Distributional Preferences, Reciprocity-Like Behavior, and Efficiency in Bilateral Exchange

By DANIEL J. BENJAMIN*

Under what conditions do distributional preferences, such as altruism or a concern for fair outcomes, generate efficient trade? I analyze theoretically a simple bilateral exchange game: each player sequentially takes an action that reduces his own material payoff but increases the other player's. Each player's preferences may depend on both his/her own material payoff and the other player's. I identify two key properties of the second-mover's preferences: indifference curves kinked around "fair" material-payoff distributions, and materials payoffs entering preferences as "normal goods." Either property can drive reciprocity-like behavior and generate a Pareto efficient outcome.
JEL: D63, J33, J41, M52, D64
Keywords: distributional preferences, fairness, altruism, gift exchange, rotten kid theorem

Under what conditions will bilateral exchange be Pareto efficient? Enforceable contracts (Coase 1960) or repeated interaction (Fudenberg & Maskin 1986) can lead to efficient exchange under some conditions. This paper addresses a third possible source of efficiency: a direct concern for the welfare of the other party, often called distributional preferences, such as altruism or a concern for fair outcomes.

The setting I analyze is a simple, two-stage bilateral exchange game, e.g., an employer-worker interaction. The game is defined in terms of "material payoffs," the players' private utilities that do not take into account any concern for the other player. Each of the two players in turn chooses how much of an action to take. For each player, a higher level of his action increases the other player's material payoff but at the cost of reducing his own material payoff. For example, by increasing the wage, an employer increases the worker's consumption but reduces profit; and by increasing effort, the worker increases the employer's profit but incurs disutility of effort. To focus on the role of distributional preferences, I assume that contracting is infeasible and that the exchange is one-shot. Hence, if both players were purely self-regarding—caring only about their own material

* Department of Economics, Cornell University, 480 Ursis Hall (E-mail: db468@cornell.edu). A previous version of this paper circulated under the title "Social Preferences and the Efficiency of Bilateral Exchange." I am grateful for comments and feedback to more people than I can list. I am especially grateful to James Choi, Steve Coate, Ed Glaeser, Ori Heffetz, Ben Ho, David Laibson, Ted O'Donoghue, Sendhil Mullainathan, Stefan Penczynski, Giacomo Ponzetto, Josh Schwartzstein, Jesse Shapiro, Andrei Shleifer, Joel Sobel, Jón Steinsson, and Jeremy Tobacman. I thank the Program on Negotiation at Harvard Law School; the Harvard University Economics Department; the Chiles Foundation; the Federal Reserve Bank of Boston; the Institute for Quantitative Social Science; Harvard's Center for Justice, Welfare, and Economics; the National Institute of Aging through grant T32-AG00186 and to the National Bureau of Economic Research and P01-AG26571 to the Institute for Social Research; the Institute for Humane Studies; and the National Science Foundation for financial support. I am grateful to Julia Galef, Dennis Shiraev, Jelena Veljic, and Jeffrey Yip for excellent research assistance, and especially Gabriel Carroll, Ahmed Jaber, and Hongyi Li, who not only provided outstanding research assistance but also made substantive suggestions that improved the paper. All mistakes are my fault.

payoff—then no gains from trade would be realized because neither player would have any reason to choose a positive amount of his action.

Instead of being purely self-regarding, each player has distributional preferences that depend on both his own and the other player's material payoff, and thus players might be willing to choose a positive action. Moreover, the second-mover's (SM's) optimal action may depend on the first-mover's (FM's) action. If so, then even if FM is purely self-regarding, it may turn out to be optimal for FM to take an action that, together with SM's optimal response, generates a Pareto improvement relative to no trade. In fact, it is possible that at the equilibrium of the game, the outcome is Pareto efficient: all potential gains from trade are realized. I identify properties of the players' preferences that may lead the outcome of their interaction to be Pareto efficient.

While much of the literature on distributional preferences assumes a particular model of distributional concerns, I study how results depend on general properties of distributional preferences that are shared by many specific models. Two properties play a particularly prominent role. The first is defined in terms of the agent's interpersonal indifference curves, which describe how the agent trades off between FM's material payoff and SM's. The property of "fairness-kinkedness"—illustrated in Figure 1a, where the axes are SM's and FM's material payoffs, π_2 and π_1—means that the agent's indifference curves are kinked at each material payoff pair along a curve. This curve, along which both players' material payoffs are increasing, is called the "fairness rule." The fairness rule describes the set of material payoff pairs that the agent considers to be "fair." Because of the kinked indifference curves, when facing a choice that requires trading off between the players' material payoffs, the agent chooses an action that exactly implements one of these fair transactions for a range of rates of tradeoff. Several leading models of distributional preferences satisfy fairness-kinkedness (e.g., Fehr & Schmidt 1999; Charness & Rabin 2002) because they embed the assumption that indifference curves are piecewise-linear and kinked at transactions where the players earn equal material payoffs, as illustrated in Figure 1b. The more general property of fairness-kinkedness, however, can accommodate non-linear indifference curves and fairness rules involving unequal material payoffs, e.g., a worker may judge as fair the material payoffs that correspond to the market rate of exchange between money and effort (Kahneman, Knetsch, & Thaler, 1986).

[FIGURE 1 ABOUT HERE]

The second property is "normality": both players' material payoffs enter the distributional preferences as "normal goods." Analogously to consumer theory, normality means that if the frontier of attainable material payoffs for the players shifts outward holding fixed the rate of tradeoff, then the agent prefers that both players get a higher material payoff. Normality seems like a natural property for distributional preferences designed to capture a concern for fairness, and indeed most existing fairness models (e.g., Fehr & Schmidt 1999, Charness & Rabin 2002) satisfy at least a weak version of it.

Throughout, I impose two assumptions that rule out potential sources of inefficiency. First, I assume that SM's distributional preferences are strong enough that FM is willing

to transact rather than take her outside option. Due to this assumption, the efficiency results should be interpreted as describing when exchange is predicted to be efficient, conditional on the players choosing to trade. Second, I assume that FM is either purely self-regarding—as when FM is a profit-maximizing firm—or has distributional preferences that are monotonically increasing in both players' material payoffs. Although existing models allow for distributional preferences to be non-monotonic, this is primarily to proxy for reciprocity by the *second* mover, and most of the evidence from simple dictator game experiments actually indicates that most people have monotonic distributional preferences (e.g., Andreoni & Miller 2002; Charness & Rabin 2002; Fisman, Kariv, & Markovits 2007). In Web Appendix A, I explore how the results are affected if this monotonicity assumption is relaxed.

The central results of the paper describe two main cases in which distributional preferences generate efficiency in bilateral exchange, and show that these are essentially the *only* two cases in which the equilibrium is efficient. In one case, normality plays a key role, and in the other, fairness-kinkedness does. First, if SM's distributional preferences satisfy normality, and if SM's action is a linear transfer of material payoff from himself to FM—e.g., SM's action is a monetary payment—then the equilibrium is efficient. Because SM faces the same linear tradeoff between the players' material payoffs regardless of FM's action, FM's action simply shifts the frontier of attainable material payoffs inward or outward. If FM's action shifts the frontier outward, then since SM's distributional preferences satisfy normality, SM will take an action that generates greater material payoff for both players. Because SM's behavior ensures that the players' material incentives are aligned, FM will take the level of her action that maximizes aggregate material surplus.

The second case does not require SM's action to be a linear transfer. If SM's distributional preferences are sufficiently fairness-kinked, then he always chooses an action that generates an outcome that is on the fairness rule. The equilibrium is efficient because, intuitively, when SM behaves in accordance with a fairness rule (such as the fairness rule shown in Figure 1a), he aligns the players' material incentives. Therefore, FM maximizes both players' material payoffs by choosing the action that induces the highest achievable point on the fairness rule, i.e., where the fairness rule intersects the frontier of attainable material payoffs. Existing laboratory evidence suggests that such fairness-rule-based behavior is plausible, and indeed the equal-split fairness rule depicted in Figure 1b often governs behavior in laboratory experiments. The result highlights the economic relevance of examining empirically how often people feel compelled to behave in accordance with rules of fair behavior in economic settings outside the laboratory.

As far as I am aware, the efficiency result involving fairness-kinkedness is novel. Other results in this paper generalize and unify results that are known for special cases, while highlighting the largely unappreciated central roles played by fairness-kinkedness and normality. The analysis also helps to bridge separate theoretical literatures on altruism, defined as a preference to increase the other player's payoff, and fairness concerns, notions of which may be captured by fairness-kinkedness or normality. For example, the efficiency result involving normality generalizes the well-known rotten kid theorem

(Becker 1974; Bergstrom 1989) and shows that, contrary to the theorem's traditional interpretation as about altruism, it is actually driven by normality.

Two recent papers take a similar approach to this paper of applying tools from classical demand theory to analyze implications of general properties of other-regarding preferences. Cox, Friedman, and Sadiraj (2008) propose axioms that generalize and extend existing models and explore the predictions of these axioms in some laboratory games. Dufwenberg, Heidhues, Kirchsteiger, Riedel, & Sobel (2011) study the implications of general properties of other-regarding preferences in a general equilibrium environment.

The rest of the paper is organized as follows. Using the rotten kid theorem and a gift-exchange game as examples, and imposing relatively specific assumptions on preferences, Section 1 illustrates and previews the main results of the paper. Section 2 lays out the more general set-up of the bilateral exchange game. Section 3 introduces the general properties that distributional preferences might satisfy. Section 4 shows how the same properties of distributional preferences that can lead to an efficient outcome—either fairness-kinkedness or normality—are also properties that give rise to reciprocity-like behavior in the bilateral exchange game. Section 5 characterizes how the set of outcomes that are efficient when players have distributional preferences relates to (and differs from) the set of efficient outcomes when both players are purely self-regarding. Section 6 derives necessary conditions for the equilibrium to be efficient, and shows that the two cases mentioned above are essentially the only cases in which distributional preferences can generate an efficient equilibrium. Section 7 provides two sets of sufficient conditions for the equilibrium to be efficient, each corresponding to one of the two cases. Section 8 discusses possible extensions of the analysis and additional testable predictions. Web Appendix A analyzes the case where FM's distributional preferences are non-monotonic, and Web Appendix B contains all proofs.

I. Model Set-Up and Illustrative Examples

In this section, I analyze two examples that preview and illustrate the main results of the paper. The set-up is a sequential bilateral-exchange environment. FM chooses the level of her action, $a_1 \in \mathbb{R}$, and then SM chooses the level of his action, $a_2 \in \mathbb{R}$. For each player, a higher level of one's action helps the other player but hurts oneself. The **material payoff functions**, $\pi_1(a_1, a_2)$ and $\pi_2(a_1, a_2)$, describe how the players' actions determine the "material payoffs" from the transaction. Material payoffs represent the purely self-regarding component of players' outcomes from the transaction but not necessarily their preferences. Preferences are represented by utility functions, $U_1(\pi_1, \pi_2)$ and $U_2(\pi_1, \pi_2)$ respectively, which may depend not only on the agent's own material payoff but also on the other player's material payoff. The equilibrium concept is subgame-perfect equilibrium.

