ELSEVIER

Journal of Economic Behavior & Organization
Vol. 65 (2008) 489–505

JOURNAL OF
Economic Behavior
& Organization

www.elsevier.com/locate/econbase

The power of conventions: A theory of social preferences

Jing Li*

Department of Economics, Duke University, 213 Social Sciences Building, Durham, NC 27705, United States

Received 7 February 2005; accepted 16 February 2006
Available online 12 October 2006

Abstract

People often act as if they care about others' welfare as well as their own (i.e. have "social preferences"). One plausible assumption is that people have preferences for social implications of their actions, determined by exogenous "conventions", in addition to the material consequences of actions. I construct games with conventions using the psychological games framework developed in Geanakoplos et al. [Geanakoplos, J., Pearce, D., Stacchetti, E., 1989. Psychological games and sequential rationality. Games and Economic Behavior 1, 60–79]. With a notion of distributional convention combining efficiency and fairness, I show equilibrium behavior reflects social preferences. The model yields tight and testable predictions consistent with a large body of experimental results, is parsimonious, and is suggestive of further studies, both experimentally and theoretically.

JEL classification: A14; D63; D64; C72

Keywords: Convention; Social preferences; Coordination; Fairness; Psychological games

1. Introduction

Social preferences refer to the phenomena that people seem to care about certain "social" goals, such as the well-being of other individuals or a "fair" allocation among members in society, in addition to their own material benefits. The evidence is ample; Camerer (2003), Kahneman and Tversky (2000) and Sobel (2005) all contain extensive accounts of both real life examples and experimental results.

* Corresponding author. Tel.: +1 919 660 1864; fax: +1 919 684 8974.
E-mail address: jingli@econ.duke.edu (J. Li).

0167-2681/$ – see front matter
doi:10.1016/j.jebo.2006.02.005

Depending on the fine details of the environment, social preferences exhibit many patterns: sometimes people reciprocate, rewarding kindness and punishing unkindness; sometimes people show unmotivated altruism; sometimes people act in the entire group's interest, even if it hurts some individuals in the group. The following experimental results are illustrative of the variety of the patterns of social preferences.

1. In an experiment of the dictator game,[1] subjects choose between pairs of (self, other) allocations. About 50 percent of the subjects choose (375, 750) over (400, 400) (Charness and Rabin, 2002).
2. Subjects first play a dictator game, choosing between (self, other) allocations of (US$ 10, US$ 10) and (US$ 18, US$ 2).[2] Then some choices were randomly selected and realized. Finally, those subjects whose decisions were not realized were given the choice of evenly splitting US$ 12 with a person whose first offer was (18, 2) or evenly splitting US$ 10 with a person whose first offer was (10, 10). The one who was not chosen for the interaction receives 0. About 74 percent of the subjects chose the latter (Kahneman et al., 1986).
3. Two players sequentially make private contributions to a public good, which is supplied either at the maximum of the two contributions (the best-shot game) or at the sum of them (the summation game). The first-mover has a smaller marginal-willingness-to-pay than the second-mover.[3] Subjects behave very differently in experiments of these two games: the first-mover typically does free ride in the best-shot game, but not in the summation game; in addition, when the first-mover contributes 0, the second-mover responds by contributing 0 almost three times more often in the summation game than in the best-shot game (Andreoni et al., 2002).
4. The ultimatum game is another famous example where theoretical prediction fails.[4] In laboratory experiments, it is rarely observed that the proposer demands the entire sum, and offers of 20–30 percent are frequently rejected. Offers of 50/50 split are observed in all experiments, often being the mode. With stakes between US$ 5 and US$ 20 and as high as US$ 100, the average offer is around 40 percent of the sum. Moreover, the rejection rate seems to depend on possible offers the proposer *did not* make. For instance, when the proposer chooses between offering 20 or 75 percent, an offer of 20 percent is rejected 33 percent of the time; however, when the proposer's choice set is changed to (20 percent, 87.5 percent), the rejection rate for an offer of 20 percent drops to 16 percent (Brandts and Sola, 2001; Camerer and Thaler, 1995; Charness and Rabin, 2002; Thaler, 1988).

It turns out to be a challenging task to explain *all* these complex patterns in a parsimonious model. The existing literature on social preferences includes two main classes of models, the distributional preferences models and the reciprocal preferences models.

Distributional preferences models assume players have preferences over final payoff *allocations*. For example, Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) explain the

[1] The dictator game is a one-player game in which the player allocates a sum of money between himself and another inactive player. Obviously, traditional game theory predicts the player takes everything himself.

[2] Consistent with the previous result, in this experiment 122 out of 161 subjects chose (10, 10) over (18, 2).

[3] Subgame perfect equilibrium predicts the same outcome in both games: the first-mover free rides and the second-mover provides the entire public good.

[4] In the ultimatum game, two players split a sum of money. The proposer moves first and makes an offer to the responder. If the responder accepts the offer then the money is divided as such. If the responder rejects it, then both players get nothing. In all subgame perfect equilibria of this game, the proposer makes an offer of no greater than the minimum share he can offer, and the responder accepts whatever she is offered.

ultimatum game results by assuming that players dislike inequality in final payoff allocations. However, these models cannot explain why players prefer an unequal payoff allocation to their own disadvantage as in Example 1. Altruism and social welfare models along the line of Andreoni (1990) and Andreoni and Miller (2002) assume players prefer a higher payoff for the opponents or the entire group of players in the game. These models can explain self-sacrificing behavior as in Examples 1 and 3, but cannot explain Pareto damaging behavior such as rejecting low offers in ultimatum games.

In fact, Example 3 clearly indicates that players' preferences over final payoff allocations alone cannot explain social preferences. When the first-mover contributes nothing, the set of payoff allocations the second-mover can generate is exactly the same in the two public-good games. Yet the second-mover makes systematically different choices. There must be something other than final payoff allocations that enters players' considerations.

In a seminal paper (Rabin, 1993), Rabin argues that it is *reciprocity* that makes the difference. Rabin assumes social preferences are driven by players' kindness towards each other: if a player believes the opponent's action is motivated by kindness toward him, he then prefers to react kindly, and vice versa. This model successfully accounts for retaliatory and altruistic behavior. Such reciprocal preferences models are intuitively appealing and further explored in Falk and Fischbacher (2006) and Dufwenberg and Kirchsteiger (2004), among others.

Rabin offers a zero-parameter model. Players strictly prefer either the kindest action or the meanest action, depending on their beliefs about the opponent. Such pure reciprocity does not explain unmotivated altruistic behavior (Example 1), or why subjects would punish, at their own cost, somebody who is mean to another person (Example 2). In addition, the simple split-the-difference fairness notion often fails to capture the context of the game. For instance, according to this fairness function, in Example 4, offering 20 percent would be strictly fairer in the first game than in the second game, if the proposer can offer 0, but then there should be fewer rejections in the first game than in the second, inconsistent with the experimental results.

One solution is to write the missing components into the utility function and adding parameters. In fact, in his appendix, Rabin suggests that to have unmotivated altruism in the model, one could add a parameter to capture the relative strength of such concerns in comparison to that of reciprocity.[5] The unsatisfactory performance of the naive fairness function could be due to its lack of an objective criterion reflecting social concerns as suggested in the distributional preferences models.

