Fairness and Reciprocity

Jonathan Levin

June 2006

Beyond the fact that laboratory play does not correspond to notions of equilibrium, many laboratory results seem strikingly at odds with the hypothesis that players act only to maximize their material payoffs.

- In the ultimatum game, responders turn down offers with positive probability and are particularly likely to turn down low offers.
- In public goods experiments where there are strong incentives to free-ride and contribute nothing, people tend to contribute non-zero amounts. This is especially true if it is possible to punish non-contributors.
- In efficiency wage experiments, agents exert effort even if they have a financial incentive not to do so, and exert more effort when paid a high wage.

How can this behavior be explained? One possibility is that players are simply confused about what is going on in the lab. Other explanations center on the idea that players are conditioned by real-world environments where there is repeated interaction and/or sanctions for behaving selfishly. Finally, a number of leading theories suggest that people care intrinsically about "fairness" — that is, either about the distribution of payoffs or about *how* the game is played. We now turn to discussing this idea.

1 Theories of Fairness: Payoff-Driven

Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) propose models in which players care about their own payoffs and also the payoffs of others. In the Fehr-Schmidt model, given a vector of material payoffs $x = (x_1, ..., x_n)$, player i's utility is:

$$u_i(x) = x_i - \alpha_i \frac{1}{n-1} \sum_{j \neq i} \max\{x_j - x_i, 0\} - \beta_i \frac{1}{n-1} \sum_{j \neq i} \max\{x_i - x_j, 0\},$$

where $0 \leq \beta_i \leq 1$ and $\alpha_i \geq \beta_i$. For just two players, this reduces to:

$$u_i(x) = x_i - \alpha_i \max\{x_j - x_i, 0\} - \beta_i \max\{x_i - x_j, 0\},$$

The basic properties of the model are:

1. Players don't like inequality ($\alpha_i, \beta_i \geq 0$). Holding a player's own material payoff fixed, he likes everyone to get the same amount.
2. Players dislike being behind more than being ahead ($\alpha_i \geq \beta_i$).
3. Players aren't willing to throw money away to reduce inequality ($\beta_i \leq 1$), but they might give away a dollar away to reduce inequality (e.g. if $n = 2$ and $\beta_i > 1/2$).

Given these preferences, we can still use Nash or subgame perfect equilibrium as our solution concept.

Example 1: The Ultimatum Game. In the ultimatum game, player 1 proposes how to split a dollar. Player 2 either accepts the split or rejects it. In the latter case, both get zero.

Claim In the unique subgame equilibrium, the proposer offers 1/2 if $\beta_1 > 1/2$ and offers $(1+\alpha_2)/(1+2\alpha_2)$ if $\beta_1 < 1/2$. The responder accepts.

If player 1 offers a split $(s, 1-s)$, player 2's best response is:

$$\begin{aligned} \text{If } s &< 1/2\text{: Accept if } 1 - s - \beta_2(1-2s) \geq 0 \\ \text{If } s &= 1/2\text{: Accept} \\ \text{If } s &> 1/2\text{: Accept if } 1 - s - \alpha_2(2s-1) \geq 0 \end{aligned}$$

Player 1 does better to offer $s = 1/2$ rather than any $s < 1/2$. For offers $s > 1/2$, player 2 will accept if:

$$s \leq s^* = \frac{1+\alpha_2}{1+2\alpha_2}.$$

Since payoffs are linear, player 1 should either offer s^* or 1/2. The former gives a higher direct payoff, but generates inequality. The equal split is better if and only if:

$$\frac{1}{2} \geq s^* - \beta_1(2s^* - 1) \qquad \Leftrightarrow \qquad \beta_1 > 1/2.$$

Thus if $\beta_1 < 1/2$, player 1 offers $s^*, 1 - s^*$ and if $\beta_1 > 1/2$, player 1 offers $1/2, 1/2$. Either way, player 2 accepts.

Example 2: The Market Game. In the Roth et. al. market game, players $1, ..., n-1$ make proposals $s_i, 1 - s_i$. Player n, the responder, then can accept or reject the lowest offer s^L. If he accepts, he gets $1 - s^L$ and the winning proposer gets s^L. If several proposers make the low offer, we randomly select one.