Example 1. The rotten kid game. FM is a child who chooses how much effort a_1 to exert to earn money for the family. Then SM, the parent, transfers to the child some amount of family income, a_2. The child's private income is $I_1 + a_2 - n(a_1)$, where

$I_1 \geq 0$ is exogenous income, and $n(a_1)$ is his cost-of-effort function (in dollars) satisfying $n' > 0$, $n'' > 0$, $\lim_{x\to-\infty} n'(x) = 0$, $n'(0) < 1$, and $\lim_{x\to\infty} n'(x) = \infty$. The parent's private income is $I_2 + a_1 - a_2$, where $I_2 \geq 0$ is an exogenous component of the parent's income. "Family income" is the sum of the child's and parent's incomes: $I_1 + I_2 + a_1 - n(a_1)$. The child's consumption is $\pi_1(a_1, a_2) = (I_1 + a_2 - n(a_1))/P_1$, where $P_1 > 0$ is the market price of consumption faced by the child. The parent's consumption is $\pi_2(a_1, a_2) = (I_2 + a_1 - a_2)/P_2$, where $P_2 > 0$ (possibly equal to P_1) is the market price of consumption faced by the parent. The child is purely self-regarding (a "rotten kid"): $U_1(\pi_1, \pi_2) = \pi_1$. The parent is altruistic: $U_2(\pi_1, \pi_2)$ is not only strictly increasing in π_2 but also in π_1. It is also assumed that $U_2(\pi_1, \pi_2)$ is twice-continuously differentiable and strictly quasi-concave, and π_1 and π_2 enter U_2 as normal goods. Finally, as a technical condition that serves only to ensure that the parent's optimal action is finite, I assume that there exist $\underline{\pi}_1 < 0$ and $\underline{\pi}_2 < 0$ such that $\lim_{\pi_2\to\infty} \frac{\partial U_2(\underline{\pi}_1,\pi_2)/\partial\pi_2}{\partial U_2(\underline{\pi}_1,\pi_2)/\partial\pi_1} = 0$ and $\lim_{\pi_1\to\infty} \frac{\partial U_2(\pi_1,\underline{\pi}_2)/\partial\pi_2}{\partial U_2(\pi_1,\underline{\pi}_2)/\partial\pi_1} = \infty$. Becker's (1974, p.1080) celebrated rotten kid theorem is:

Proposition 1 (Rotten kid theorem). *In the equilibrium of the rotten kid game, the child chooses the level of a_1 that maximizes family income.*

The rotten kid theorem is generally interpreted as showing that an efficient outcome can occur within the family due to the parent being altruistic (e.g., Becker, 1974). Bergstrom (1989) pointed out that Becker's example hinges on the assumption that the material payoffs are quasi-linear in a_2 but continued to describe the theorem as a result about altruism. Although typically not defined explicitly, altruism is usually understood as meaning that preferences depend positively on the material payoff of the other person. I will refer to this property of distributional preferences as "monotonicity."

The analysis in this paper will show that the rotten kid theorem is *not* driven by monotonicity, but rather by the combination of the material payoffs being quasi-linear in a_2 with the "normality" assumption: π_1 and π_2 enter U_2 as normal goods. Indeed, Theorem 3 in Section VII is a generalization of Proposition 1 in which monotonicity is relaxed. Moreover, I will argue that normality captures a kind of concern for fair distribution; in Example 1, normality means that when family income increases, the parent prefers that both players share in the material gains. Such a concern for fairness appears to be widespread in interactions between unrelated individuals (e.g., Kahneman, Knetsch, & Thaler 1986). Therefore, rather than as a result about altruism within the family, the rotten kid theorem should be interpreted as a result about fairness preferences that may be relevant to a wider range of settings.

Example 2. Gift-exchange game with a profit-maximizing firm. FM is a firm who chooses a worker's salary, a_1. Then SM, the worker, chooses his level of effort, a_2. The firm's profit is $\pi_1(a_1, a_2) = a_2 - a_1$. The worker's material payoff is $\pi_2(a_1, a_2) = a_1 - c(a_2)$, where $c(a_2)$ is his cost-of-effort function satisfying $c(0) = 0$,

$c' > 0$, $c'' > 0$, $\lim_{x\to-\infty} c'(x) = 0$, $c'(0) < 1$, and $\lim_{x\to\infty} c'(x) = \infty$. Since the material payoff functions are quasi-linear in a_1, any transaction (a_1, a_2) where $c'(a_2) = 1$ is Pareto efficient in terms of the material payoffs. The firm is profit maximizing: $U_1(\pi_1, \pi_2) = \pi_1$. Both players anticipate the subgame-perfect equilibrium, and if either would earn negative utility from the game, then they do not transact and instead each get an outside-option material payoff of 0. The worker has distributional preferences that are piecewise linear and hence kinked. The preferences weight the firm's and workers's material payoffs differently, depending on which player earns more:

$$U_2(\pi_1, \pi_2) = \begin{cases} \sigma\pi_1 + (1-\sigma)\pi_2 & \text{if } \pi_1 > \pi_2 \\ \rho\pi_1 + (1-\rho)\pi_2 & \text{if } \pi_1 \leq \pi_2 \end{cases}, \tag{1}$$

where $\sigma < 1$ is the relative weight on the firm's material payoff when the firm is ahead, and $\rho \in (\sigma, 1]$ is the relative weight on the firm's material payoff when the worker is ahead. For example, the set of parameter values that correspond to Fehr & Schmidt's (1999) inequity-aversion model is $\sigma < 0 < \rho < 1$, while Charness & Rabin (2002) argue that $0 < \sigma < \rho < 1$. Either way, the equilibrium is efficient if the worker's distributional preferences are "sufficiently kinked," as made precise in the following proposition:

Proposition 2. *In the gift-exchange game with a profit-maximizing firm, there exists $\overline{\sigma} > 0$ such that if $\sigma < \overline{\sigma}$ and $\rho \geq \frac{1}{2}$, then the equilibrium transaction is Pareto efficient in terms of the material payoffs.*

For intuition, it is clearest to begin with the special case $\sigma \leq 0$. In that case, $\rho \geq \frac{1}{2}$ is not only sufficient but also necessary for the equilibrium transaction to be Pareto efficient in terms of the material payoffs.[1] When the worker exerts less than the efficient level of effort, a marginal increase in effort increases the firm's material payoff more than it reduces the worker's. Since $\rho \geq \frac{1}{2}$, the worker when ahead puts at least as much weight on the firm's material payoff as his own, but since $\sigma \leq 0$, the worker when behind puts non-positive weight on the firm. Consequently, for any salary at which the worker ends up exerting less than the efficient level of effort, the worker would increase his effort exactly up to (and not beyond) the level that equates the firm's material payoff with his own. The players' material incentives are therefore aligned, and the firm maximizes its own material payoff by setting the salary level that induces the efficient level of effort.

The situation is more complex when $\sigma > 0$ because the worker may be willing to increase his effort beyond the level that equates the material payoffs, in which case the players' material incentives are no longer aligned. However, if σ is small enough, then at relatively high salaries (which induce high effort and hence a high marginal cost of effort) the worker still increases his effort only up to the level that equates the material

[1] One might wonder whether $\rho \geq \frac{1}{2}$ is empirically plausible. If material payoff functions are quasi-linear in money (as assumed in Example 2), then ρ can be estimated from experimental participants' allocations of money. Fehr & Schmidt (1999, Table III and p.864) suggest that about 40% of subjects have $\rho \geq \frac{1}{2}$. Drawing on a broader set of experimental games, Charness & Rabin's (2002, Table VI, row 5) estimates are also consistent with a sizeable minority of participants satisfying $\rho \geq \frac{1}{2}$.

payoffs. Even though a relatively low salary may evoke effort beyond the equal-payoff level, if σ is small enough, the effort will be low enough that the firm could earn higher profit by offering the higher salary that induces the efficient level of effort.

Theorem 4 in Section VII generalizes Proposition 2 to a more general class of fairness-kinked distributional preferences, in which the preferences are convex rather than piecewise-linear, and the kinks do not necessarily occur at equal material payoffs. Unlike in Example 1, where quasi-linearity of the material payoffs is crucial, in Example 2 it merely simplifies stating sufficient conditions for efficiency; Theorem 4 allows for more general, convex material-payoff functions. I will argue that the fairness-kinkedness of the distributional preferences represents another kind of concern for fair distribution (different from normality): a motivation to follow a "rule" of fair behavior described by the set of material payoffs where the kinks occur. In Example 2, the rule is to equalize the players' material payoffs. Thus, Example 2 illustrates a type of efficiency result that can arise from fairness preferences that is distinct from Example 1. Theorems 3 and 4 also generalize the examples in another way: they show in each case that the equilibrium is—in addition to being Pareto efficient in terms of material payoffs—also Pareto efficient in terms of overall preferences.

II. The Bilateral Exchange Game

In this section, I generalize the games in the examples from the previous section; in the next section, I generalize the distributional preferences. In addition to the rotten kid game and the gift-exchange game, the bilateral exchange environment I introduce in this section includes as special cases the trust game (Berg, Dickhaut, & McCabe 1995); a two-player, sequential, public goods game; and a version of the hold-up problem where, after FM makes a costly irreversible investment, SM has all the bargaining power in determining how the surplus is divided.

FM chooses the level of her action, a_1, and then SM chooses the level of his action, a_2. To ensure that all optimal actions are interior and thereby simplify exposition, I assume that $a_1, a_2 \in \mathbb{R}$.[2] The outcome of the game is a **transaction**, (a_1, a_2). As in many exchange settings in the field, I assume that the players could alternatively choose not to transact. In that case, both players receive an outside-option payoff as if the action pair had been $(0, 0)$. The outside-option material payoffs are normalized to zero: $\pi_1(0, 0) = \pi_2(0, 0) = 0$.

The material payoff functions are twice-continuously differentiable and have these properties, which I will always assume:

A1. *Each player's action increases the other player's material payoff while reducing his or her own:* $\frac{\partial \pi_1}{\partial a_1} < 0$, $\frac{\partial \pi_2}{\partial a_1} > 0$, $\frac{\partial \pi_1}{\partial a_2} > 0$, *and* $\frac{\partial \pi_2}{\partial a_2} < 0$.

[2] In applications, it is instead typical to assume that $a_1 \in [0, \overline{A}_1]$ and $a_2 \in [0, \overline{A}_2]$ for some upper bounds $\overline{A}_1$ and $\overline{A}_2$. My assumption that the action spaces are unbounded has the drawback that it necessitates technical conditions (such as A4 below) to ensure the existence of optimal actions. If the action space were closed and bounded, then these technical conditions could be eliminated, but the propositions would have to separately deal with cases where optimal actions are not interior.

A2. *There are (material) gains from trade:* $\frac{-\partial \pi_1(0,0)/\partial a_1}{\partial \pi_1(0,0)/\partial a_2} < \frac{\partial \pi_2(0,0)/\partial a_1}{-\partial \pi_2(0,0)/\partial a_2}$.

A3. *The functions* $\widetilde{\pi}_1(a_1, a_2)$ *and* $\widetilde{\pi}_2(a_1, a_2)$*, defined by* $\widetilde{\pi}_1(a_1, a_2) \equiv \pi_1(-a_1, a_2)$ *and* $\widetilde{\pi}_2(a_1, a_2) \equiv \pi_1(a_1, -a_2)$*, are both weakly concave; and at least one is strictly concave in at least one of its arguments.*

A4. *(Technical condition) Fixing any* $\widehat{a}_1$ *and* $\widehat{a}_2$*, each of the mappings from one agent's action to a real number given by* $\pi_1(\widehat{a}_1, a_2)$*,* $\pi_2(\widehat{a}_1, a_2)$*,* $\pi_1(a_1, \widehat{a}_2)$*, and* $\pi_2(a_1, \widehat{a}_2)$*, is surjective.*

A2 means that there exist some transactions involving positive actions for both players such that both earn a positive material payoff: for any sufficiently small, positive actions $da_1 > 0$ and $da_2 > 0$ such that FM's material payoff equals 0, i.e., $\frac{\partial \pi_1(0,0)}{\partial a_1} da_1 + \frac{\partial \pi_1(0,0)}{\partial a_2} da_2 = 0$, SM's material payoff is strictly positive: $\frac{\partial \pi_2(0,0)}{\partial a_1} da_1 + \frac{\partial \pi_2(0,0)}{\partial a_2} da_2 > 0$. A3 helps guarantee that the equilibrium is unique. Since the action spaces are unbounded, A4 helps ensure that optimal actions exist.

The players maximize their utility functions, which may depend on the material payoffs received by both players (according to properties described in the next section). The solution concept is subgame-perfect equilibrium. Because payoffs and preferences are common knowledge, both players correctly anticipate the equilibrium of the game. Therefore, if either player would get negative utility from trading, then the players do not trade.

III. Distributional Preferences

An agent with **distributional preferences** has preferences that depend on both players' outcomes. When studying behavior in experiments, the typical approach is to define distributional preferences over the players' incremental *monetary* payoffs earned in the experiment. In many field settings, however, the players' actions affect at least one commodity other than money, such as effort. In order to analyze such settings, I define distributional preferences over the (full) *material* payoffs from the transaction. This formulation specializes to preferences over incremental monetary payoffs in experiments where the players' actions only affect their earnings.