Charness and Rabin suggest a comprehensive model along this line. Their model uses six parameters to summarize how players weigh social preferences in their total preferences, how they weigh fairness in comparison to efficiency, and how they punish those opponents whom *they believe* to fail to be sufficiently concerned about other people according to some social standard. The model formalizes the most important heuristic patterns of social preferences observed in experiments. In particular, it combines distributional preferences with reciprocal preferences. The model fits important experimental results in the literature. However, with so many parameters, the model is rather unrestrictive in making interpretations and predictions, while at the same time it is too restrictive in that it does not allow for heterogeneity across players or any other forms of social preferences.

In this paper, I explore an alternative perspective to social preferences. In each of the above examples, the players seem to share some normative standpoint of what each of them "ought to" do

[5] I thank an associate editor for bringing to my attention the discussion about more general utility functions in Rabin's appendix.

given what could be done. I refer to this normative standpoint as "the convention". Players prefer to conform to the convention and prefer their opponents to conform to the convention as well. In the dictator game and the ultimatum game, it is conventional for the proposer to share the money with the opponent evenly. In the public-good game, it is conventional for players to contribute as long as they benefit from the public good and as long as their contributions count. Under common knowledge of these conventions, subjects are willing to give up some material benefits in order to conform to the convention (Experiment 1), to choose an opponent who conforms to the convention (Experiment 2 and the best-shot game in Experiment 3), and when having to interact with an opponent who has the opportunity to conform to the convention but chooses not to, to refuse to conform to the convention themselves (the summation game in Experiments 3 and 4).

A notable departure of the above story from the distributional preferences models is that players do not care about others' welfare *per se*. Rather, they only try to conform to conventions, which may incorporate some received notions about how the resources should be allocated among players. On the other hand, this explanation is also subtly different from reciprocal preferences models in that players do not care about the opponents' intentions towards *themselves*; instead, they care about the opponents' intentions towards the convention: how much the opponents conform to the convention compared to themselves.

I assume players receive payoffs from the social implications of their actions according to the convention, and the payoffs come from two additively separable components: *conformity effects*, players prefer to conform to the convention, and *interaction effects*, players prefer their opponents to conform to the convention to a degree at least as much as themselves. The total payoffs are the weighted average of the material payoffs and such social payoffs. The weight is interpreted as the salience of the convention one perceives in a game, which can be heterogeneous across players.

Intuitively, a convention induces a ranking over all possible actions in terms of their "appropriateness" or the degree of "right and wrong". The higher an action is ranked, the more desirable it is in terms of its social implications. Whether an action is appropriate depends on what appears to be the relevant context, which in turn depends on the player's feasible alternatives and beliefs about the opponent's action. Therefore, I model conventions as rankings of all actions *conditional on the player's belief* about the opponent's action. Assuming common knowledge of the convention and payoffs,[6] using the psychological games framework developed in Geanakoplos et al. (1989) (henceforth GPS), I construct games incorporating conventions for two-person normal-form games.

Conventions are exogenous in this model. They are part of the definition in the psychological game. In principle, conventions could reflect political ideal, religion, tradition and so on, and do not necessarily depend on payoffs. The same material game can be associated with different conventions, depending on the contexts of the game. In economic context, it seems the most relevant conventions are criteria regarding allocations of the payoffs. Thus, I am most interested in *distributional convention*, which is based on payoff allocations and reflects some social standard based on efficiency and fairness criteria. Social preferences emerge naturally in equilibrium of games with distributional convention. With only one parameter summarizing the (heterogeneous) attitude towards conventions across players and two parameters summarizing the distributional

[6] That the convention is common knowledge is a critical assumption for this model because for players to take into account the social implications of each other's actions, the payoffs derived from the convention have to be common knowledge. On the other hand, that the convention induces a unique ranking over one's strategy space is a simplifying assumption that rules out uncertainty of payoffs. It should be possible to relax the latter assumption, but it is less obvious when it comes to the former. For more discussions along this line, see Section 4.

convention, the model generates tight predictions consistent with a large body of experimental results. The separation of conventions and players' attitudes towards conventions makes it possible to isolate the effects of changes in conventions and heterogeneity across players on equilibrium behavior, which is suggestive for further experimental study. The general model could incorporate a wide range of social effects. In the class of games I am most interested in, namely games with distributional conventions, the model is comparable to Charness and Rabin, but with a more flexible structure, fewer parameters and heterogeneous players.

The paper proceeds as follows. Section 2 presents the basic model. Section 3 focuses on models with distributional convention. Two examples are discussed at length: the symmetric two-by-two games (including the prisoner's dilemma game, the pure coordination game and the chicken game) and the public-good games (including the summation game and the best-shot game). Section 4 concludes. Proofs not found in the text are collected in Appendix in Supplementary data.

2. A model

Fix a two-person normal form game. Let S_i and Σ_i, $i = 1, 2$, denote player i's finite set of pure strategies and set of mixed strategies, respectively. Following Rabin, I use b and c to denote the first and second-order beliefs, respectively. That is, $b_i \in \Sigma_i$ is j's belief about i's strategy, and $c_i \in \Sigma_i$ is i's belief about j's belief about i's strategy.[7] Let $\pi_i(\sigma)$ be the material payoff player i receives if strategy profile $\sigma = (\sigma_1, \sigma_2) \in \Sigma_1 \times \Sigma_2$ is played. I assume π_i is bounded.

Suppose besides the material payoffs, players also care about some exogenous convention that is applicable in the particular context of the game. To incorporate such considerations, I first expand the game to include a description of the convention players have in mind.

There are two subtleties concerning conventions. First, for any action, there is typically no *absolute* social implication attached to it. Rather, an action may be regarded as right or wrong, depending on what else one could have done.[8] For example, a doctor who recommends an expensive physical therapy is doing the right thing if there is no better treatment available, while is not if he is aware of an effective and cheap new drug but chooses not to mention it to the patient. Secondly, in strategic situations, the right thing to do depends on what the opponent is doing. For an obvious example, compare the different driving conventions in the UK and in the US. Therefore, I model the convention as belief-dependent rankings over one's strategy space, and normalize them to the unit interval.

Definition 1. Fix a two-person normal form game G. A *convention* of G is a function $\omega : \Sigma_1 \times \Sigma_2 \to [0, 1]^2$ such that, for i, $j = 1, 2$,

1. for all $b_j \in S_j$, either $\omega_i(s_i; b_j) = 1$ for all $s_i \in S_i$; or $\max_{s_i \in S_i} \omega_i(s_i; b_j) = 1$ and $\min_{s_i \in S_i} \omega_i(s_i; b_j) = 0$;
2. for all $\sigma_i \in \Sigma_i$, $b_j \in \Sigma_j$, $\omega_i(\sigma_i; b_j) = \sum_{s_i} \sum_{s_j} \sigma_i(s_i) b_j(s_j)\, \omega_i(s_i; s_j)$.

Condition 1 says ω_i ranks i's pure actions conditional on i's belief of j's action and the ranking is normalized to the unit interval, so that dependence on S_i is suppressed. The number $\omega_i(s_i; b_j)$

[7] I assume all beliefs are degenerate. See footnote 10 for a related discussion.

