predict that *any* second mover would play A if given the choice. A rational inequity averse first mover should then move l in order to get a larger material payoff and reduce disadvantageous inequity.

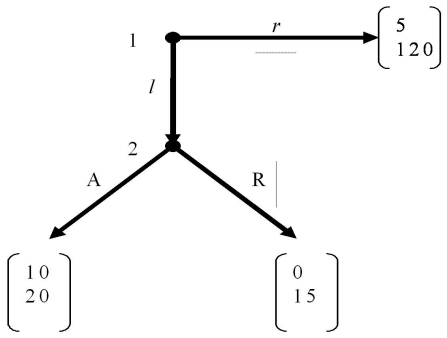


Figure 6: Punishing a disadvantaged opponent.

Unless the inequality parameter δ is very close to one, (5, 120) is the unique fairmax distribution of this material game so that the E-norm prescribes to play *r*. Suppose then that player 1 violates the norm and plays *l*. If the second mover is warm and his anger parameter α is large enough, he will punish the first mover. That is, he will play R. As a result, a rational first mover may decide to play *r* if her priors μ are large enough.

Another prediction of inequity aversion models is that punishment never takes place if it is so costly that it does not reduce disadvantageous inequity. Suppose, for example, that reducing the opponent's payment in one monetary unit costs exactly one unit as well. Then no inequity averse agent would punish the other player. On the contrary, we predict that a very aggressive warm player -i.e., $\alpha = 1$- would indeed punish a *transgressor*. Experimental evidence strongly supports our prediction -see Falk, Fehr and Fischbacher (2000) for details.

Finally, inequity aversion models predict some punishment towards an *advantaged* opponent if it is cheap enough. Since our model predicts *no* punishment towards non-transgressors, independently of their relative status, it is clearly at odds with that idea. Experimental evidence seems to be at conflict too: In previously mentioned game Berk23 of Charness and Rabin (2002) subject B chooses between (B, other) allocations of dollars (2, 8) and (0, 0). Because punishment -i.e., choosing (0, 0)- reduces inequity, inequity aversion models predict that a significant proportion of subjects choose (0, 0). Contrary to that, *all* participants chose (2, 8). Consult Fehr and Schmidt (2002) for more evidence.

6 Concluding remarks

We have shown that a large body of experimental evidence, including very different phenomena like generous and punishing behavior, may be explained by a relatively simple utility theory in which agents experience different emotional responses depending on how they and others act. Roughly speaking, our claim is that aggressive emotions like anger and moral emotions like or shame are strong psychological forces that enforce reciprocity, understanding by that concept two things: (1) People adhere to norms if they expect others to respect them as well, and (2) people punish those who violate binding norms. Further,

and since these emotions are activated by deviations from the norm, they induce path-dependent preferences.

Because it is very simple, the E-norm we propose is also too unrealistic. On one side, a norm that allows *any* move once a player transgresses it is a rather eccentric norm. Even if someone has committed a misdeed, actual norms of fairness still commend to be kind with those who previously respected them, and that may heavily restrict the set of decent choices. Another important issue is that societies have norms regulating revenge and punishment, something that the E-norm does not consider. For instance, proportionality concerns are widespread -i.e., many people believe that the punishment imposed on a deviator should be proportional to the damage that her deviation caused (*Lex Talionis*). Further, the E-norm has the problem that it is not strategic, that is, it makes prescriptions at any information set h without taking into account what the mover at h *expects* others are subsequently going to do -i.e., their intentions. Due to this, the E-norm may commend a move that, if the other player is selfish, ends up reaching a very unfair outcome.

To illustrate this, consider the *material game* tree represented at Figure 7. As (3, 3) is the only fairmax allocation of this game, the E-norm commends player 1 to move *l*. Nevertheless, a cold player 2 would then choose action A so reaching outcome (0, 4), which is most unfair. We believe that many people would argue that, unless one were sure enough that the opponent is a well-principled person who plays R, action *r* is at least as fair as action *l*. All this can be introduced in the model.