In this section, I specify general properties that FM's and SM's respective distributional preferences could satisfy. I begin by defining the two properties that will play a central role in generating an efficient equilibrium and then turn to properties that primarily serve as regularity conditions.

The first property, which I call "fairness-kinkedness," formalizes kinked indifference curves without building in piecewise-linearity or the restriction that the kinks occur at 50-50 split allocations. While Fehr & Schmidt (1999) interpret the kinks in their model as reflecting loss aversion in social comparisons, Charness & Rabin (2002) treat the kinks in their own model as just a byproduct of the simplifying assumption of piecewise-linearity. In any event, the kinks around 50-50 splits account for some of the descriptive accuracy of these models in laboratory experiments. In particular, as Fehr & Schmidt (1999)

note, the kinks are the feature of the model that enables it to explain why in dictator games, subjects often give exactly half of the money to the other player (see Camerer 2003 for a review). Moreover, the kinks can explain why many of the same people who choose exactly even splits in a dictator game also choose to assign equal monetary payoffs to themselves and another player in modified dictator games, where the "price" of increasing one player's payoff by \$1 is less than \$1 (e.g., Andreoni & Miller 2002). No smooth distributional preferences could explain equal-split behavior in both cases. Hence, a kink in the indifference curve can be interpreted as describing a "rule" for how to allocate payoffs in the sense that over some range of prices, the prescribed behavior is insensitive to the price (see Andreoni & Bernheim, 2009, for an alternative model based on signaling).

Let a strictly increasing function $f(\pi_2)$ describe what the agent considers to be a "fair" material payoff for FM for each possible material payoff for SM. For fairness-kinked preferences, the graph of f—which I call the **fairness rule**—is the set of material payoff pairs where the indifference curves are kinked. Existing models with kinks embed the "equal-split rule" into preferences, defined by $f(\pi_2) = \pi_2$. Generalizing f allows preferences to capture adherence to whatever rule of fair behavior might be relevant to a particular setting.[3] Using labels suitable for SM, let $D_f \equiv \{(\pi_1, \pi_2) | \pi_1 > f(\pi_2)\}$ denote the region of **disadvantageously unfair** transactions, where FM's material payoff is higher and SM's material payoff is lower than dictated by the fairness rule; and let $A_f \equiv \{(\pi_1, \pi_2) | \pi_1 < f(\pi_2)\}$ denote the region of **advantageously unfair** transactions for SM. Figure 1a illustrates these regions. (In all figures, I put π_1 on the y-axis because in the simple case in which FM is self-regarding, solving for equilibrium amounts to maximizing π_1.)

Definition 1. *U is **fairness-kinked** if (a) $U(\pi_1, \pi_2)$ is twice-continuously differentiable except along a fairness rule f; (b) for all $(\pi_1, \pi_2) \in D_f$, $\partial U/\partial \pi_2 > 0$; and (c) for all $(\pi_1, \pi_2) \in A_f$, $\partial U/\partial \pi_1 > 0$.*

For example, the piecewise-linear distributional preferences (1) are fairness-kinked if (*a*) $\sigma \neq \rho$, (*b*) $\sigma < 1$, and (*c*) $\rho > 0$, as in both Fehr & Schmidt (1999) and Charness & Rabin (2002).

A second property, "normality," can also capture a concern for fairness. In consumer theory, an agent's preferences have the "normal good" property with respect to a particular good if, for prices held fixed, the agent chooses to consume more of that good when his income increases. Normality can be defined analogously for distributional preferences, but in this context, the "goods" are the material payoffs of the players. For some price $p > 0$ and income $I \in \mathbb{R}$, define $\widetilde{\pi}_1(p; I)$ and $\widetilde{\pi}_2(p; I)$ by $(\widetilde{\pi}_1, \widetilde{\pi}_2) = \arg\max_{\{(\pi_1,\pi_2):\pi_1+p\pi_2=I\}} U(\pi_1, \pi_2)$. Assume that $\widetilde{\pi}_1(p; I)$ and $\widetilde{\pi}_2(p; I)$

[3] While 50-50 splits often serve as a benchmark for what is fair in contexts where payoffs are monetary—such as in negotiations, asymmetric joint ventures among corporations, share tenancy in agriculture, and bequests to children (Andreoni & Bernheim 2009)—there are exceptions, e.g., financial contracts often apportion profit according to unequal percentages that are standard in the industry. Moreover, in settings involving two commodities or a commodity in exchange for money, the rate of pay that is considered fair is often determined by prevailing market prices or recent experiences (Kahneman, Knetsch, & Thaler 1986).

are finite, real-valued functions (which will be implied by the other assumptions on U, given below).

Definition 2. *For $i = 1, 2$, U is **(weakly) locally normal in** π_i **at** $(p; I)$ if $\widetilde{\pi}_i(p; I)$ is (weakly) increasing in I at $(p; I)$. U is **(weakly) normal in** π_i if U is (weakly) locally normal in π_i at $(p; I)$ for all $p > 0$ and $I \in \mathbb{R}$. U is **(weakly) normal** if U is (weakly) normal in both π_1 and π_2.*

Following Becker (1974), it is common in models of altruism to assume that distributional preferences satisfy not only monotonicity but also normality. However, while monotonicity is intrinsic to the notion of altruism, the connection between normality and altruism is questionable. Instead, normality is more naturally interpreted as capturing a concern for fairness. It amounts to assuming that FM's material payoff and SM's material payoff enter the utility function as complements.[4] Normality is not assumed explicitly in existing fairness models, but it is a byproduct of most of the specific functional forms that are adopted. While seemingly a natural assumption, it has strong implications, as will be seen.

Turning to regularity conditions, a standard assumption about preferences is monotonicity: utility is strictly increasing in each player's material payoff.

Definition 3. *U is **monotonic** if $U(\pi_1, \pi_2)$ is strictly increasing in both π_1 and π_2.*

Monotonicity is the defining feature of altruism, and all models of altruism assume it.

Some models of distributional preferences aimed at capturing a concern for fairness also satisfy monotonicity, such as Charness & Rabin's (2002), but some do not (e.g., Fehr & Schmidt 1999; Bolton & Ockenfels 2000). In particular, these latter models assume that people are "behindness averse," preferring to reduce the other player's payoff when that player's payoff is higher than their own. For example, in the piecewise-linear model (1), behindness aversion corresponds to $\sigma < 0$.

To allow for this kind of non-monotonicity, I define a weaker property that I call "joint-monotonicity."[5]

Definition 4. *U is **joint-monotonic** if for any (π_1, π_2) and any $\varepsilon > 0$, there is some $(\widehat{\pi}_1, \widehat{\pi}_2)$ such that $0 < \widehat{\pi}_1 - \pi_1 < \varepsilon$, $0 < \widehat{\pi}_2 - \pi_2 < \varepsilon$, and $U(\widehat{\pi}_1, \widehat{\pi}_2) > U(\pi_1, \pi_2)$.*

The definition states that for any material payoff pair, there is an arbitrarily close alternative material payoff pair giving more to *both* players that the agent strictly prefers. It

[4] To be precise, at any material payoff pair where $\partial U(\pi_1, \pi_2)/\partial \pi_1 > 0$ and $\partial U(\pi_1, \pi_2)/\partial \pi_2 > 0$, the statement about behavior "U is locally normal in π_i" is equivalent to the following statement about complementarity in preferences: $\frac{\partial}{\partial \pi_i}\left(\frac{\partial U/\partial \pi_i}{\partial U/\partial \pi_{-i}}\right) < 0$ (Quah, 2007, Theorem S1 and Proposition S1). The conditions $\partial U(\pi_1, \pi_2)/\partial \pi_1 > 0$ and $\partial U(\pi_1, \pi_2)/\partial \pi_2 > 0$ may not hold at every material payoff pair when U is joint-monotonic (as defined below) and not monotonic. However, the analysis will show that normality is a relevant property for SM's distributional preferences (not FM's), and Lemma 1 will establish that $\partial U(\pi_1, \pi_2)/\partial \pi_1 > 0$ and $\partial U(\pi_1, \pi_2)/\partial \pi_2 > 0$ hold at an optimum for SM.

[5] In studying other-regarding preferences in a general equilibrium environment, Dufwenberg, Heidhues, Kirchsteiger, Riedel, & Sobel (2011) independently propose a "social monotonicity" property, which is similar to my joint-monotonicity property, except that it is a restriction on both players' distributional preferences. I discuss the relationship between social monotonicity and joint monotonicity in Appendix A.

implies local non-satiation but additionally requires that it is possible to find a more-preferred allocation in a particular direction, a direction which jointly increases both players' material payoffs. The interpersonal indifference curves depicted in Figures 1a and 1b represent distributional preferences that violate monotonicity but satisfy joint-monotonicity. While ruling out pure spitefulness and pure self-hating, joint-monotonicity allows for behindness aversion. More generally, it permits the possibility that an agent might prefer to reduce either one or the other player's material payoff to reach what the agent considers to be a fairer allocation.

In much of the analysis, I will assume that SM's distributional preferences are joint-monotonic but that FM either is purely self-regarding or has monotonic distributional preferences. Given that some of the existing models allow for behindness aversion in order to describe behavior in experiments, this assumption about FM might seem suspect. There are two distinct justifications for it. First, while there is debate over whether behindness aversion should be assumed, most direct evidence from experiments in fact indicates that most subjects' distributional preferences satisfy monotonicity.[6] Advocates of behindness aversion primarily argue that it should be assumed because it provides a tractable shortcut for capturing reciprocity-like behavior by a second mover (e.g., Fehr & Schmidt 2004, p.10; Fehr & Schmidt 2003), which is valuable because models of reciprocity itself (e.g., Rabin 1993) are notoriously difficult to work with. The assumption that FM has monotonic preferences is compatible with this argument in favor of assuming that SM's preferences are joint-monotonic. Second, in an exchange situation in which FM is a profit-maximizing firm, it is appropriate to assume that FM is purely self-regarding. In Web Appendix A, I discuss the more complex case where FM is assumed to have merely joint-monotonic preferences.

The final property, quasi-concavity, is familiar from consumer theory and social choice.

Definition 5. *U is **quasi-concave** if for any two distinct material payoff pairs, (π_1, π_2) and $(\widehat{\pi}_1, \widehat{\pi}_2)$, such that $U(\pi_1, \pi_2) \leq U(\widehat{\pi}_1, \widehat{\pi}_2)$, $U(\pi_1, \pi_2) < U(\lambda\pi_1 + (1-\lambda)\widehat{\pi}_1, \lambda\pi_2 + (1-\lambda)\widehat{\pi}_2)$ for any $\lambda \in [0, 1]$. U is **weakly quasi-concave** if the strict inequality is replaced by a weak inequality.*

For distributional preferences, quasi-concavity means that along an interpersonal indifference curve, the higher FM's material payoff, the less of SM's material payoff the decision-maker is willing to give up to increase FM's material payoff (and similarly with "FM" and "SM" switched). Equivalently, it means that the upper level sets of U are convex. Every model of distributional preferences that I am aware of satisfies

[6]The debate has largely centered on the question of whether subjects care more about "efficiency" (in this context, meaning the sum of monetary payoffs) or "equity" (meaning equality of monetary payoffs), and the experimental findings are contradictory (e.g., Engelmann & Strobel 2004; Fehr, Naef, & Schmidt 2006). The question of whether subjects' distributional preferences are monotonic is related but distinct. Almost all of the experiments involving simple allocation decisions by adult subjects find that most people do have monotonic distributional preferences (Charness & Grosskopf 2001; Kritikos & Bolle 2001; Andreoni & Miller 2002; Charness & Rabin 2002; Fisman, Kariv, & Markovits 2007; Cox & Sadiraj 2010). The exceptions in which a majority of subjects violate monotonicity are: Bazerman, Loewenstein, & White (1992), who report evidence from hypothetical choices; Bolton & Ockenfels (2006), from an experiment in which subjects vote over allocations; and Pelligra & Stanca (2013), from an Internet survey where the dictator games have a small chance of being played out for real money.

quasi-concavity (e.g., Bolton & Ockenfels, 2000) or weak quasi-concavity (e.g., Fehr & Schmidt 1999; Charness & Rabin 2002).[7]