[8] Andreoni et al. document experimental results indicating the relevance of unchosen but available alternatives in subjects' fairness consideration. Also see Brandts and Sola (2001) and Falk et al. (2000).

is referred to as the *social index* of i taking the action s_i given his set of strategies S_i and belief b_j. For example, in the driving convention example, under the belief that the opponent drives on the left side of the road (in the UK), the action "driving at the left side" is attached a social index of 1 and the alternative "driving at the right side" is attached a social index of 0; the social indices of the actions are reversed if the player believes the opponent drives on the right side of the road (in the US). Through its dependence on the set of strategies, the convention is sensitive to the overall structure of a game, including the availability of strictly dominated strategies. In the doctor's example, suppose in addition to the effective drug, there are two equally ineffective physical treatments, A and B and A is even more expensive than B. In the absence of A, recommending B is attached a social index of 0; but if A is also an option for the doctor, then B would be attached a positive social index since A is now the 0 action. I allow for the possibility that one's entire set of pure actions is equally appropriate, which I refer to as the "trivial convention" case, and model it by giving all actions a social index of 1.[9] Condition 2 extends the specification to mixed strategies by taking expectations of the corresponding social indices.[10]

I explore the idea that in principle, one would like to conform to conventions, but such effort is greatly affected by whether others conform to conventions too. Given common knowledge of the convention, i's belief of the social index of j's action is $\omega_j(b_j; c_i)$, a function of i's first and second-order beliefs. Let $f_i(\sigma_i, b_j, c_i) = \max\{0, \omega_i(\sigma_i; b_j) - \omega_j(b_j; c_i)\}$. This function represents i's belief about how much more his own action conforms to the convention comparing to his opponent's. I consider the following utility specification:

$$u_i(\sigma_i, b_j, c_i; \omega) = \pi_i(\sigma_i, b_j) + \theta_i[g_i(\omega_i(\sigma_i; b_j)) + h_i(f_i(\sigma_i, b_j, c_i))] \tag{1}$$

where $g_i, h_i : [0, 1] \to \mathbb{R}$ are continuous, and satisfy

1. *Conformity*: g_i is increasing in ω_i;
2. *Interaction*: h_i is decreasing in f_i;
3. *Total concavity*: $g_i + h_i$ is concave in ω_i.

Eq. (1) says players get utility from both material consequences and social implications of their actions, with the total payoff being a weighted sum of the two. The scalar $\theta_i \in [0, \infty)$ represents how salient the convention is to i in the game. The larger θ_i is, the more i takes social implications of actions into account when making decisions. A player with $\theta = 0$ represents a classic agent who only cares about the material payoffs, while a player with $\theta \to \infty$ is an extremely sensitive person who only cares about the convention and ignores material consequences of actions.[11]

While θ_i captures the magnitude of social payoffs, the functions g_i and h_i capture the pattern of preferences for conventions. The function g_i reflects the *conformity effects* of conventions: taking the appropriate action makes one happy. The function h_i reflects the *interaction effects* of conventions: one prefers the opponent to conform to the conventions, and the less the opponent conforms to the conventions compared to oneself, the more disutility one receives from the

[9] Under the trivial convention, the social implications of actions do not affect players' decisions.

[10] There are some conceptual subtleties when it comes to mixed strategies in this framework. For instance, one can imagine that believing the opponent is mixing 50/50 and assigning a 50/50 chance to the opponent playing either of two pure strategies would make a difference when making a value judgment about the social implications of the opponent's action. The definition seems most appropriate if mixed strategy is interpreted as randomization over pure strategies.

[11] However, such a player may take the socially "wrong" action in equilibrium; see examples in Section 3.

interaction.[12,13] Notice that since f_i is nonincreasing in $\omega_j(b_j; c_i)$, for any G, the *associated social game* described by $g + h$ exhibits strategic complementarity in the social indices of equilibrium strategies.[14]

For any two-person normal form game G, Eq. (1) defines the induced *game with conventions* ω, denoted by (G, ω). Under Eq. (1), one's utility depends not only on material payoffs, but also on the social index of one's own action, which is belief-dependent, and one's belief about the social index of the opponent's action. Since the first and second-order beliefs enter the utility functions directly, the game is a *psychological game*. Adapting the psychological Nash equilibrium defined in GPS, I consider the following equilibrium concept.

Definition 2. A strategy profile $\sigma^* = (\sigma_1^*, \sigma_2^*) \in \Sigma$ is a *social equilibrium* in (G, ω) if, for i, $j = 1, 2$,

$$\sigma_i^* \in \arg\max_{\sigma_i \in \Sigma_i} u_i(\sigma_i, \sigma_j^*, \sigma_i^*; \omega).$$

That is, a social equilibrium in (G, ω) is a Nash equilibrium that satisfies an additional consistency condition that all beliefs correspond to actual strategies. Since the social index is linear in probability, under total concavity, standard argument delivers the existence of social equilibrium.

Theorem 1. *Social equilibrium exists for all* θ_1, θ_2.

However, the set of social equilibria in (G, ω) can differ greatly from the set of Nash equilibria in G. Obviously, when $\theta_i = 0$, $i = 1, 2$, social equilibrium reduces to Nash equilibrium. On the other hand, given the convention, not every Nash equilibrium may be approached by social equilibria.

For example, consider the game depicted in Fig. 1. This game has two pure strategy Nash equilibria: (U, L) and (D, R). Suppose the social indices are such that $\omega_1(D, L) = \omega_2(L, U) = 1$; that is, it is appropriate for player 1 to play D when player 2 plays L and for player 2 to play L when player 1 plays U. As long as $g_1(1) - g_1(0)$ and θ_1 are popsitive, (U, L) cannot be a social equilibrium. Intuitively, if a Nash equilibrium in G involves material payoff ties for some player, then social implications of the relevant actions become pivotal in (G, ω).

[12] Reciprocity modeled in Rabin and concern withdrawal modeled in Charness and Rabin reflect the idea that one's belief about how kind the opponent is affects how kind one wants to be towards the opponent, which is similar to the interaction effects modeled here. Rabin assumes beliefs determine the sign and magnitude of payoffs derived from being kind to the opponent; Charness and Rabin let beliefs affect what actions are considered kind. The interaction effects in this model are closer to Rabin's in the sense that beliefs (rather than the convention itself) affect one's incentives to conform to the convention.

Intuitively, conventions, social norms, or ethical principles are by definition commonly-agreed behavioral codes. Their strength lies in the fact that they are respected by all members in society. The interaction effects capture the intuitive idea that one would like to conform to a behavioral code only if all other members conform to it as well. For example, while most people would treat even a stranger politely, they tend to be less polite with a rude acquaintance.

[13] The "objective social standard" in Charness and Rabin refers to *the weight* people put on social payoffs, the counterpart of θ_i in this model. Players punish those opponents whose actions *seem to* suggest a θ_i lower than some θ^*, despite of the assumption of homogeneous preferences, by adjusting the rankings of actions to reflect a lack of concern for such opponents. For comparison, in this model, the convention itself is the objective standard. Players "punish" the opponents in equilibrium by not conforming to the convention themselves if the opponents' equilibrium actions are ranked low. In some sense, players in this model are "consequentialists:" they don't care about how important social payoffs are in their opponents' decisions; instead, they only care about eventually how "correct" their actions are.

[14] A game has strategic complementarity if higher action by one player implies the other players gain more from taking higher actions as well.

	L	R
U	0, 0	1, −4
D	0, 3	2, 5

Fig. 1. Nash equilibrium and social equilibrium.