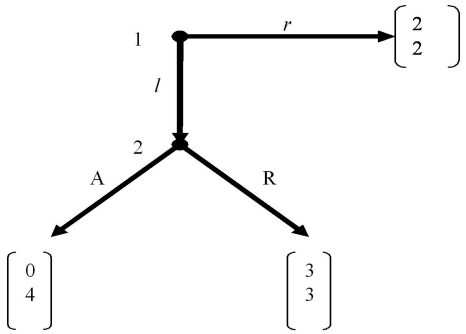


Figure 7: A Risky Move.

Assuming that the psychological cost triggered by a deviation does not depend on

the particular deviation one makes is extreme. It seems more reasonable to assume that such cost grows with the *undeserved* harm our actions impose on the others. In particular, actions leading to unfair outcomes that favour the *opponent* should not induce any remorse at all -think of an ultimatum game proposer who offers the whole cake to the responder! Further, one could add a hypothesis of nonlinearity and some heterogeneity which would be useful, for instance, to replicate the fact that dictator game offers are usually scattered along the interval $[0, M/2]$.

This paper has concentrated on the study of fairness norms -i.e., norms regulating behavior in order to reach a fair distribution of material resources. Nevertheless, the definition of norm that we have given here is general enough to include many other types of norms. Think of norms regulating dressing, eating, or communication. For instance, parents instruct their children that telling lies is, most of the cases, a bad act that should embarrass them if performed. Accordingly, most of us feel badly when breaking that rule or anger at those who violate it. This emotional responses help to enforce sincere communication, and this can be easily accommodated within the setting offered by this paper. On the contrary, other models of social preferences and reciprocity are badly suited to explain these phenomena because they define a 'bad' action -if they define it at all- only by making reference to its expected material consequences; something that, obviously, communication does not affect.

To finish, the field of application of our model should not be restricted to the lab. Indeed, social norms and emotions strongly influence human action in many 'real life' settings like voting, law compliance, bargaining, team performance, and conflict, to cite a few. The next step should go in the direction of studying such influence -Lindbeck et al. (1999) is an example of this line of research.

References

[1] Andreoni, James and John H. Miller, 2002. "Giving according to GARP: An experimental test of the Consistency of Preferences for Altruism." Econometrica, 70(2), pp. 737-53.

[2] Blount, S., 1995. “When Social Outcomes Aren’t Fair:.The Effect of Causal Attributions on Preferences.” Organizational Behavior and Human Decision Process, 63, pp. 131-144.

[3] Bolton, Gary E. and Axel Ockenfels, 2000. “ERC: A Theory of Equity, Reciprocity, and Competition.” American Economic Review, 90(1), pp. 166-93.

[4] Brandts, Jordi and Carles Solà, 2001. “Reference Points and Negative Reciprocity in Simple Sequential Games.” Games and Economic Behavior, 36, pp. 138-157.

[5] Burnham, Terence, 1999. “Testosterone and Negotiation: An Investigation into the Role of Biology in Economic Behavior.” Harvard University, JFK School of Government.

[6] Camerer, Colin F., 2003. *Behavioral Game Theory. Experiments in Strategic Interaction.* Russel Sage Foundation-Princeton University Press.

[7] Charness, Gary and Brit Grosskopf, 2001. “Relative Payoffs and Happiness: An Experimental Study.” Journal of Economic Behavior and Organization, 45, 301-328.

[8] Charness, Gary and Matthew Rabin, 2002. “Understanding Social Preferences with Simple Tests.” Quarterly Journal of Economics, 117, 817-869.

[9] Cox, James C., and Cary A. Deck, 2005. “On the nature of reciprocal motives.” Economic Inquiry, 43(3), 623-635.

[10] Dufwenberg, Martin and Uri Gneezy, 2000. “Measuring Beliefs in an Experimental Lost Wallet Game.” Games and Economic Behavior, 30, 163-182.

[11] Dufwenberg, Martin and Georg Kirchsteiger, 2004. “A Theory of Sequential Reciprocity.” Games and Economic Behavior, 47, 268-98.