While the above properties will be listed explicitly when assumed in the propositions, the following two technical assumptions (TAs) will be maintained implicitly throughout. TA1 ensures that the indifference curves (which are what matter for behavior) are kinked if and only if U is kinked.[8]

TA1. *At any point where U is differentiable, U has non-vanishing first derivative: there is no (π_1, π_2) such that $\partial U/\partial \pi_1 = \partial U/\partial \pi_2 = 0$ at (π_1, π_2).*

Whenever U is *not* purely self-regarding, I impose another technical assumption:

TA2. *If U is not purely self-regarding, then there exist $\underline{\pi}_1 < 0$ and $\underline{\pi}_2 < 0$ such that*

$$\lim_{\pi_2 \to \infty} \sup_{\Delta_1, \Delta_2 > 0} \frac{\frac{U(\underline{\pi}_1, \pi_2 + \Delta_2) - U(\underline{\pi}_1, \pi_2)}{\Delta_2}}{\frac{U(\underline{\pi}_1 + \Delta_1, \pi_2) - U(\underline{\pi}_1, \pi_2)}{\Delta_1}} \leq 0, \textit{ and either } \lim_{\pi_1 \to \infty} \inf_{\Delta_1, \Delta_2 > 0} \frac{\frac{U(\pi_1, \underline{\pi}_2 + \Delta_2) - U(\pi_1, \underline{\pi}_2)}{\Delta_2}}{\frac{U(\pi_1 + \Delta_1, \underline{\pi}_2) - U(\pi_1, \underline{\pi}_2)}{\Delta_1}} \leq 0 \textit{ or } = \infty.$$

TA2 would be satisfied if, as in Example 1 in Section I, $\lim_{\pi_2 \to \infty} \frac{\partial U(\underline{\pi}_1, \pi_2)/\partial \pi_2}{\partial U(\underline{\pi}_1, \pi_2)/\partial \pi_1} = 0$ (caring exclusively about FM) and $\lim_{\pi_1 \to \infty} \frac{\partial U(\pi_1, \underline{\pi}_2)/\partial \pi_2}{\partial U(\pi_1, \underline{\pi}_2)/\partial \pi_1} = \infty$ (caring exclusively about SM), but TA2 also allows either of these limits to be weakly negative (putting negative weight on the player with the very high payoff) and does not assume differentiability. For any given bilateral exchange game, $\underline{\pi}_1$ and $\underline{\pi}_2$ can be chosen to be small enough that TA2 has little economic content, but TA2 helps ensure the existence of optimal actions by helping to make the set of individually-rational transactions compact.

Finally, I normalize the utility levels so that the outside option gives both players zero utility: $U_1(0, 0) = U_2(0, 0) = 0$. As a tie-breaker with the outside option, I assume that if an agent also expects to get zero utility from trading, then the agent chooses to trade.

[7] The reason some models only satisfy weak quasi-concavity is that the utility function is assumed to be piecewise-linear, as in (1). Since piecewise-linearity is clearly intended as a simplifying assumption and does not drive any of the explanatory power of the models for laboratory behavior, adopting quasi-concave versions of these models is consistent with their spirit. In the analysis, quasi-concavity serves mainly as a regularity condition to help ensure uniqueness of optimal behavior.

[8] TA1 is needed because the assumptions are stated in terms of U (rather than made directly on the indifference curves) and because monotonicity will be weakened. When U is monotonic, the interpersonal indifference curves are kinked if and only if U is kinked. However, when U is joint-monotonic, there may be saddle points, (π_1, π_2) with $\partial U/\partial \pi_1 = \partial U/\partial \pi_2 = 0$, where the indifference curves can be kinked even though U is smooth. For example, the function

$$U(x, y) = \begin{cases} x^3 + y^3 & \text{if } x > 0, y > 0 \\ y^3 & \text{if } x > 0, y \leq 0 \\ x^3 & \text{if } x \leq 0, y > 0 \\ x^3 + y^3 & \text{if } x \leq 0, y \leq 0 \end{cases}$$

is twice-continuously differentiable, but has a kinked indifference curve at $U(x, y) = 0$ given by $\min\{x, y\} = 0$.

IV. Reciprocity-Like Behavior in the Bilateral Exchange Game

In this section, partly to build intuition for the efficiency results and partly because it is of independent interest, I show that normality and/or fairness-kinkedness are the properties of distributional-preference that generate reciprocal behavior in bilateral exchange games. I will refer to such behavior as "reciprocity-like" because it is not generated by true reciprocity as modeled, e.g., by Rabin (1993). I define reciprocity-like behavior as follows: SM's optimal response $a_2(a_1)$ to FM's action a_1 is an increasing function of a_1.

For analyzing SM's behavior here and in later sections, it will be useful to introduce notation and terminology for a consumer-theory-like conceptualization of the bilateral exchange game. Denote a material-payoff "consumption bundle" as the vector $\boldsymbol{\pi}(a_1, a_2) \equiv (\pi_1(a_1, a_2), \pi_2(a_1, a_2))$. Given FM's action a_1, SM's choice of action a_2 can be thought of as selecting a pair of material payoffs on the **(material payoff) budget curve** $B(a_1) = \{\boldsymbol{\pi}(a_1, a_2)\}_{a_2 \in \mathbb{R}}$. FM's choice of a_1 can be thought of as a decision of which budget curve to offer to SM. To facilitate the analogy with consumer theory, it is useful to consider the budget *line* that locally approximates the budget curve. At a transaction (a_1, a_2) that identifies a point $(\pi_1(a_1, a_2), \pi_2(a_1, a_2))$ on the budget curve $B(a_1)$, the equation for the budget line is $\pi_1 + p\pi_2 = I$, where $p = p(a_1, a_2) \equiv -\frac{d\pi_1}{d\pi_2}\Big|_{B(a_1)}$ is the local slope of the budget curve—the **price** of π_1 in terms of π_2—and $I = I(a_1, a_2) \equiv \pi_1(a_1, a_2) + p(a_1, a_2)\pi_2(a_1, a_2)$ is the corresponding level of "income" that would allow SM to just "afford" the point on the budget curve. Figure 2 depicts a budget curve and the approximating budget line at SM's optimal action. Finally, I refer to the transaction $(\widehat{a}_1, a_2(\widehat{a}_1))$ as a **fairness-rule optimum** if SM's distributional preferences are fairness-kinked and his optimum occurs on the fairness rule: $\boldsymbol{\pi}(\widehat{a}_1, a_2(\widehat{a}_1)) \in \text{graph}(f)$. This occurs when

$$\lim_{\boldsymbol{\pi} \to \boldsymbol{\pi}(\widehat{a}_1, a_2(\widehat{a}_1)), \boldsymbol{\pi} \in D_f} \left(\frac{\partial U_2(\boldsymbol{\pi})}{\partial \pi_2} - p(\widehat{a}_1, a_2(\widehat{a}_1)) \frac{\partial U_2(\boldsymbol{\pi})}{\partial \pi_1} \right) \geq 0$$

and

$$\lim_{\boldsymbol{\pi} \to \boldsymbol{\pi}(\widehat{a}_1, a_2(\widehat{a}_1)), \boldsymbol{\pi} \in A_f} \left(\frac{\partial U_2(\boldsymbol{\pi})}{\partial \pi_2} - p(\widehat{a}_1, a_2(\widehat{a}_1)) \frac{\partial U_2(\boldsymbol{\pi})}{\partial \pi_1} \right) \leq 0,$$

where these inequalities describe the local slope of SM's indifferences curves in the regions of disadvantageous and advantageous unfairness, respectively, relative to the price at $(\widehat{a}_1, a_2(\widehat{a}_1))$. I call the transaction a **strict fairness-rule optimum** if both of these inequalities are strict.

[FIGURE 2 ABOUT HERE]

Under what conditions is SM's behavior reciprocity-like? It is widely believed that behindness-aversion is the property that enables the inequity-aversion model to generate

such behavior. That is indeed true in the much-studied ultimatum game (Güth, Schmittberger, & Schwarze 1982), in which a second mover can either accept or reject a first mover's offer of some division of \$10. If the second mover rejects, both players get \$0. If the first mover's offer is \$5/\$5, then the second mover will accept the offer because it is just as fair as \$0/\$0 and gives him a higher payoff. In contrast, if the offer would leave the second mover behind, then due to behindness aversion, the second mover may prefer the equal outcome from rejecting, even though both players get a lower payoff.

In bilateral exchange games (including the gift-exchange game and the trust game), however—or more generally, any game where the budget curve is both downward-sloping and continuous—behindness aversion does *not* generate reciprocity-like behavior. This follows from Lemma 1, which shows that as long as SM's distributional preferences are joint-monotonic, then even if they are not monotonic, his behavior is indistinguishable from an agent whose preferences are monotonic.

Lemma 1. *Suppose U_2 is joint-monotonic and quasi-concave. For any a_1, SM has a unique optimal best response, $a_2(a_1)$, that is a continuous function of a_1. Moreover, if U_2 is continuously differentiable at some $(\widehat{a}_1, a_2(\widehat{a}_1))$, then $\partial U_2/\partial \pi_1 > 0$ and $\partial U_2/\partial \pi_2 > 0$ at $(\widehat{a}_1, a_2(\widehat{a}_1))$.*

The lemma states that even if SM's distributional preferences are merely joint-monotonic, as long as his optimum occurs on a smooth region of his indifference curves, his utility at his optimal action will be increasing in both players' material payoffs. Intuitively, SM cannot be optimizing if, at his supposed optimum, he preferred to reduce one of the player's payoffs; since the price of π_1 in terms of π_2 is positive, he would be able to get higher utility by either increasing or reducing his action. Graphically, Figure 2 illustrates that since the budget curve is always downward-sloping in the space of material payoffs, the tangency point with the indifference curve must occur on a downward-sloping region of the indifference curve.

Lemma 1 implies that the generalization from monotonicity to joint-monotonicity for SM is irrelevant for his behavior in a neighborhood of his optimum—and therefore, peeking ahead a bit, for his behavior in a neighborhood of an equilibrium. Even if SM's distributional preferences are fairness-kinked, either his optimum occurs on a smooth region of his indifference curves, in which case the result applies, or his optimum occurs at a kink, in which case the weakening of monotonicity to joint-monotonicity does not matter because non-monotonicities away from the kink are not relevant for behavior.[9]

Rather than behindness aversion, either normality or fairness-kinkedness is a property of distributional preferences that can generate reciprocity-like behavior in the bilateral exchange game, as shown by Proposition 3.

Proposition 3.

[9] In the range of economic settings captured by the bilateral exchange game, Lemma 1 implies that if SM had the option of "punishing" FM for taking a low action by choosing a material payoff pair that is materially-dominated by some point on the budget curve, then (unlike in the ultimatum game) he would never do it. Hence, if such behavior were observed, it would be mistaken to attribute it to SM's distributional preferences and instead should presumably be attributed to negative reciprocity.

1) *Suppose U_2 is joint-monotonic, quasi-concave, and fairness-kinked. Suppose that $(\widehat{a}_1, a_2(\widehat{a}_1))$ is a strict fairness-rule optimum. Then $(a_1, a_2(a_1))$ is a strict fairness-rule optimum for all a_1 in a neighborhood of $\widehat{a}_1$, and $a_2(a_1)$ is increasing in a_1 at $\widehat{a}_1$. Furthermore, U_2 is locally normal in π_1 and π_2 at $(p(\widehat{a}_1, a_2(\widehat{a}_1)); I(\widehat{a}_1, a_2(\widehat{a}_1)))$.*

2) *Suppose $\frac{\partial}{\partial a_1}\left(\frac{\partial \pi_1/\partial a_2}{\partial \pi_2/\partial a_2}\right) \leq 0$ and U_2 is joint-monotonic and quasi-concave. If U_2 is weakly locally normal at $(p(\widehat{a}_1, a_2(\widehat{a}_1)); I(\widehat{a}_1, a_2(\widehat{a}_1)))$, then $a_2(a_1)$ is increasing in a_1 at $\widehat{a}_1$. Hence if U_2 is weakly normal in π_1, then $a_2(a_1)$ is increasing in a_1.*

The first part considers a situation where SM's preferences are fairness-kinked and FM's behavior induces a strict fairness-rule optimum. In that case, if FM slightly increases her action, thereby slightly shifting and changing the slope of the budget curve, then SM's new optimum will occur at another fairness-rule optimum. Since the increase in FM's action increases π_2 but reduces π_1, SM must increase his action in order to keep the players' material payoffs on the fairness rule. A special case of this result has been proved previously for inequity aversion and particular material payoff functions (Fehr, Klein, & Schmidt 2007, p.147).