In general, little can be said about which Nash equilibria are robust to the introduction of conventions, except in situations where all players have strict material interests at stake: a strict Nash equilibrium is a social equilibrium as long as the salience parameter θ_1, θ_2 are sufficiently small regardless of the convention involved.

It is perhaps of more interest to examine what happens when the convention is very salient to both players. Clearly, when θ_1, θ_2 are sufficiently large, players effectively play the associated social game described by $g + h$. By strategic complementarity, if there exists a (pure) strategy profile where both players' actions are social best responses to the opponent's action, then it is a social equilibrium regardless of the material consequences.

3. Distributional convention

There are many social principles guiding different aspects of human life. In the rest of the paper, I focus attention on the principles most relevant to economic activities, namely efficiency and fairness, where the efficiency principle applies to the total social surplus, and the fairness principle applies to the allocation of payoffs between players, dubbed distributional convention.[15]

Let $\alpha_e(s)$ index the degree of efficiency in the outcome $(\pi_1(s), \pi_2(s))$ and $\alpha_{fi}(s)$ index the degrees of i's fairness towards j in this allocation. I define the social index to be the normalization of the weighted average: $\alpha_i(s) = t\alpha_e(s) + (1-t)\alpha_{fi}(s)$, where $t \in [0, 1]$ is the weight society attaches to efficiency in evaluating the desirability of the outcomes.

The task is to define sensible measures for efficiency and fairness. Choosing an index for efficiency is less controversial. I use the following measure:

$$\alpha_e(s) = \begin{cases} \dfrac{\sum_i \pi_i(s) - \min_{s\in S} \sum_i \pi_i(s)}{\max_{s\in S} \sum_i \pi_i(s) - \min_{s\in S} \sum_i \pi_i(s)} & \text{if } \max\limits_{s\in S} \sum\limits_i \pi_i(s) \neq \min\limits_{s\in S} \sum\limits_i \pi_i(s) \\ 0 & \text{if } \max\limits_{s\in S} \sum\limits_i \pi_i(s) = \min\limits_{s\in S} \sum\limits_i \pi_i(s). \end{cases}$$

Choosing a measure for fairness, however, is more complicated. In two-person games, equity seems to be a fairly good proxy. But equity is a symmetric measure and would indicate that a person who chooses (self, other) allocations of (1, 3) over (2, 2) violates fairness, and violates it as much as choosing (3, 1). This contradicts the usual perception of being fair. It seems "fairness", in its common usage, incorporates not only the notion of equity, but also a sense of being "generous" or "humble" towards others. In the above example, one can argue that a person choosing (1, 3) over (2, 2) is unequal to himself, but is generous to the opponent, which is considered "fair". Let

$$\alpha_q(s) = \begin{cases} 1 & \text{if } \pi_i = \pi_j = 0 \\ 1 - \dfrac{|\pi_i - \pi_j|}{|\pi_i| + |\pi_j|} & \text{otherwise} \end{cases}$$

[15] Distributional convention reflects social criteria about the desirability of different payoff allocations. The specification is similar to distributional preferences as discussed in Charness and Rabin and in other distributional models. The difference is that instead of being interpreted as individual preferences, here it is interpreted as a social ranking of outcomes. Also see footnote 19 for the connection with the literature.

index the degree of equity of outcomes. It decreases in the difference between two payoffs, equals 1 whenever two players receive identical payoffs, and equals 0 whenever one player receives 0 or negative payoffs while the other player receives positive payoffs.[16] Let

$$\alpha_{gi}(s) = \begin{cases} 0 & \text{if } \pi_i = \pi_j = 0 \\ \max\left\{\frac{\pi_j - \pi_i}{|\pi_i| + |\pi_j|}, 0\right\} & \text{otherwise} \end{cases}$$

index i's generosity consideration. The index is zero whenever one grabs more than the opponent and increases by the amount the opponent leads. With this adjustment, choosing an unequal allocation to one's own disadvantage is not unfair. I define fairness as a weighted average of equity and generosity consideration: $\alpha_{fi}(s; r) = r\alpha_q(s) + (1 - r)\alpha_{gi}(s)$ where $r \in [0, 1]$. For $r = 1/2$, it is fair whenever one gives the opponent weakly more than one's own payoffs. The fairness index α_{fi} increases in π_j for all $\pi_j < \pi_i$ and reaches maximum for all $\pi_j \geq \pi_i$. For $r < 1/2$, the fairness criterion values generosity: the more one gives, the fairer one is (the fairness index increases in $\pi_j - \pi_i$). Finally, for $r > 1/2$, letting others lead is not encouraged, and the fairness index obtains the maximum at $\pi_j = \pi_i$. In this case, the fairness index reflects inequality aversion.[17,18]

The social index for i taking pure action s_i, believing j plays pure action b_j is

$$\omega_i(s_i, b_j) = \begin{cases} 1 & \text{if } \max\limits_{s_i \in S_i} \alpha_i(s_i, b_j) = \min\limits_{s_i \in S_i} \alpha_i(s_i, b_j) \\ \dfrac{\alpha_i(s_i, b_j) - \min\limits_{s_i \in S_i} \alpha_i(s_i, b_j)}{\max\limits_{s_i \in S_i} \alpha_i(s_i, b_j) - \min\limits_{s_i \in S_i} \alpha_i(s_i, b_j)} & \text{otherwise} \end{cases} \tag{2}$$

where $\alpha_i(s_i, b_j) = t\alpha_e(s_i, b_j) + (1 - t)[r\alpha_q(s_i, b_j) + (1 - r)\alpha_{gi}(s_i, b_j)]$ and $\alpha_e, \alpha_q, \alpha_{gi}$ are as defined above.[19]

The presence of strategic complementarity in preferences for distributional convention gives rise to behavior reflecting social preferences in social equilibrium. Intuitively, there are two trade-offs in games with conventions: the trade-off between material payoffs and social payoffs and the trade-off between utility derived from conformity effects and the disutilities derived from interaction effects. If conformity effects dominate, one would prefer to sacrifice the material payoffs in order to conform to distributional convention in social equilibrium, resulting in altruistic type behavior, while if interaction effects dominate, one would prefer to take a low-ranking action at the cost of material payoffs, which then is interpreted intuitively as retaliatory behavior.[20] It is easy to see that given the distributional convention, the "most" efficient and fair outcome, had it

[16] The index seems reasonable for games with nonnegative payoffs. It works less satisfactorily when there are negative payoffs involved.

[17] As long as $r > 0$, letting others lead is ranked higher than leading oneself by the same amount.

[18] The generosity consideration introduces an intrinsic asymmetry of social implications of players' actions in any action profile: $\pi_j(s) > \pi_i(s)$ implies $\alpha_{fi}(s) > \alpha_{fj}(s)$, and this difference decreases in r.

[19] The distributional convention is similar to the distributional preferences studied in Charness and Rabin; in particular, the combination of efficiency and fairness bears similarity to the combination of efficiency and maximin in their model. For different values of t and r, the distributional convention replicates some other criteria suggested in the literature, too. For example, for $t = 1$, the distributional convention reflects the social welfare criterion; for $t = 0$, $r > 1/2$, it reflects inequality aversion.

[20] Notice that the latter observation links the relative strength of conformity effects and the interaction effects in one's preference for conventions with retaliatory behavior in social equilibrium in games with distributional convention.