Claim In the unique subgame perfect equilibrium, at least two proposers offer $s = 0$ and the responder accepts.

First note that the responder would certainly accept a low offer $s^L \leq 1/2$, since then:

$$\begin{aligned} u(1 - s^L, s^L, 0, ..., 0) &= (1 - s^L) - \beta\frac{n-2}{n-1}(1 - s^L) - \beta\frac{1}{n-1}(1 - 2s^L) \\ &= (1 - s^L)(1 - \beta) + \beta\frac{1}{n-1}s^L > 0. \end{aligned}$$

Now, if the lowest offer was greater than 1/2, a losing proposer would get at most zero (if the offer were to be rejected), but could get positive utility by offering 1/2. So in equilibrium, we must have $s^L \leq 1/2$ and the offer accepted. Moreover, in equilibrium at least two proposers must offer s^L — otherwise the winning proposer would deviate to a slightly higher offer.

If the lowest offer is $s^L > 0$, a losing proposer will have utility:

$$u(0, s^L, 1 - s^L, 0, ..., 0) = -\alpha\frac{1}{n-1}.$$

By offering just below s^L, he could have:

$$u(s^L, 1 - s^L, 0, ..., 0) = s^L - \alpha\frac{1}{n-1}(1 - 2s^L) - \beta\frac{n-2}{n-1}s^L > -\alpha\frac{1}{n-1}.$$

Since proposers prefer to win at just below s^L rather than lose at s^L, competition implies that $s^L = 0$ in equilibrium.

Fehr and Schmidt show the theory can also explain lab findings in some other games such as public good games (see the assignment). When it comes to applying the theory outside the lab, however, there are some tough conceptual issues that await further research:

1. How should one define the relevant reference group? Should it include just a person's closest peers or a broader sampling?
2. How should one define material payoffs? Are "payoffs" total wealth or just changes in wealth? Should endowments matter in a definition of fair outcomes?

2 Theories of Fairness: Intention-Driven

Rabin (1993) proposes an alternative model of fairness motivations in which players care about both material payoffs and about the *intentions* of other players. People want to reward those who are nice to them, and hurt those who are mean to them. Rabin's players end up playing what Geanakoplos, Pearce and Stacchetti (1989) call a "psychological game" — one where payoffs depend on actions and on beliefs about actions.

Rabin develops his theory for two player games, with action sets A_1, A_2. Starting with the material payoffs $\pi_i : A_1 \times A_2 \to \mathbb{R}$, Rabin defines "fairness" functions. Suppose player i believes player j is choosing b_j. Then by choosing a_i, player i is choosing a payoff pair from the set:

$$\Pi(b_j) = \left\{\pi_i(a_i, b_j), \pi_j(a_i, b_j) :, \pi_j) : a_i \in A_i\right\}.$$

Let $\pi_j^h(b_j), \pi_j^l(b_j)$ be the be player j's highest and lowest payoff among points that are Pareto-efficient in $\Pi(b_j)$. Let:

$$\pi^e(b_j) = \frac{1}{2}\left(\pi_j^h(b_j) + \pi_j^l(b_j)\right)$$

be the equitable payoff, and let $\pi_j^{\min}(b_j)$ be j's minimum payoff in $\Pi(b_j)$. Player i's kindness to player j in choosing a_i is then:

$$f_i(a, b_j) = \frac{\pi_j(a, b_j) - \pi_j^e(a, b_j)}{\pi_j^h(a, b_j) - \pi_j^{\min}(a, b_j)}$$

Now, suppose player i believes that player j believes that i will play c_i. Let $\tilde{f}_j(c_i, b_i) = f_j(c_i, b_j)$ be player i's *belief* about how kind player j is being to him, given that i believes j will play b_i. Then player i will choose his action a_i to maximize:

$$U_i(a_i, b_j, c_i) = \pi_i(a_i, b_j) + \tilde{f}_j(c_i, b_i)\left[1 + f_i(a_i, b_j)\right].$$

Definition 1 *A pair of strategies* (a_1, a_2) *is a* ***fairness equilibrium*** *if, for* $i = 1, 2$ *and* $j \neq i$,

$$a_i \in \arg\max_{a \in S_i} U_i(a, b_j, c_i)$$

with $c_i = b_i = a_i$.