The first part of Proposition 3 also shows that, although distinct, fairness-kinkedness and normality are related: at a strict fairness-kinked optimum, U_2 is locally normal. If the budget curve shifts outward with the slope unchanged, SM's new optimum will occur at another fairness-rule optimum and hence both players' material payoffs increase.

The second part of Proposition 3 states that when SM's preferences are normal, a sufficient condition for reciprocity-like behavior is $\frac{\partial}{\partial a_1}\left(\frac{\partial \pi_1/\partial a_2}{\partial \pi_2/\partial a_2}\right) \leq 0$. This condition means that an increase in FM's action weakly lowers the price for SM of increasing FM's payoff. This assumption is satisfied if, as in trust game experiments, both players' material payoff functions are additively-separable in the actions.[10] It is also satisfied by the material payoff functions typically used in gift-exchange game experiments, and indeed Fehr, Kirchsteiger, & Riedl (1998, pp.7-8) prove the result for this case.[11]

The intuition for Part 2 of the proposition can be understood in terms of income and substitution effects on the material-payoff "consumption bundle" that are induced by a small increase in FM's action. Since FM's material payoff becomes cheaper—due to the condition $\frac{\partial}{\partial a_1}\left(\frac{\partial \pi_1/\partial a_2}{\partial \pi_2/\partial a_2}\right) \leq 0$—the substitution effect gives SM an incentive to increase π_1 relative to π_2, and therefore to increase his action. If the income effect is positive, then since SM's distributional preferences are normal, SM's incentive to increase π_1 is reinforced, and consequently SM prefers to increase his action. If instead the income effect

[10] More generally than additive-separability, the assumption is satisfied if the actions enter the material payoff functions as complements in the sense that the transformed material payoff functions, $\widetilde{\pi}_1(a_1, a_2)$ and $\widetilde{\pi}_2(a_1, a_2)$, defined by $\widetilde{\pi}_1(a_1, a_2) \equiv \pi_1(-a_1, a_2)$ and $\widetilde{\pi}_2(a_1, a_2) \equiv \pi_1(a_1, -a_2)$, are both weakly supermodular in (a_1, a_2).

[11] Specifically, following Fehr, Kirchsteiger, & Riedl (1993), in order to rule out negative payoff values, gift-exchange experiments typically use as material payoff functions: $\pi_1(a_1, a_2) = (k_1 - a_1)a_2$ and $\pi_2(a_1, a_2) = a_1 - c(a_2) - k_2$, where $c(\cdot)$ is increasing and strictly convex, $k_1 > 0$ and k_2 are constants, and $a_1 \leq k_1$ and $a_2 \geq 0$ have restricted domain.

is negative, then both the substitution effect and income effect give SM an unambiguous incentive to decrease π_2, which again makes him prefer to increase his action.[12]

V. Characterizing Efficient Transactions

In this section I address what is meant by an "efficient" transaction when agents have distributional preferences. There are two possible generalizations of Pareto efficiency, depending on whether the players' welfare is measured by material payoffs or by utilities:

Definition 6. *A transaction (a_1, a_2) is **utility Pareto efficient (UPE)** if there is no other transaction $(\widehat{a}_1, \widehat{a}_2)$ such that $U_1(\boldsymbol{\pi}(\widehat{a}_1, \widehat{a}_2)) \geq U_1(\boldsymbol{\pi}(a_1, a_2))$ and $U_2(\boldsymbol{\pi}(\widehat{a}_1, \widehat{a}_2)) \geq U_2(\boldsymbol{\pi}(a_1, a_2))$, at least one inequality strict.*

Definition 7. *A transaction (a_1, a_2) is **materially Pareto efficient (MPE)** if there is no other transaction $(\widehat{a}_1, \widehat{a}_2)$ such that $\pi_1(\widehat{a}_1, \widehat{a}_2) \geq \pi_1(a_1, a_2)$ and $\pi_2(\widehat{a}_1, \widehat{a}_2) \geq \pi_2(a_1, a_2)$, at least one inequality strict.*

If a transaction (a_1, a_2) is MPE, then I will also refer to the resulting material payoff pair $\boldsymbol{\pi}(a_1, a_2)$ as MPE; analogously for UPE. A transaction is MPE if and only if at that transaction, the material-payoff marginal rates of substitution are equal: $\frac{\partial \pi_1(a_1,a_2)/\partial a_1}{\partial \pi_1(a_1,a_2)/\partial a_2} = \frac{\partial \pi_2(a_1,a_2)/\partial a_1}{\partial \pi_2(a_1,a_2)/\partial a_2}$. In general, the level of a_1 that corresponds to an MPE transaction depends on a_2. By discussing Pareto efficiency exclusively in terms of monetary payoffs, analyses of laboratory experiment have implicitly focused on MPE.

Which generalization of Pareto efficiency is the right social welfare criterion? If the U's represent the players' "true" preferences, then UPE is appropriate. However, if fair-minded behavior is caused by (unmodeled) social pressure and the U's are a reduced-form representation of the resulting behavior, then the π's may actually represent the players' "true" preferences. In that case, MPE is the appropriate welfare criterion.[13]

To characterize MPE and UPE and their relationship to each other, a few definitions will be useful. Let

$$(\overline{a}_1, \overline{a}_2) \equiv \arg\max_{(a_1,a_2)} U_1(\boldsymbol{\pi}(a_1, a_2))$$

be called **FM's favorite transaction**, her most-preferred transaction among the feasible transactions. I will sometimes also call the resulting material payoff pair, $(\overline{\pi}_1, \overline{\pi}_2) \equiv \boldsymbol{\pi}(\overline{a}_1, \overline{a}_2)$, FM's favorite transaction. Let

$$\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right) \equiv \arg\max_{(a_1,a_2)} U_2(\boldsymbol{\pi}(a_1, a_2))$$

be called **SM's favorite transaction**, his most-preferred transaction among the feasible transactions, with corresponding material payoff pair $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$. Theorem 1 describes

[12] In terms of the notation defined in Section 6, the income effect is positive when the small increase in FM's action occurs from a level below $\overline{\overline{a}}_1$, and the income effect is negative when FM's initial action is above $\overline{\overline{a}}_1$.

[13] Sen (1973) and Köszegi & Rabin (2008) similarly argue that in some situations, behavior—as represented by the U's—may not be the correct basis for judging welfare. For example, Sen (1973, pp.253-254) writes: "mores and rules of behaviour drive a wedge between behaviour and welfare...basing normative criteria, e.g., Pareto optimality, on [behaviour-derived] *as if* preferences poses immense difficulties."

the relationship between MPE and UPE when FM has monotonic distributional preferences (see Web Appendix A for the more general case where FM's preferences are joint-monotonic).[14]

Theorem 1. *Suppose U_1 is monotonic and quasi-concave, and suppose U_2 is joint-monotonic and quasi-concave. FM's and SM's favorite transactions, $(\overline{a}_1, \overline{a}_2)$ and $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$, exist and are unique. The set of UPE material payoff pairs coincides exactly with the set of material payoff pairs on the MPE frontier between $(\overline{\pi}_1, \overline{\pi}_2)$ and $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$.*

Figure 3 illustrates that the set of UPE material payoff pairs is the subset of the MPE frontier between $(\overline{\pi}_1, \overline{\pi}_2)$ and $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$. (The figure is drawn with $\overline{\pi}_1 > \overline{\overline{\pi}}_1$ and $\overline{\pi}_2 < \overline{\overline{\pi}}_2$, but the theorem also holds if these inequalities are reversed.) To understand why the theorem is true, first note that any UPE material payoff pair must be MPE: for any non-MPE material payoff pair, there is an alternative material payoff pair that gives more to both players that SM prefers because his preferences are joint-monotonic, and FM prefers because her preferences are monotonic. Next, note that for any two material payoff pairs on the MPE frontier, each player prefers the pair closer to his favorite transaction. Therefore, a pair on the frontier that gives higher material payoff to FM than $\overline{\pi}_1$ cannot be UPE because *both* players prefer $(\overline{\pi}_1, \overline{\pi}_2)$. Similarly, a pair on the frontier that gives higher material payoff to SM than $\overline{\overline{\pi}}_2$ cannot be UPE because both players prefer $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$.[15]

[FIGURE 3 ABOUT HERE]

While there are many MPE transactions, Lemma 2 shows a surprising result: in the bilateral exchange game, SM's favorite transaction is the *only* MPE transaction that is possible for FM to induce! As above, let $a_2(a_1)$ denote SM's best-response function.

Lemma 2. *Suppose U_2 is joint-monotonic and quasi-concave. Then there exists a unique $\widehat{a}_1$ such that the resulting transaction $(\widehat{a}_1, a_2(\widehat{a}_1))$ is MPE. This transaction is SM's favorite transaction $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$, and it is UPE.*

To understand Lemma 2, note that at any MPE material payoff pair where SM's action is a best response, SM's indifference curve must be tangent to the MPE frontier (as shown in Figure 4 below). Such a tangency point must be SM's favorite transaction. Given this result, I will hereafter refer to SM's favorite transaction as "the" efficient transaction.

[14] Dufwenberg, Heidhues, Kirchsteiger, Riedel, & Sobel (2011) independently prove a result related to Theorem 1; their Theorem 3 implies that when at least one player has monotonic distributional preferences, material Pareto efficiency is a necessary condition for utility Pareto efficiency. I discuss the relationship between Theorem 1 and their result in more detail in Web Appendix A.

[15] If one or both of the players is purely self-regarding, then Theorem 1 does not technically apply but extends straightforwardly: The set of UPE material payoff pairs remains coincident with the set of material payoff pairs on the MPE frontier between $(\overline{\pi}_1, \overline{\pi}_2)$ and $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$, but depending on which player is purely self-regarding, $(\overline{\pi}_1, \overline{\pi}_2) \equiv (\infty, -\infty)$, $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right) \equiv (-\infty, \infty)$, or both.

VI. Necessary Conditions for An Efficient Equilibrium

This section describes necessary conditions for the efficient transaction to be the equilibrium of the bilateral exchange game. The main result will be that there are essentially only two cases: one involving SM's action being a locally linear transfer of material payoffs, and the other involving SM's distributional preferences being fairness-kinked.

As an initial step, Lemma 3 establishes that under the maintained assumptions TA1 and TA2, an equilibrium of the game exists.

Lemma 3. *An equilibrium exists. Moreover, if* $U_1\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right) \geq 0$, *then an equilibrium exists in which the players exchange rather than taking their outside options.*

The equilibrium will involve FM choosing her outside option if SM's optimal response to every possible a_1 resulted in negative utility for FM. Lemma 3 states that a sufficient condition for trade to occur in equilibrium is that FM prefers SM's favorite transaction to her own outside option: $U_1\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right) \geq 0$. This condition is sufficient because, from Lemma 2, there exists an action for FM that induces SM's favorite transaction.

As another preliminary step, Proposition 4 states formally a corollary of Lemma 2: SM's favorite transaction is the only candidate for an equilibrium that is MPE.

Proposition 4. *Suppose* U_1 *and* U_2 *are joint-monotonic and quasi-concave. If the equilibrium* $(a_1, a_2(a_1))$ *is MPE, then* $(a_1, a_2(a_1))$ *is SM's favorite transaction, and* $U_1(\pi(a_1, a_2(a_1))) \geq 0$.

Proposition 4 additionally states that, besides being a sufficient condition for trade to occur, $U_1\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right) \geq 0$ is also a necessary condition for the equilibrium to be MPE.

If FM is self-regarding, then the condition $U_1\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right) \geq 0$ has a straightforward interpretation: SM's distributional preferences involve sufficient positive regard for FM that SM's favorite transaction is better for FM than not trading. If instead SM were too selfish, then $\overline{\overline{\pi}}_1$ would be so small that FM would prefer her outside option to $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$. Later, when providing sufficient conditions for an efficient equilibrium, $U_1\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right) \geq 0$, as well as the other necessary conditions, will be maintained assumptions. Although $U_1\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right) \geq 0$ is not an assumption directly on primitives, it is a straightforward condition to check once the players' material payoff functions and distributional preferences have been specified.