	C	D
C	c, c	b, a
D	a, b	d, d

Fig. 2. Symmetric 2×2 games.

existed, can be supported in social equilibrium as long as the convention is sufficiently salient for both players.[21]

I apply the model to two classes of games that have attracted much attention in both experimental and theoretical work: symmetric 2×2 games, including the prisoner's dilemma game, the coordination game and the chicken game; and public-good games, including the summation game and the best-shot game. I restrict my attention to pure strategy social equilibrium throughout.

3.1. Symmetric 2×2 *games*

Fig. 2 depicts the symmetric 2×2 game, where "C" stands for cooperate and "D" stands for defect. Let the payoff entries represent the monetary payoffs players receive, and that $a, b, c, d > 0$, $a > b$, $c > d$. Three particularly important games in this class include the prisoner's dilemma game ($b < d < c < a$), the pure coordination game ($b < d < c$, $a < c$) and the chicken game ($c < a$, $d < b$, $d < c$). I am interested in whether and under what conditions (C, C) is a (unique) social equilibrium in the induced game with distributional convention.

The strategy profile (C, C) is a Nash equilibrium in the coordination game, but not unique; it is not a Nash equilibrium in the prisoner's dilemma game and the chicken game. These predictions have been long regarded as inaccurate and counterintuitive, especially in the prisoner's dilemma game, where the unique equilibrium outcome is Pareto dominated by the cooperation outcome. In fact, in laboratory experiments of one-shot or finitely repeated prisoner's dilemma games, cooperation outcomes are frequently observed, while the ratio of cooperation outcomes typically depends on fine details of the payoff structure of the game (Cooper et al., 1996; Dawes and Thaler, 1988; Georee and Holt, 2001).

An intuitive explanation is that the game is in fact not prisoner's dilemma from the subjects' perspective due to their considerations of things other than the material payoff entries.[22] The challenge is to "expand" the game appropriately to account for the dependence of the equilibrium in the "real" game on fine details of the original game.

As a benchmark, consider the pure coordination game with distributional convention. It is easy to see that Nash equilibria in this game are robust to the introduction of distributional convention for all values of t, r and θ_1, θ_2, because the distributional convention is perfectly aligned with material interests in the game: for all t, r, $i = 1, 2$, $\omega_i(C, C) = \omega_i(D, D) = 1$. The interesting thing is that now (C, D) and (D, C) can be social equilibria, too.

For concreteness, consider the following utility function:

$$u_i(s_i, b_j, c_i; \omega) = \pi_i(s_i, b_j) + \theta_i[-(1 - \omega_i)^2 - 4 \max\{\omega_i(s_i, b_j) - \omega_j(b_j, c_i)\}]. \tag{3}$$

The quadratic loss function $-(1 - \omega_i)^2$ captures the conformity effects, and the linear loss function $-4 \max\{\omega_i(s_i, b_j) - \omega_j(b_j, c_i)\}$ captures the interaction effects. Let $c = 6$, $a = b = 0$, $d = 3$.

[21] Recall footnote 18. For a strategy profile s such that $\omega_i(s_i; s_j) = 1$ for $i = 1, 2$ to exist under distributional convention, a necessary condition is that r cannot be too small.

[22] For an inspiring discussion about fundamental modeling issues in game theory, see Weibull (2004).

Then (D, C) is a social equilibrium as long as $0 - \theta_1 \geq 6 - 4\theta_1$, $0 - \theta_2 \geq 3 - 4\theta_2$, or $\theta_1 \geq 2$ and $\theta_2 \geq 1$. Intuitively, one could supply a retaliation story for this equilibrium: 1 believes 2 believes 1 plays D, which means 2 is disregarding the convention, perhaps in order to hurt 1. If the convention is sufficiently salient to 1 ($\theta_1 \geq 2$), even though playing C is materially beneficial, the disutility from conforming to the convention given 2 does not is so much that 1 would rather play D. Similarly for 2. Since 2's material stake is smaller than 1, the required salience for 2 is also lower. Moreover, the equilibrium depends on fine material payoff details in the game: as the material stakes $a - c$ or $d - b$ increase, it is harder and harder to sustain (D, C) as a social equilibrium because the required salience values of θ_1 and θ_2 increase in $a - c$ and $d - b$, respectively. Now consider the prisoner's dilemma game and the chicken game with distributional convention. Let $c = 6$, $a = 8$, $b = 4$, $d_1 = 5$ (the prisoner's dilemma) and $d_2 = -5$ (the chicken game). Obviously, for all t, r, $\omega_i(C, C) = 1$, $i = 1, 2$ in both games. That is, it is conventional to cooperate if the opponent cooperates. It follows that in both games, (C, C) is social equilibrium provided $\theta_i \geq 2$, $i = 1, 2$. In this social equilibrium, players appear to be altruistic towards each other. Moreover, the higher a is, the harder it is to sustain (C, C) as a social equilibrium: first of all, higher a increases the material temptation $a - c$; second, for sufficiently high a, the outcome (a, b) could be so efficient that the distributional convention would rank D above C, in which case, for all θ_1, θ_2, (C, C) cannot be a social equilibrium.[23]

Whether there is other social equilibrium depends on how the distributional convention ranks C and D when the opponent plays D. Suppose efficiency matters in the distributional convention ($t > 0$) and the generosity consideration is not over-emphasized in fairness criterion ($r \geq 1/2$ is a sufficient condition). Then it is conventional to cooperate even if the opponent defects: $\omega_i(C; D) = 1$, $i = 1, 2$. Under such convention, in the prisoner's dilemma game, (D, D) is social equilibrium for all values of θ_1, θ_2. The reason is similar to that given for (D, C) being the social equilibrium in the pure coordination game: both players believe the opponent does not conform to the convention, and hence given the interaction effects in social payoffs, it is beneficial to defect not only in terms of material payoffs, but also in terms of social payoffs. On the other hand, (C, D) or (D, C) can never be social equilibrium. This is because this strategy profile has a social index profile of (0, 1), where 1 does not conform to the convention but 2 does, given each other's action. But then 2 would like to deviate: given the interaction effects, 2 prefers not to conform to the convention when 1 does not (besides, it is also materially beneficial to do so).

Similarly, one can show that, for different combinations of values of θ_1, θ_2, the sets of pure strategy social equilibria in the chicken game are as follows:

	$\theta_2 \in [0, 1]$	$\theta_2 \in [1, 2]$	$\theta_2 \in [2, 3]$	$\theta_2 \in [3, \infty)$
$\theta_1 \in [0, 1]$	$\{(C, D), (D, C)\}$	$\{(C, D)\}$	$\emptyset$	$\emptyset$
$\theta_1 \in [1, 2]$	$\{(D, C)\}$	$\emptyset$	$\emptyset$	$\emptyset$
$\theta_1 \in [2, 3]$	$\emptyset$	$\emptyset$	$\{(C, C)\}$	$\{(C, C)\}$
$\theta_1 \in [3, \infty)$	$\emptyset$	$\emptyset$	$\{(C, C)\}$	$\{(C, C), (D, D)\}$

It is worth noting that when $(\theta_1, \theta_2) \in (2, \infty)^2 \setminus [3, \infty)^2$, (C, C) is the *unique* social equilibrium. To see why this is the case, notice that in this game, "punishing" an opponent who does not conform to the convention is more costly than "rewarding" an opponent who conforms to the convention ($d - b > a - c$). This creates room for C to be the dominant strategy for a suitably

[23] When $a + b > 2c$, there is a trade-off between efficiency and fairness. Suppose $c = 10$, $a = 1000$, $b = 9$. Is (10,10) necessarily a better outcome than (1000,9)? It depends on how society evaluates efficiency versus equity, namely, the value of $t/((1 - t)r)$. For details, see the proofs for Section 3.1 in Appendix in Supplementary data.

convention-conscientious player, who values social payoffs enough to prefer to sacrifice $a - c$ in order to "reward" a "worthy" opponent, yet does not value it enough to be willing to give up $d - b$ to "punish" an "unworthy" opponent.