[12] Edgeworth, Francis Y., 1881. *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences.* London: Kegan Paul.

[13] Elster, Jon, 1999. *Alchemies of the Mind. Rationality and the Emotions.* Cambridge University Press.

[14] Falk, Armin; Ernst Fehr and Urs Fischbacher, 2000. "Informal Sanctions." Working paper, University of Zürich.

[15] Falk, Armin; Ernst Fehr and Urs Fischbacher, 2003. "On the Nature of Fair Behavior." Economic Inquiry, 41(1), 20-26.

[16] Falk, Armin and Urs Fischbacher. "A Theory of Reciprocity." Forthcoming in Games and Economic Behavior.

[17] Fehr, Ernst and Klaus Schmidt, 1999. "A Theory of Fairness, Competition and Cooperation." Quarterly Journal of Economics, 114(3), 817-68.

[18] Fehr, Ernst and Klaus Schmidt, 2002. "Theories of Fairness and Reciprocity-Evidence and Economic Applications." in M. Dewatripont, L. Hansen and S. Turnovsky, eds., *Advances in Economics and Econometrics-8th World Congress, Econometric Society Monographs*, Cambridge, UK: Cambridge University Press.

[19] Frijda, N., 1986. *The Emotions.* Cambridge University Press.

[20] Geanakoplos, J., D. Pearce and E. Stacchetti, 1989. "Psychological Games and Sequential Rationality." Games and Economic Behavior 1, 60-79.

[21] Gintis, Herbert, 2000. "Strong Reciprocity and Human Sociality." Journal of Theoretical Biology, 206, 169-179.

[22] Güth, Werner, 1995. "On Ultimatum Bargaining Experiments–A Personal Review." Journal of Economic Behavior and Organization, 27, 329-344.

[23] Holt, Charles A., 1995. "Industrial Organization: A Survey of Laboratory Research." in John H. Kagel and Alvin E. Roth, eds., *Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.

[24] Kahneman, D., J. L. Knetsch, and R. Thaler, 1986. "Fairness as a Constraint on Profit Seeking: Entitlements in the Market." American Economic Review, 76, 728-741.

[25] Levine, David K., 1998. "Modeling Altruism and Spitefulness in Experiments." Review of Economic Dynamics, 1, 593-622.

[26] Lindbeck, A., S. Nyberg, J. W. Weibull, 1999. "Social Norms and Economic Incentives in the Welfare State." The Quarterly Journal of Economics, 114(1), 1-35.

[27] López-Pérez, Raúl, 2005. "Guilt and Shame in Games." Unpublished paper.

[28] McKelvey, Richard. D., and Thomas R. Palfrey, 1992. "An Experimental Study of the Centipede Game." Econometrica, 60, 803-836.

[29] Offerman, Theo, 2002. "Hurting hurts more than helping helps." European Economic Review 46, 1423–1437.

[30] Offerman, Theo, Joep Sonnemans, and Arthur Schram, 1996. "Value Orientations, Expectations, and Voluntary Contributions in Public Goods." The Economic Journal 106, 817-845.

[31] Palfrey, T., and H. Rosenthal, 1991. "Testing Game Theoretic Models of Free-Riding: New Evidence on Probability Bias and Learning." in T. Palfrey, ed., *Laboratory Research in Political Economy*. Ann Arbor, MI: University of Michigan Press, pp. 239-68.

[32] Prasnikar, Vesna and Alvin E. Roth, 1992. "Considerations of Fairness and Strategy: Experimental Data from Sequential Games." Quarterly Journal of Economics, 107(3), pp. 865-88.

[33] Rabin, Matthew, 1993. "Incorporating Fairness into Game Theory and Economics." American Economic Review 83, 1281-1302.

[34] Slonim, Robert and Alvin E. Roth, 1998. "Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic." Econometrica, 66, 3, 569-596.

[35] Sonnemans, Joep, Arthur Schram and Theo Offerman, 1999. "Strategic Behavior in Public Good Games: When Partners Drift Apart." Economics Letters 62, 35-41.