Now consider some examples of fairness equilibria.

Example 1 Battle of the Sexes. Payoffs are:

	F	B
F	$2x, x$	$0, 0$
B	$0, 0$	$x, 2x$

Both (F, F) and (B, B) are Nash equilibria, and also fairness equilibria. However, if x is small, (F, B) and (B, F) are also fairness equilibria. The reason is that each player feels the other is trying to hurt him by miscoordinating, so he wants to respond similarly.

Example 2 Prisoner's Dilemma. Payoffs are:

	C	D
C	$4x, 4x$	$0, 6x$
D	$6x, 0$	$0, 0$

In the prisoner's dilemma, (C, C) will be a fairness equilibrium if x is small enough (so the material stakes are small), while (D, D) is always a fairness equilibrium.

Example 3 Hawk-Dove. Payoffs are:

	H	D
H	$-2x, -2x$	$0, 2x$
D	$2x, 0$	x, x

Here, if x is small, the Nash equilibria (H, D) and (D, H) are not fairness equilibrium. The player playing D feels that the other is trying to hurt her, and wants to be mean back, which means playing H.

Rabin also proves a few more general results. In particular, say that (a_1, a_2) is a *mutual-max* outcome if for $i = 1, 2$, $a_i \in \arg\max_{a \in S_i} \pi_j(a, a_j)$. Similarly, one can define a *mutual-min* equilibrium. Rabin shows that if (a_1, a_2) is a Nash equilibrium and also a mutual-max or mutual-min profile, then it must be a fairness equilibrium. Moreover, if stakes are small, then a mutual-max or mutual-min profile will be a fairness equilibrium even if it is not Nash.

3 Open Questions

1. Some recent experimental papers by Fehr, Rabin and others suggest very strongly that a convincing model of fairness should include some role for intentions. But Rabin's paper seems somewhat hard to make operational outside of a limited class of games. The Fehr-Schmidt model has the attractive feature of being very simple, but it misses the idea that intentions matter. This leaves the problem open for future work.

2. Recent work on this problem has focused on the laboratory, but earlier research by George Akerlof, Robert Frank and others has considered the implications of fairness or concerns about relative status in broader economic contexts. This seems like another area where there is significant room for research.

3. Perceptions of whether a situation is fair can be very sensitive to framing. As an example, Gneezy (2003) discusses the results of several experiments in which it is found that using small monetary incentives is counter-productive relative to using large incentives or no incentives at all. An explanation he offers is that people are either insulted by small compensation, or interpret it as meaning that the task is odious without feeling that it makes the task worthwhile. But is seems quite easy to imagine situations where offering a small "token of appreciation" for a service could elicit a large response. This suggests that when it comes to incentive systems, the way in which a system is framed can be of crucial importance. Again, this is something economists have made minimal progress in understanding.

References

[1] Bolton, Gary and Axel Ockenfels, (2000) "ERC: A Theory of Equity, Reciprocity and Competition," *Amer. Econ. Rev.,* 90, 166–193.

[2] Fehr, Ernst and Simon Gachter (2001) "Do Incentive Contracts Crowd Out Voluntary Cooperation?" mimeo.

[3] Fehr, Ernst and Klaus Schmidt (1999) "A Theory of Fairness, Competition, and Cooperation," *Quarterly J. Econ.,* 114, 817–868.

[4] Geanakoplos, John, David Pearce and Ennio Stacchetti (1989) "Psychological Games and Sequential Rationality," *Games and Econ. Behav.*, 1, 60–79.

[5] Gneezy, Uri (2003) "The W-effect of Incentives," Working Paper.

[6] Rabin, Matthew (1993) "Incorporating Fairness into Game Theory and Economics," *Amer. Econ. Rev.*, 83, 1281–1302.