On the other hand, retaliatory behavior in equilibrium depends on the relative strength of conformity effects and interaction effects in players' social payoffs. For example, consider

$$u_i(s_i, b_j, c_i; \omega) = \pi_i(s_i, b_j) + \theta_i[-2(1 - \omega_i)^2 - \max\{\omega_i(s_i, b_j) - \omega_j(b_j, c_i)\}]$$

With this utility function, i would prefer to conform to the convention even if the opponent does not conform to it. It is easy to see that now in the pure coordination game, (C, D) or (D, C) cannot be social equilibria. For all values of θ_1, θ_2, the set of social equilibria in this game is $\{(C, C), (D, D)\}$. In the prisoner's dilemma game, suppose it is always conventional to cooperate, regardless of the opponent's action, then (D, D) is social equilibrium only when the convention is not very salient for both players ($\theta_i \leq 1,\ i = 1, 2$); (C, D) could be social equilibrium if the convention is salient for 1 but not for 2 ($\theta_1 \geq 1, \theta_2 \leq 1$), similarly for (D, C). In this case, it is actually possible to achieve (C, C) as the *unique* social equilibrium even in the prisoner's dilemma game ($\theta_i > 1,\ i = 1, 2$). In the chicken game, again suppose the convention ranks C above D regardless of the opponent's action, then (D, D) is never social equilibrium, (C, D) is social equilibrium provided the convention is not very salient for 2 ($\theta_2 \leq 1$), similarly for (D, C). In this case, (C, C) is the unique social equilibrium whenever the convention is sufficiently salient for both players ($\theta_i > 1,\ i = 1, 2$).

The above analysis illustrates how equilibrium behavior depends on fine details of material payoff structure in the game, the prevailing convention (t, r), the salience of conventions for each player (θ_1, θ_2), and the relative strength of conformity effects and interaction effects in the social payoffs (g_i, h_i). The predictions are much richer and more realistic than in the traditional analysis, yet the complex behavioral patterns can be organized concisely as the following results.

Theorem 2. *In the symmetric* 2×2 *game with distributional convention, the strategy profile* (C, C) *is a social equilibrium if and only if one of the following conditions holds*:

1. $2c \geq a + b$ *and* $\theta_i \geq \frac{a-c}{g_i(1)-g_i(0)}$ *for* $i = 1, 2$;
2. $2c < a + b$, $\frac{(1-t)r}{t} > \frac{(a+b-2c)(a+b)}{(a+b-2d)(a-b)}$ *and* $\theta_i \geq \frac{a-c}{g_i(1)-g_i(0)}$ *for* $i = 1, 2$;
3. $2c < a + b$, $\frac{(1-t)r}{t} < \frac{(a+b-2c)(a+b)}{(a+b-2d)(a-b)}$ *and* $g_i(1) - g_i(0) < h_i(0) - h_i(1)$, $\theta_i \geq \frac{a-c}{g_i(0)+h_i(0)-(g_i(1)+h_i(1))}$ *for* $i = 1, 2$.

This result gives the necessary and sufficient conditions for coordination outcome to be a social equilibrium: either it is conventional to cooperate when the opponent cooperates, and such convention is sufficiently salient for both players (conditions 1 and 2); or it is conventional to defect when the opponent cooperates, and the interaction effects dominate the conformity effects in players' social payoffs (condition 3).

The second result summarizes the situations where (C, C) can be obtained as the unique social equilibrium.

Theorem 3. *In the symmetric* 2×2 *normal-form game with distributional convention, suppose* $d > b$. *Then* (C, C) *is the unique social equilibrium if and only if*

1a. $\frac{(1-t)r}{t} > \frac{(a+b-2c)(a+b)}{(a+b-2d)(a-b)}$; $\frac{(1-t)(2r-1)}{t} < \frac{|a+b-2d|(a+b)}{(2c-\min\{a+b,2d\})(a-b)}$ *if* $2c \geq a+b$ *and* $\frac{(1-t)(2r-1)}{t} < \frac{(a+b)}{(a-b)}$ *otherwise*;

1b. $g_i(1) - g_i(0) > h_i(0) - h_i(1)$ *for* $i = 1, 2$;

1c. $\theta_i > \max\left\{\frac{a-c}{g_i(1)-g_i(0)}, \frac{b-d}{g_i(0)+h_i(0)-(g_i(1)+h_i(1))}\right\}$ *for* $i = 1, 2$.

Suppose $d < b$. *Then* (C, C) *is the unique social equilibrium provided*,

2a. $\frac{(1-t)r}{t} > \frac{(a+b-2c)(a+b)}{(a+b-2d)(a-b)}$; $\frac{(1-t)(2r-1)}{t} < \frac{|a+b-2d|(a+b)}{(2c-2d)(a-b)}$ *if* $2c \geq a+b$; *and* $\frac{(1-t)(2r-1)}{t} < \frac{(a+b)}{(a-b)}$ *otherwise*;

2b. *for* $i = 1, 2$, $\theta_i \in \left(\frac{a-c}{g_i(1)-g_i(0)}, \infty\right)$ *if* $g_i(1) - g_i(0) > h_i(0) - h_i(1)$; *and* $\theta_i \in \left(\frac{a-c}{g_i(1)-g_i(0)}, \frac{b-d}{g_i(0)+h_i(0)-(g_i(1)+h_i(1))}\right)$ *otherwise*;

or,

3a. $\frac{(1-t)r}{t} > \frac{(a+b-2c)(a+b)}{(a+b-2d)(a-b)}$; $\frac{(1-t)(2r-1)}{t} > \frac{|a+b-2d|(a+b)}{(2c-2d)(a-b)}$ *if* $2c \geq a+b$; *and* $\frac{(1-t)(2r-1)}{t} > \frac{(a+b)}{(a-b)}$ *otherwise*;

3b. $g_i(1) - g_i(0) > h_i(0) - h_i(1)$ *for* $i = 1, 2$;

3c. $\theta_i \in \left(\frac{a-c}{|g_i(0)+h_i(0)-(g_i(1)+h_i(1))|}, \frac{b-d}{g_i(1)-g_i(0)}\right)$ *for* $i = 1, 2$.