Theorem 2 is a central result of this paper. It states that a necessary condition for the equilibrium to be efficient is that at least one of three possibilities must be true.

Theorem 2. *Suppose* U_1 *and* U_2 *are joint-monotonic and quasi-concave, and* U_2 *is either twice-continuously differentiable or fairness-kinked. If the equilibrium* $(a_1, a_2(a_1))$ *is MPE, then at least one of the following must be true:*

1) $(a_1, a_2(a_1))$ *is FM's favorite transaction.*

2) $dp(a_1, a_2(a_1))/da_1 = 0$.

3) *U_2 is fairness-kinked, and $(a_1, a_2(a_1))$ is a fairness-rule optimum.*

Possibility (1) (the least interesting) is that FM and SM share the same favorite transaction. That transaction would then be the equilibrium, and it would be efficient. Possibility (2) is that FM's action does not affect the slope of the budget curve at the equilibrium transaction. Possibility (3) is that SM's indifference curve is fairness-kinked at the equilibrium transaction.

Once possibility (1) is excluded, to understand why possibilities (2) and (3) are the only situations where the equilibrium could be MPE, consider a deviation by FM from her equilibrium action to some alternative action. Figure 4 illustrates, but instead of showing the budget curves that SM actually faces at the original, equilibrium material-payoff pair and the new point, it shows the budget lines that approximate the budget curves.

[FIGURE 4 ABOUT HERE]

SM's response to the change in the budget line can be characterized by the Slutsky decomposition into an income effect and a substitution effect. The magnitude of the income effect depends on how much the budget line shifts due to the change in FM's action, holding constant the original, equilibrium price. Since the original budget line is tangent to the MPE frontier, FM's original action is the action that maximizes income at the original price; hence if FM's deviation is small, then by the envelope theorem, the income effect is second order.

Since the income effect is second order, the substitution effect must equal zero. Otherwise, by marginally deviating from the equilibrium action, FM could cause SM to choose a material payoff pair that either—depending on the direction FM chooses to deviate—gives FM a higher material payoff and SM a lower material payoff than at the original material payoff pair, or vice-versa. Since FM's favorite transaction does not coincide with SM's favorite transaction, FM would prefer one of these over the original material payoff pair, violating the assumption that the original action was an equilibrium.

Possibilities (2) and (3) correspond to the two possible ways that the substitution effect can equal zero. The budget lines may locally be parallel shifts, in which case there is no change in relative price; that is (2). Alternatively, the optimal material payoff pair may occur at a kink in SM's indifference curves, in which case SM's optimal pair does not change in response to a Slutsky-compensated change in price; because any kink must be on the fairness rule by assumption, that situation is (3).

I have stated possibility (2) as $dp(a_1, a_2(a_1))/da_1 = 0$ in order to make transparent its link to the intuition that the substitution effect is zero. Yet, as stated, it raises the question: for what material payoff functions is it satisfied? In a paper about the special case of the rotten kid theorem, Dijkstra (2007, his Lemma 1) answered this question: $dp(a_1, a_2(a_1))/da_1 = 0$ at SM's favorite transaction if and only if the material payoff

functions are **locally conditionally transferable** at SM's favorite transaction, i.e., in a neighborhood of $(\overline{\overline{a}}_1, \overline{\overline{a}}_2)$, $\frac{\partial \pi_1(a_1,a_2)/\partial a_2}{\partial \pi_2(a_1,a_2)/\partial a_2} = -k$ for some constant $k > 0$.[16]

The fact that possibility (2) corresponds to locally parallel shifts of the budget curves makes clear why normality of SM's distributional preferences will play an important role. In fact, under possibility (2), if FM is purely self-regarding, local normality of U_2 in π_1 at SM's favorite transaction is another necessary condition for the equilibrium to be MPE.

VII. Sufficient Conditions for An Efficient Equilibrium

The previous section showed that there are exactly two interesting cases in which the equilibrium could be efficient: (1) the budget lines that approximate the budget curves are parallel shifts, or (2) SM's interpersonal indifference curve is kinked at the equilibrium. This section explores these cases in more detail, giving sufficient conditions for the equilibrium to be efficient in each case.

The intuition in both cases is fundamentally the same: SM's behavior aligns the players' material incentives by ensuring that the players' material payoffs increase or decrease together as FM varies her action. FM will choose the action that maximizes both players' material payoffs if FM's distributional preferences are monotonic, leading to an efficient equilibrium.

A. Efficient Case I: Budget Curves Are Parallel Shifts

As discussed in Section VI, the budget curves are parallel shifts locally if and only if the material payoff functions are locally conditionally transferable. In that case, as long as U_2 is locally normal, both players' material payoffs increase or decrease together as FM varies her action. If these conditions hold in a neighborhood of the efficient transaction, then the action that generates the efficient outcome will be a local optimum for FM. Global analogs of the local assumptions ensure that the players' material incentives are aligned over the entire range of FM's possible actions.

The natural condition to guarantee that the budget curves facing SM are parallel shifts everywhere is that the material payoff functions are **globally conditionally transferable**: for some functions G, H, and Z and constant $k > 0$, $\pi_1(a_1, a_2) = -G(a_1) + Z(a_1, a_2)$ and $\pi_2(a_1, a_2) = H(a_1) - kZ(a_1, a_2)$. If so, and if FM is purely self-regarding or has monotonic distributional preferences, then (global) normality of U_2 is sufficient to ensure that the equilibrium is unique and occurs at the efficient transaction.[17]

[16] In an influential paper, Bergstrom (1989) argued but did not prove that global conditional transferability, as defined in Section VII, is necessary for possibility (2). Dijkstra's (2007) result shows that that conjecture was incorrect. Dijkstra's "Condition 2" characterizes exactly the class of material payoff functions that is locally conditionally transferable at $(\overline{\overline{a}}_1, \overline{\overline{a}}_2)$, but the condition is difficult to interpret. Here I provide an intuitive example of material payoff functions that are not globally conditionally transferable but that are locally conditionally transferable at $(\overline{\overline{a}}_1, \overline{\overline{a}}_2)$. Consider $\pi_1(a_1, a_2) = Z(a_1, a_2)$ and $\pi_2(a_1, a_2) = H(a_1) - F(Z(a_1, a_2))$, where $F' > 0$ and $F'' \neq 0$. These material payoff functions could describe a setting where an investor (FM) invests an amount of money a_1 and pays a trustee (SM) an amount $H(a_1)$ to oversee the investment, and then the trustee allocates the accumulated capital between the investor and himself by choice of a_2.

[17] If FM is purely self-regarding, the assumption of normality of U_2 can be weakened to normality of U_2 in π_1.

Theorem 3. *Suppose U_2 is joint-monotonic, quasi-concave, and normal. Suppose the material payoff functions are globally conditionally transferable. If U_1 is monotonic or purely self-regarding, and if $U_1\left(\boldsymbol{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\right) \geq 0$, then the unique equilibrium transaction is the efficient transaction $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$.*

Figure 5 illustrates Theorem 3.

[FIGURE 5 ABOUT HERE]

For fixed material payoff functions, as long as the specified assumptions of the theorem hold, the conclusion does not depend on exactly how selfish or altruistic SM is, or whether U_2 is kinked or smooth; FM will choose the same action in any case, since with globally conditionally transferable material payoffs, there is a unique efficient a_1 such that the budget curve coincides with the MPE frontier. Thus, loosely speaking (since there is no uncertainty in the model), as Becker (1974) notes in the context of the rotten kid theorem, FM would choose the efficient action even if she were uncertain about SM's distributional preferences and hence uncertain about exactly which action SM will choose.

A special class of material payoff functions that satisfies global conditional transferability is quasi-linearity in a_2: $\pi_1(a_1, a_2) = -G(a_1) + a_2$ and $\pi_2(a_1, a_2) = H(a_1) - ka_2$ (as in Example 1 from Section I). These material payoff functions are often used to model situations where *SM's action* is a monetary transfer. This would describe settings where FM is a seller who provides a service, and SM is a (fair-minded) customer who decides how much to pay for the service. Quasi-linearity in a_2 would *not* describe environments where FM's action is a transfer of money to SM, such as when FM is a profit-maximizing employer who pays a wage, and SM is a (fair-minded) worker who exerts effort. Therefore, Theorem 3 could apply in the former case but not the latter.

B. Efficient Case II: SM's Distributional Preferences Are Fairness-Kinked

The logic for how fairness-kinked distributional preferences can lead to an efficient equilibrium requires that the efficient transaction be a strict fairness-rule optimum. In that case, as shown in Proposition 3, SM behaves in accordance with the fairness rule for any small change in FM's action. As long as FM is self-regarding or has monotonic distributional preferences, this condition ensures that the efficient transaction is a local optimum for both players.

To ensure that the efficient outcome is the equilibrium, a natural approach would be to write down sufficient conditions for SM's optimum to occur on the fairness rule for *any* action by FM. Unfortunately, such conditions would probably have to be quite strong. For example, if there are no restrictions on the shape of the budget curves, then in order to ensure that SM's optimum occurs at a kink, both the advantageously unfair and

disadvantageously unfair portions of his indifference curves would have to be upward-sloping. This would mean that SM cares so much about fairness that, starting from any fair transaction, he would never prefer to increase just one player's material payoff.

Instead, I seek sufficient conditions that are not implausibly restrictive *and* relatively straightforward to check. Analogous to the $\sigma < \overline{\sigma}$ assumption in Example 2 from Section I, the idea of the sufficient conditions is to ensure that SM is not so generous in the region of disadvantagenous inequality that FM can earn higher utility by deviating to a low action. With piecewise-linear distributional preferences for SM and with a purely self-regarding FM, making the single parameter σ sufficiently small sufficed. Here, several assumptions are needed to do the same job.

Let $(\widehat{a}_1, \widehat{a}_2)$ denote the (unique) transaction satisfying $\pi_1(\widehat{a}_1, \widehat{a}_2) = \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$, $U_2\left(\boldsymbol{\pi}(\widehat{a}_1, \widehat{a}_2)\right) = 0$, and $\widehat{a}_1 < \overline{\overline{a}}_1$. That is, $\widehat{a}_1$ is the smallest action that keeps SM from taking his outside option and that could possibly give FM a material payoff of at least $\pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. I assume:

S1. *SM's distributional preferences are "sufficiently kinked" at the efficient transaction:* $\lim_{\boldsymbol{\pi}\to\boldsymbol{\pi}\left(\overline{\overline{a}}_1,\overline{\overline{a}}_2\right),\boldsymbol{\pi}\in D_f}\left(\frac{\partial U_2(\boldsymbol{\pi})}{\partial \pi_2} - p(\overline{\overline{a}}_1, \widehat{a}_2)\frac{\partial U_2(\boldsymbol{\pi})}{\partial \pi_1}\right) > 0$.

S2. *$\pi_1(a_1, a_2)$ and $\pi_2(a_1, a_2)$ are each additively separable in the actions.*

S3. *U_2 is normal.*

S4. *FM gets higher material payoff from her own favorite transaction than from SM's favorite transaction:* $\pi_1(\overline{a}_1, \overline{a}_2) > \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$.

S5. *U_1 is weakly quasi-concave.*

S1 means that if FM chose $\widehat{a}_1$, SM's optimal response would give FM a lower material payoff than $\pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. Combined with S2 and S3, it implies that FM would also earn a lower material payoff than $\pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ for any action between $\widehat{a}_1$ and $\overline{\overline{a}}_1$. Specifically, as FM's action increases, S2 implies that the cost (in units of π_2) to SM of choosing an action that yields $\pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ is rising, and S3 ensures that SM's willingness to pay for π_1 is falling.[18] If FM is purely self-regarding, then S1-S3 are sufficient to ensure that $\overline{\overline{a}}_1$ is FM's global optimum. If FM has monotonic distributional preferences, however, then it is possible that FM could prefer to deviate to an action that gives her a *lower* material payoff. S4 and S5 are realistic assumptions that together rule out that possibility.

Theorem 4. *Suppose U_2 is joint-monotonic, quasi-concave, and fairness-kinked. Assume S1-S5. If U_1 is monotonic or purely self-regarding, if $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ is a strict fairness-rule optimum, and if $U_1\left(\boldsymbol{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\right) \geq 0$, then the unique equilibrium transaction is the efficient transaction $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$.*

[18] While S3 is exactly what is needed to ensure that SM's willingness to pay for π_1 is falling (see footnote 4), S2 seems stronger than what is required, but I do not know if a less restrictive assumption will suffice.

I emphasize that while S3 imposes normality, its role in Theorem 4 is to help rule out that other actions give FM a higher material payoff than $\overline{\overline{a}}_1$; normality is not required for the fundamental logic, described at the beginning of this subsection, for how SM behaving in accordance with a fairness rule aligns the players' material incentives. Figure 6a illustrates the efficient equilibrium when SM has fairness-kinked distributional preferences, and Figure 6b shows a way the equilibrium could fail to be efficient if the assumptions of Theorem 4 are not satisfied.

[FIGURE 6 ABOUT HERE]

Unlike Theorem 3, Theorem 4 does not require the material payoff functions to be locally conditionally transferable and so applies to non-monetary trades, such as barter or exchange of favors. Moreover, as long as SM adheres to *some* fairness rule, the equilibrium will be efficient, even if the fairness rule is non-linear or self-serving.

I suggested above that Theorem 3 would hold even if FM were uncertain about exactly what SM's distributional preferences are. Theorem 4, in contrast, requires that FM know what fairness rule SM is following. Otherwise, FM would not know which action would induce SM's favorite transaction. Therefore, loosely speaking, there is "social value" in having SM's fairness rule be common knowledge. Social norms like 50-50 splits or other fairness conventions may serve the function of being fairness rules that are common knowledge.

VIII. Discussion

This paper gives conditions under which distributional preferences alone give rise to efficient exchange. However, in one-shot interactions, efficient exchange is usually thought to be enabled by contracts. Therefore, the results in this paper raise the question: why do people so often write contracts? I conclude by briefly discussing four answers that may be fruitful avenues for research.

One answer suggested from within the logic of the model is that FM *prefers* a contract, even when the equilibrium of the bilateral exchange game would be efficient. Suppose a contract implements the Nash bargaining solution. It will select a UPE transaction that is in between FM's favorite transaction and SM's favorite transaction, depending on the agents' relative bargaining power. At any of these transactions, FM gets higher utility than she does at SM's favorite transaction. Therefore, FM would *always* be better off with a contract rather than relying on SM's distributional preferences, as long as writing and enforcing a contract is not too costly.

A second answer may be that FM is uncertain about SM's distributional preferences. Indeed, Fehr & Schmidt (1999) argue that heterogeneity in distributional preferences and the resulting asymmetric information helps explain behavior in experiments. However, as noted in the discussions after Theorems 3 and 4, FM's uncertainty regarding some features of SM's preferences do *not* matter for FM's action and the efficiency of equilibrium. Nonetheless, the overall degree of SM's non-selfishness can matter; recall that

SM's favorite transaction being sufficiently generous was a maintained assumption in the analysis (see the discussions following Lemma 3 and Proposition 4).

The asymmetric-information game cannot be analyzed in full generality without assumptions about distributional preferences under uncertainty. The essential logic for how uncertainty regarding selfishness may reduce efficiency, however, can be seen in a simple example. As in Example 2 in Section I, suppose that $\pi_1(a_1, a_2) = a_2 - a_1$ and $\pi_2(a_1, a_2) = a_1 - c(a_2)$. Now suppose FM is a risk-neutral, profit-maximizing firm; and SM is purely self-regarding with probability $1 - p$, and behaves in accordance with the equal-split fairness rule with probability p: choosing $a_2(a_1)$ to satisfy $\pi_1(a_1, a_2) = \pi_2(a_1, a_2)$. Assume that $a_2 \in [0, \infty)$ so that if SM is self-regarding, then $a_2(a_1) \equiv 0$. In equilibrium, FM's first-order condition solves $c'(a_2(a_1)) = 2p - 1$. Thus, if $p < 1$, SM's equilibrium action falls short of the efficient level. Intuitively, by choosing a lower level of a_1, FM can get some of the gains from trade when SM turns out to be fair-minded while insuring against losing too much if SM turns out to be self-regarding.

A third possible reason to write contracts (instead of relying on distributional preferences to generate efficiency) is that more complex mechanisms of other-regarding behavior that are left out of the model—such as signaling (e.g., Andreoni & Bernheim 2009) and intentions-based reciprocity (e.g., Rabin 1993)—might cause the efficiency predictions to break down. For example, Netzer & Schmutzler (2013) study a bilateral exchange game in which FM is purely self-regarding, and SM puts positive weight on FM's material payoff only to the extent he believes FM has behaved kindly toward him. They argue that when FM is purely self-regarding and SM's behavior is driven by such intentions-based reciprocity, the equilibrium is generically materially Pareto *in*efficient because SM is unwilling to reciprocate high actions by FM, which are interpreted as attempts at material-payoff maximization rather than as kindness. It is unclear whether this conclusion would hold in the more general case where intentions-based reciprocity operates in combination with distributional preferences, as in Falk & Fischbacher (2006).[19]

A fourth answer is that key assumptions underlying the efficiency results may be faulty. One such assumption is that distributional preferences are defined over material payoffs (rather than, say, separately over monetary payoffs and non-monetary payoffs). The combination of this assumption with the assumption of either normality or fairness-kinkedness has an implication that fundamentally underlies the efficiency results: SM's strategy ensures that FM's material payoff *net* of the cost incurred by her choosing a higher action is increasing in the efficiency of the action. This implication has major ramifications even beyond those that I have focused on. For example, it means that distributional preferences alone can completely solve the hold-up problem! When both players are purely self-regarding, the hold-up problem arises when FM and SM bargain over surplus after FM has already incurred the sunk cost of investing to generate the surplus, and thus FM is forced to share the *gross* returns with SM. Anticipating this, it may

[19] Moreover, Charness & Rabin (2002) argue that intentions-based reciprocity becomes operative only in response to a first-mover's unkind behavior, while distributional preferences alone govern a second-mover's behavior when FM has behaved kindly. If so, then the analysis in this paper applies without modification to bilateral exchange settings where both parties are gaining from the transaction because the intentions-based reciprocity is never activated.

not be profitable for FM to make a socially efficient investment. However, if FM and SM share the *net* returns due to SM's reciprocity-like behavior, then the hold-up problem disappears.

To assess how well these assumptions—distributional preferences defined over material payoffs, normality, and fairness-kinkedness—approximate reality, each should be tested empirically. Their implication—that FM's material payoff is increasing in net returns—can also be examined empirically because it means that SM's action will depend not only on the benefit that SM receives from FM's action but also on the cost incurred by FM. Consider two situations. In both, FM provides a service where higher quality requires higher effort, and then SM chooses how much to tip. In the second situation, it is more costly for FM to provide any given level of effort, but the two situations are otherwise identical, with the same efficient level of effort and resulting quality. A testable prediction of the model is that SM would tip more in the second situation. An alternative hypothesis, which also seems natural, is that SM's tip depends only on the quality of service, and so SM would tip the same amount in both situations. I am not aware of existing evidence that tests these conflicting hypotheses

REFERENCES

Andreoni, James, and B. Douglas Bernheim. 2009. "Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects." *Econometrica*, 77(5): 1607–1636.

Andreoni, James, and John Miller. 2002. "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism." *Econometrica*, 70(2): 737–753.

Bazerman, Max H., George F. Loewenstein, and Sally Blount White. 1992. "Reversals of Preference in Allocation Decisions: Judging an Alternative Versus Choosing Among Alternatives." *Administrative Science Quarterly*, 37(2): 220–240.

Becker, Gary S. 1974. "A Theory of Social Interactions." *Journal of Political Economy*, 82(6): 1063–93.

Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. "Trust, Reciprocity, and Social History." *Games and Economic Behavior*, 10: 122–142.

Bergstrom, Theodore C. 1989. "A fresh look at the Rotten Kid theorem – and other household mysteries." *Journal of Political Economy*, 97(5): 1138–59.

Bolton, Gary E., and Axel Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review*, 90(1): 166–193.

Bolton, Gary E., and Axel Ockenfels. 2006. "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments: Comment." *American Economic Review*, 96(5): 1906–1911.

Camerer, Colin F. 2003. *Behavioral Game Theory.* Princeton, NJ:Princeton University Press.

Charness, Gary, and Brit Grosskopf. 2001. "Relative payoffs and happiness: an experimental study." *Journal of Economic Behavior and Organization*, 45: 301–328.

Charness, Gary, and Matthew Rabin. 2002. "Understanding Social Preferences With Simple Tests." *Quarterly Journal of Economics*, 117(3): 817–869.

Coase, Ronald H. 1960. "The problem of social cost." *Journal of Law and Economics*, 3: 1–44.

Cox, James C., and Vjollca Sadiraj. 2010. "Direct Tests of Individual Preferences for Efficiency and Equity." *Economic Inquiry*.

Cox, James C., Daniel Friedman, and Vjollca Sadiraj. 2008. "Revealed Altruism." *Econometrica*, 76(1): 31–69.

Dijkstra, Bouwe R. 2007. "Samaritan versus Rotten Kid: Another Look." *Journal of Economic Behavior and Organization*, 64: 91–110.

Dufwenberg, Martin, Paul Heidhues, Georg Kirchsteiger, Frank Riedel, and Joel Sobel. 2011. "Other-Regarding Preferences in General Equilibrium." *Review of Economic Studies*, 78: 640–66.

Engelmann, Dirk, and Martin Strobel. 2004. "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments." *American Economic Review*, 94(4): 857–869.

Falk, Armin, and Urs Fischbacher. 2006. "A theory of reciprocity." *Games and Economic Behavior*, 54: 293–315.

Fehr, Ernst, Alexander Klein, and Klaus M. Schmidt. 2007. "Fairness and Contract Design." *Econometrica*, 75(1): 121–154.

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*, 114(3): 817–868.

Fehr, Ernst, and Klaus M. Schmidt. 2004. "The Role of Equality, Efficiency, and Rawlsian Motives in Social Preferences: A Reply to Englemann and Strobel." University of Zurich Institute for Empirical Research in Economics Working Paper Number 179.

Fehr, Ernst, and Klaus Schmidt. 2003. "Theories of Fairness and Reciprocity: Evidence and Economic Applications." In *Advances in Economic Theory, Eighth World Conference of the Econometric Society, Vol. 1.* , ed. M. Dewatripont, L.P. Hansen and S. Turnovski, 208–257. Cambridge, U.K.: Cambridge University Press.

Fehr, Ernst, Georg Kirchsteiger, and Arno Riedl. 1993. "Does Fairness Prevent Market Clearing? An Experimental Investigation." *Quarterly Journal of Economics*, 108(2): 437–459.

Fehr, Ernst, Georg Kirchsteiger, and Arno Riedl. 1998. "Gift Exchange and Reciprocity in Competitive Experimental Markets." *European Economic Review*, 42: 1–34.

Fehr, Ernst, Michael Naef, and Klaus M. Schmidt. 2006. "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments: Comment." *American Economic Review*, 96(5): 1912–1917.

Fisman, Raymond, Shachar Kariv, and Daniel Markovits. 2007. "Individual Preferences for Giving." *American Economic Review*, 97(5): 1858–1876.

Fudenberg, Drew, and Eric Maskin. 1986. "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information." *Econometrica*, 54(3): 533–554.

Güth, Werner, R. Schmittberger, and B. Schwarze. 1982. "An Experimental Analysis of Ultimatum Bargaining." *Journal of Economic Behavior and Organization*, 3: 367–388.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. 1986. "Fairness as a Constraint on Profit Seeking Entitlements in the Market." *American Economic Review*, 76(4): 728–41.

Kritikos, Alexander, and Friedel Bolle. 2001. "Distributional concerns: equity- or efficiency-oriented." *Economics Letters*, 73: 333–338.

Köszegi, Botond, and Matthew Rabin. 2008. "Choices, situations, and happiness." *Journal of Public Economics*, 92: 1821–1832.

Netzer, Nick, and Armin Schmutzler. 2013. "Explaining Gift-Exchange – The Limits of Good Intentions." University of Zurich Socioeconomic Institute Working Paper No. 0919.