Conditions 1a–c are necessary and sufficient for (C, C) to be the unique social equilibrium in symmetric 2×2 games with distributional convention satisfying the condition that it is materially beneficial to play D when the opponent plays D:[24] the distributional convention ranks C above D regardless of the opponent's action (condition 1a), the conformity effects dominate the interaction effects in players' social payoffs (condition 1b), and the convention is sufficiently salient for both players (condition 1c). Conditions 2a and b and 3a–c are sufficient conditions for (C, C) to be the unique social equilibrium in games satisfying $d < b$, for different prevailing conventions:[25] if it is conventional to play C regardless of the opponent's action (condition 2a), then for (C, C) to be the unique social equilibrium, such conventions need to be salient but not overwhelmingly so (condition 2b), while if it is conventional to play C when the opponent plays C and to play D when the opponent plays D (condition 3a), then not only does the convention need to be suitably salient (condition 3b), but also the conformity effects need to dominate the interaction effects in social payoffs (condition 3c).

Given the large amount of experimental evidence of retaliatory behavior, I am particularly interested in equilibrium behavior of players whose preferences for conventions are such that the interaction effects dominate the conformity effects.

Corollary 4. *In the symmetric* 2×2 *game with distributional convention*, *suppose* $h_i(0) - h_i(1) > g_i(1) - g_i(0)$ *for* $i = 1, 2$. *Then*,

[24] Note this includes the pure coordination games and the prisoner's dilemma games.

[25] This includes the chicken game.

1. (C, C) *is a social equilibrium if and only if*: (1) $2c \geq a + b$ *or* $\frac{(1-t)r}{t} > \frac{(a+b-2c)(a+b)}{(a+b-2d)(a-b)}$ *for* $2c < a + b$ *and* (2) $\theta_i \geq a - c$, $i = 1, 2$;
2. (C, C) *is the unique equilibrium only if*: (1) $d - b > a - c$; (2) $\frac{(1-t)r}{t} > \frac{(a+b-2c)(a+b)}{(a+b-2d)(a-b)}$; (3) $\frac{(1-t)(2r-1)}{t} < \frac{|a+b-2d|(a+b)}{(2c-2d)(a-b)}$ *if*: $2c \geq a + b$; *and* $\frac{(1-t)(2r-1)}{t} < \frac{(a+b)}{(a-b)}$ *otherwise*.

The result states that when interaction effects dominate conformity effects in both players' social payoffs, (1) the necessary and sufficient conditions for (C, C) to be a social equilibrium is that it is conventional to cooperate when the opponent cooperates and such convention is sufficiently salient for both players and (2) only in the chicken game it is possible to achieve (C, C) as a unique social equilibrium, for which a necessary condition is that it is conventional to cooperate regardless of one's beliefs about the opponent's action.

3.2. The public-good games

Two players simultaneously decide how much to contribute to the public good. If player i contributes $c_i \leq m$ to the public good C, his material payoff is $\pi_i(c_i; c_j) = m - c_i + (p_i + 1) \ln C(c_i; c_j)$. I examine two variants of the game: the summation game where $C = c_1 + c_2 + 1$, and the best-shot game where $C = \max\{c_1, c_2\} + 1$.[26,27] Let $p_1 \gg p_2$ so that $(p_1, 0)$ is the only Nash equilibrium in both games. The public-good games are extensively explored in the literature; see, for example, Harrison and Hirshleifer (1989), Prasnikar and Roth (1992), Andreoni et al. (2002), among others. It is found that typically subjects do not play the Nash equilibrium in experiments of the summation game, while they do in the best-shot game.

Intuitively, this is because in the two games, 2's free-riding behavior has different social implications. Given 1 contributes a positive amount $c_1 > 0$, the efficiency index of 2's action increases in c_2 in the interval $[0, p_1 + p_2 + 1 - c_1]$ in the summation game, while decreases in c_2 in the best-shot game. In other words, according to the efficiency criterion, when 1 contributes much to the public good, for 2, free-riding is the right thing to do in the best-shot game, but not in the summation game; in fact, it is the least appropriate action in that situation.

The distributional convention captures this subtlety. For simplicity, let $t = 1$ (i.e., efficiency is the only underlying principle of distributional convention). Under such convention, the right thing to do is to contribute as much as it takes to reach the social optimal level of the public good (i.e., $p_1 + p_2 + 1$). Thus, there exists some $c_1^* \in (0, p_1)$ such that for all $c_1 \in [c_1^*, p_1 + p_2 + 1]$, $\omega_1^S(c_1; 0) > \omega_2^S(c_1; 0) = 0$ while $\omega_1^B(c_1; 0) < \omega_2^B(c_1; 0) = 1$, where ω_i^S, ω_i^B denote the social index of i's action in the summation game and the best-shot game, respectively. Since the social index of 1's contribution increases monotonically in the interval $[0, p_1 + p_2 + 1]$ conditional on $c_2 = 0$, and 1's material payoff monotonically increases in the interval $[0, p_1]$, it is the interaction effects that could potentially upset such a social equilibrium. But for c_1 that falls in the above range, the interaction effects are only present in the summation game. Consequently, in the best-shot game, a much wider range of θ_1, θ_2 are susceptible to a social equilibrium in which 2 free rides.

[26] In the literature, the best-shot game typically refers to a dynamic public-good game where players move sequentially and the public good is supplied at the maximum of all private contributions. I deal with the normal-form version to avoid complications arising from sequential moves.

[27] The typical utility function used in this class of games is $\pi_i(c_i; c_j) = m - c_i + p_i \ln C$, where $C = c_1 + c_2$ or $C = \max\{c_1, c_2\}$. The modification is because I need a bounded function.

For a numerical example, let $t = 1$, $p_1 = 9$, $p_2 = 3$, $m = 20$, and consider the utility function (3) (see Section 3.1). The equilibrium characterizations turn out to fit the experimental results very well (for example, as in Experiment 3 from the introduction (Andreoni et al., 2002)). I summarize the findings below.

First, in the best-shot game, for all θ_1, θ_2, (9, 0) is a social equilibrium, while $(c_1, 0)$ where $4.8 \leq c_1 < 9$ is never a social equilibrium. In other words, if the prevailing convention attaches high weight to social efficiency, then it is always a social equilibrium for 2 to free ride and 1 to contribute the entire public good; moreover, in every social equilibrium where 2 free rides, 1 would never contribute less than 1/2 of his willingness-to-pay. On the other hand, 2 prefers to free ride as long as he believes 1 contributes at least 3.

To the contrary, in the summation game, only for $\theta_1 = \theta_2 = 0$, (9, 0) is a social equilibrium. In fact, if the convention is salient for 1 ($\theta_1 \geq 6$), he never contributes more than a negligible amount ($c_1 < 0.89$) in any social equilibrium in which 2 free rides. On the other hand, 2 typically does not want to free ride if he believes 1 contributes a positive amount, even if 1's contribution already exceeds 2's own marginal-willingness-to-pay. For example, given that 1 contributes 3, 2 prefers to free ride only if his salience is extremely low ($\theta_2 \leq 0.23$). Intuitively, in these social equilibria, 2's social index is 0 while 1's is positive, and hence 1 is subject to the interaction effects while 2 is subject to the conformity effects. It is possible to have such outcome as a social equilibrium only if neither player puts much weight on social payoffs (i.e., has a low salience).

In the set of social equilibria in which 2 free rides ($c_2^* = 0$), 1's equilibrium contribution as a function of the salience parameter $c_1^*(\theta_1)$ has very different properties in the two games. In the best-shot game, c_1^* increases in θ_1, while in the summation game, it decreases in θ_1. To see this, notice that given 2 free rides, the sign of the social payoffs for 1 are different in the two games: it is negative in the summation game due to the interaction effects, while it is positive in the best-shot game due to the conformity effects. In particular, in the summation game, 1 contributes less than his marginal willingness-to-pay, while in the best-shot game, 1 contributes at least his marginal willingness-to-pay.

Finally, for $\theta_1 \geq 6.7$, $\theta_2 \geq 1.97$, (0, 0) is a social equilibrium in both games. In this equilibrium, both players' actions have a social index of 0. As discussed before, a necessary condition for such a social equilibrium is that the interaction effects dominate the conformity effects in both players' social payoffs.

4. Concluding remarks

I propose a simple model to account for social preferences. The model takes "conventions" as given and hypothesizes that people prefer to conform to conventions and prefer the opponent to conform to conventions as well. Formalizing distributional convention as a belief-dependent ranking over the whole strategy space according to some combination of efficiency and fairness principles, I show that equilibrium behavior in games incorporating distributional convention reflects social preferences. For concrete examples, I show that the model makes sharp predictions in symmetric 2×2 games and public-good games that are consistent with experimental evidence.

The simplicity and parsimony of the model make it particularly appealing empirically. The separation of heterogeneous salience of convention and the convention itself, and the fact that the convention parameters are obviously experimentally manipulatable make it easy to test the model empirically. For example, in symmetric 2×2 games, holding everything else constant, the model predicts that equilibrium outcome varies with the material payoff details, for example, $a - c$; in

particular, the model predicts distinctly different behavior when the material payoff structure is that of the chicken game. In the public-good game, the model predicts different sets of social equilibria in the two games for fixed salience parameters θ_1, θ_2.

Conceptually, the model also differs from the main body of the social preferences literature in that in a sense, inter-dependent utilities are not the primitives of the model. In this model, the presence and nature of inter-dependent utilities only reflect the presence and nature of exogenous conventions. In a social equilibrium, by taking into account the actions' social implications, which are evaluated according to some convention that depends on overall payoff allocations, players act as if they care about each other. The nature and pattern of "social preferences" reflect the nature and pattern of the prevailing distributional convention in the game. By manipulating the distributional convention in a game, one could change the pattern and/or degree of such inter-dependence.[28]

The model embraces an "instrumental view" of conventions in one-shot games that is reminiscent of the repeated games or evolutionary arguments for social preferences. One can view a convention as a coordination device: it suggests an appropriate action (or actions) to each player in each and every context; given the strategic complementarity of the social payoffs, for sufficiently convention-conscientious players, a fixed point of the "social best response" correspondence is a social equilibrium, achieving the outcome the convention targets. Furthermore, the concept of conventions is reminiscent of notions such as social norms, ethics, morals and so on. The model connects social preferences to questions such as these: what social norms are sustainable in the long run? What determines the emergence and fading of particular moral principles? This suggests that a thorough understanding of social preferences must be based on a theory of evolution of social norms.

There are a number of interesting extensions one can study. Multi-person games and dynamic games are the most obvious ones. How do people respond to an environment where multiple opponents take actions with different social implications? How do people draw inferences when such inferences affect their utilities directly in dynamic games? How to evaluate the social implications of an action in these complex environments? These are open questions that invite both experimental and theoretical examinations. The current model provides a framework that highlights the additional questions that need to be answered and sheds light on possible experimental design.

In real life, people do not always share the same convention. For example, people from different cultural backgrounds may respect different ethical principles or attach different weights to even the same set of ethical principles; hence they evaluate an action's social implications differently. One could imagine that a bargaining impasse could result from each bargainer evaluating the social implications of actions using the convention most favorable to his own material benefits. Such issues could be dealt with in a tractable way using the framework presented in this paper, for example, by introducing payoff uncertainties into games with multiple conventions.

Acknowledgements

I am indebted to Bart Lipman and Larry Samuelson for invaluable guidance in conducting economics research in general and in writing this paper specifically. I thank Jim Andreoni, Pierpaolo Battigalli, Ming Li, Bill Sandholm, an associate editor and an anonymous referee for very useful comments. I also wish to thank fellow campers at the 2002 Russell Sage Behavioral Economics

[28] For example, by manipulating the framing of the game.

Camp for inspiring discussions and comments, especially Hanming Fang, David Laibson, George Lowenstein and Matthew Rabin.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jebo.2006.02.005.

References

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

Andreoni, J., Miller, J.H., 2002. Giving according to GARP: an experimental test of the consistency of preferences for altruism. Econometrica 70, 737–753.

Andreoni, J., Brown, P.M., Vesterlund, L., 2002. What produces fairness? Some experimental evidence. Games and Economic Behavior 40, 1–24.

Bolton, G., Ockenfels, A., 2000. ERC: a theory of equity, reciprocity, and competition. American Economic Review 90, 166–193.

Brandts, J., Sola, C., 2001. Reference points and negative reciprocity in simple sequential games. Games and Economic Behavior 36, 138–157.

Camerer, C., 2003. Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press, Princeton.

Camerer, C., Thaler, R.H., 1995. Anomalies: ultimatum, dictators and manners. Journal of Economic Perspectives 9 (2), 209–219.

Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Quarterly Journal of Economics 117, 817–869.

Cooper, R., DeJong, D.V., Forsythe, R., Ross, T.W., 1996. Cooperation without reputation: experimental evidence from prisoner's dilemma games. Games and Economic Behavior 12, 187–218.

Dawes, R.M., Thaler, R.H., 1988. Anomalies: cooperation. The Journal of Economic Perspectives 2 (3), 187–197.

Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games and Economic Behavior 47, 198–268.

Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games and Economic Behavior 54, 293–315.

Falk, A., Fehr, E., Fischbacher, U., 2000. Testing theories of fairness. Working paper.

Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. Quarterly Journal of Economics 114, 817–868.

Geanakoplos, J., Pearce, D., Stacchetti, E., 1989. Psychological games and sequential rationality. Games and Economic Behavior 1, 60–79.

Georee, J.K., Holt, C.A., 2001. Ten little treasures of game theory and ten intuitive contradictions. American Economic Review 91, 1402–1422.

Harrison, G.W., Hirshleifer, J., 1989. An experimental evaluation of weakest link/best shot models of public goods. Journal of Political Economy 97, 201–225.

Kahneman, D., Knetsch, J.L., Thaler, R.H., 1986. Fairness and the assumptions of economics. Journal of Business 59, 285–300.

Kahneman, D., Tversky, A. (Eds.), 2000. Choices, Values, and Frames. Cambridge University Press, Cambridge.

Prasnikar, V., Roth, A.E., 1992. Considerations of fairness and strategy: experimental data from sequential games. The Quarterly Journal of Economics 107, 865–888.

Rabin, M., 1993. Incorporating fairness into game theory and economics. American Economic Review 83, 1281–1301.

Sobel, J., 2005. Interdependent preferences and reciprocity. Journal of Economic Literature 43 (2), 392–436.

Thaler, R.H., 1988. The Winner's Curse: Paradoxes and Anomalies of Economic Life. Free Press, New York.

Weibull, J., 2004. Testing game theory. In: Huck, S. (Ed.), Advances in Understanding Strategic Behavior: Game Theory, Experiments and Bounded Rationality, Essay in Honour of Werner Güth. Palgrave MacMillan, New York, pp. 85–104.