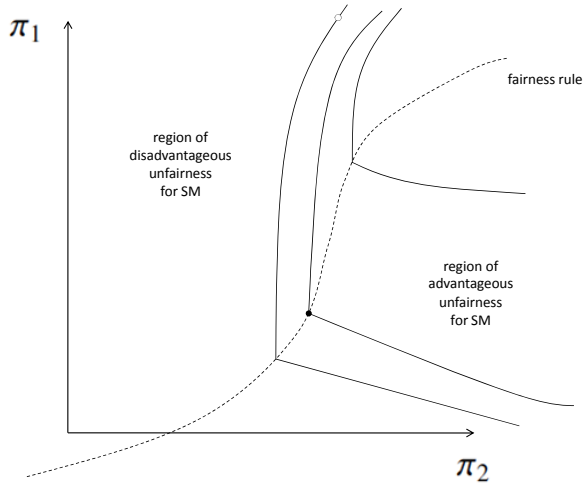
Pelligra, Vittorio, and Luca Stanca. 2013. "To Give or Not To Give? Equity, Efficiency and Altruistic Behavior in an Artefactual Field Experiment." *Journal of Socio-Economics*, 46: 1–9.

Quah, John K.-H. 2007. "Supplement to 'The Comparative Statics of Constrained Optimization Problems'." *Econometrica*, 75(2): 401–431.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review*, 83(5): 1281–1302.

Sen, Amartya. 1973. "Behaviour and the Concept of Preference." *Economica*, 40(159): 241–259.

[1a]



[1b]

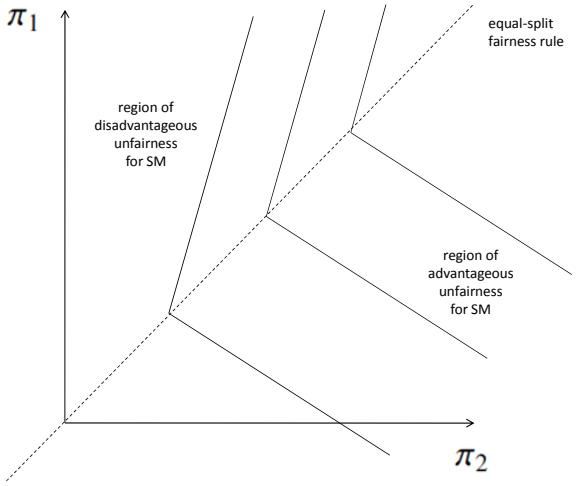


Figure 1. Interpersonal indifference curves. Panel (a): Fairness-kinked preferences. The fairness rule is an upward-sloping locus of material payoff pairs. The indifference curves are kinked at each material payoff pair on the fairness rule. These particular preferences are joint-monotonic but not monotonic; due to the non-monotonicity, the black point is preferred to the white point that gives higher material payoffs to both players. Panel (b): Inequity-averse preferences. The fairness rule is the set of 50-50 splits, and the indifference curves are piecewise-linear. Inequity-averse preferences are joint-monotonic but not monotonic.

[2]

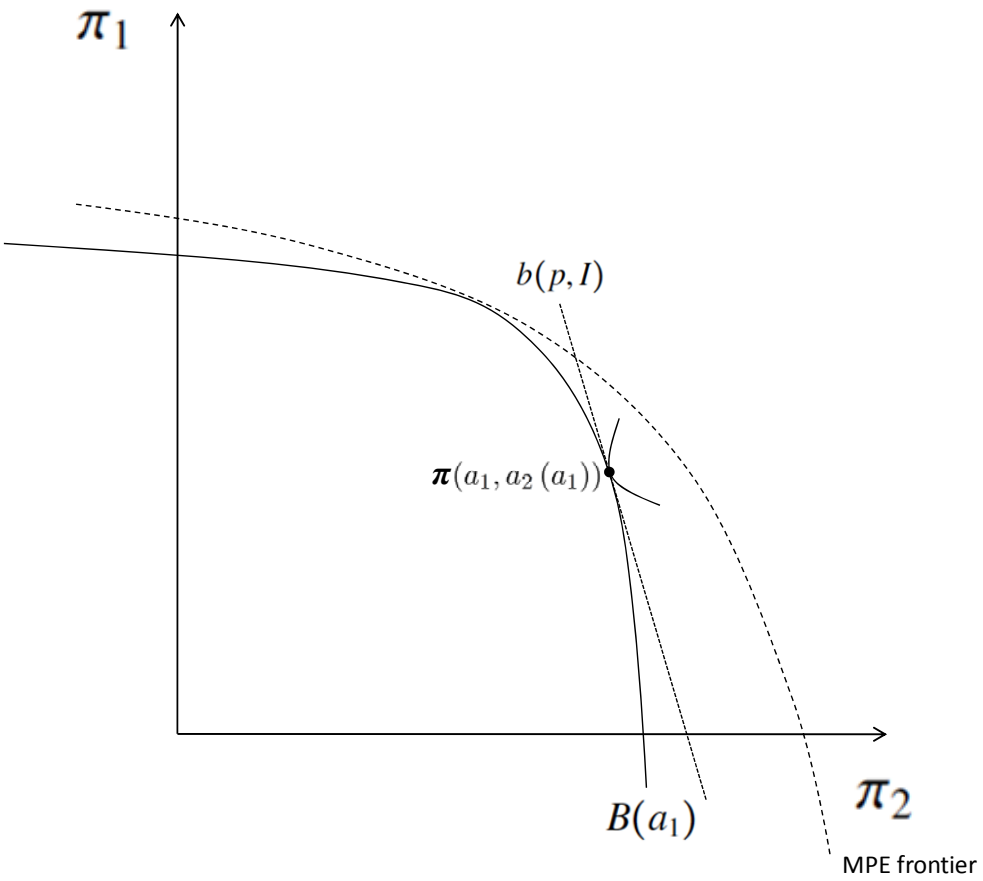


Figure 2. SM's optimal choice on the budget curve $B(a_1)$, $a_2(a_1)$, is the same as SM's optimal choice on the budget line $b(p, I)$ that first-order approximates the budget curve.

[3]

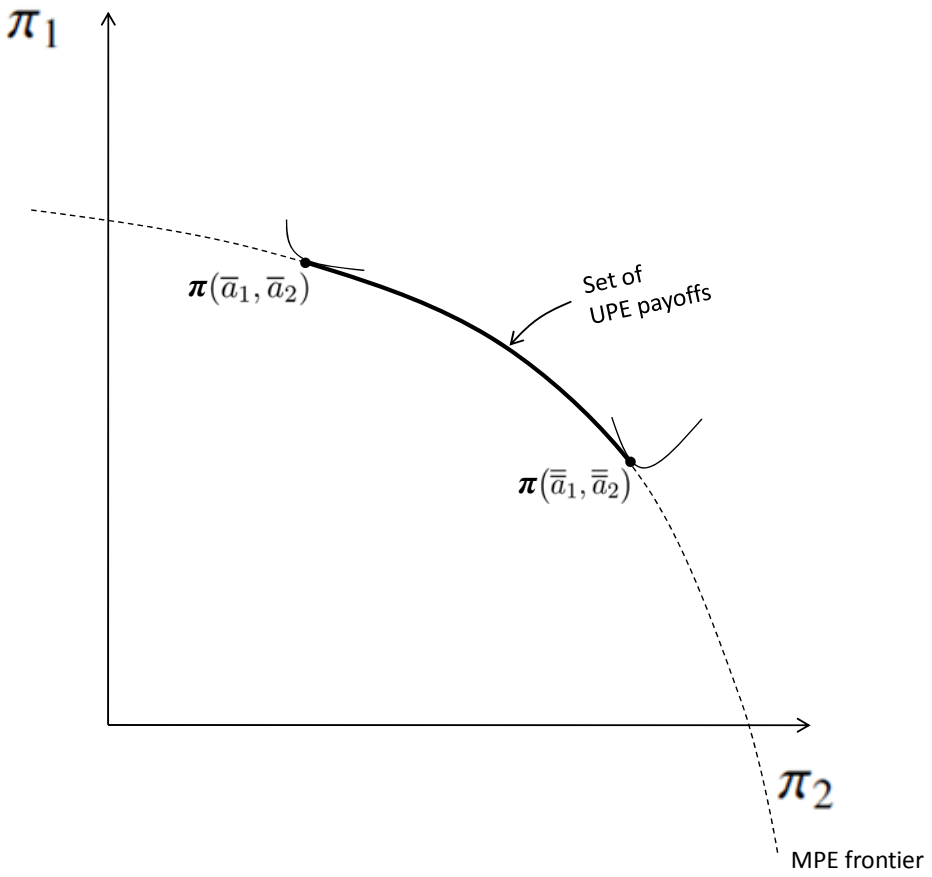


Figure 3. Relationship between utility Pareto efficiency and material Pareto efficiency. SM's distributional preferences are joint-monotonic, and FM's are monotonic.

[4]

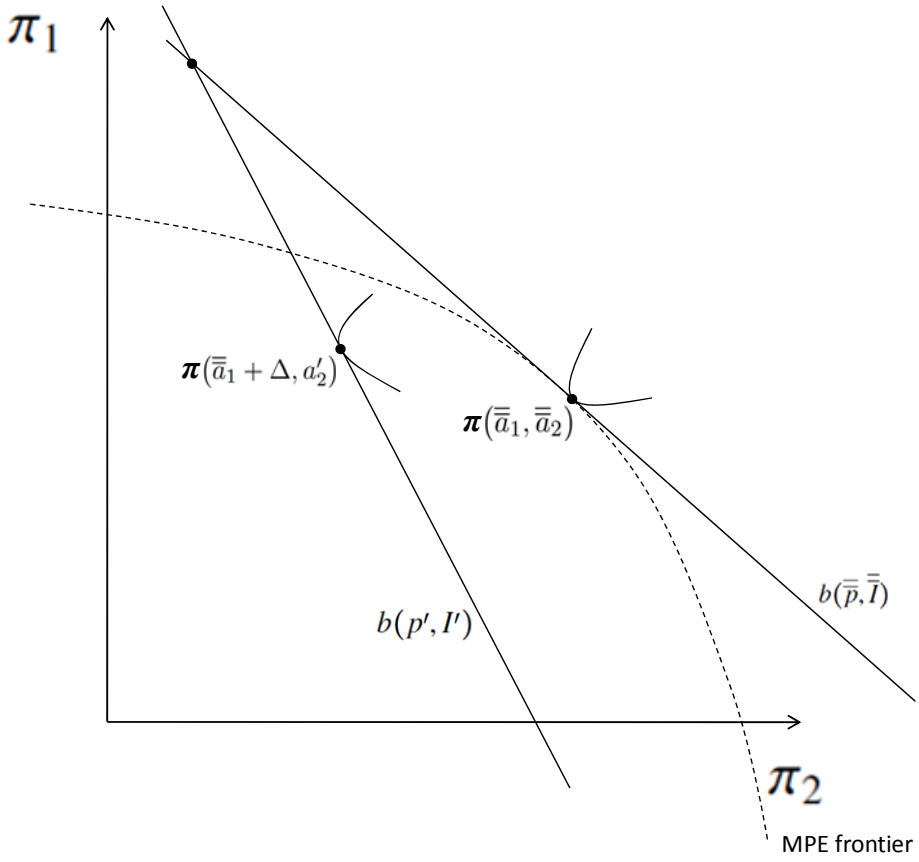


Figure 4. The effect of FM deviating from the action $\bar{\bar{a}}_1$ that would generate an MPE outcome. At the old material payoff pair, $\boldsymbol{\pi}\left(\bar{\bar{a}}_1, \bar{\bar{a}}_2\right)$, the budget line is tangent is to the MPE frontier. At the new material payoff pair, $\boldsymbol{\pi}\left(\bar{\bar{a}}_1 + \Delta, a_2'\right)$, there is a different budget line. The movement from $\boldsymbol{\pi}\left(\bar{\bar{a}}_1, \bar{\bar{a}}_2\right)$ to $\boldsymbol{\pi}\left(\bar{\bar{a}}_1 + \Delta, a_2'\right)$ can be decomposed into an income effect and a substitution effect.

[5]

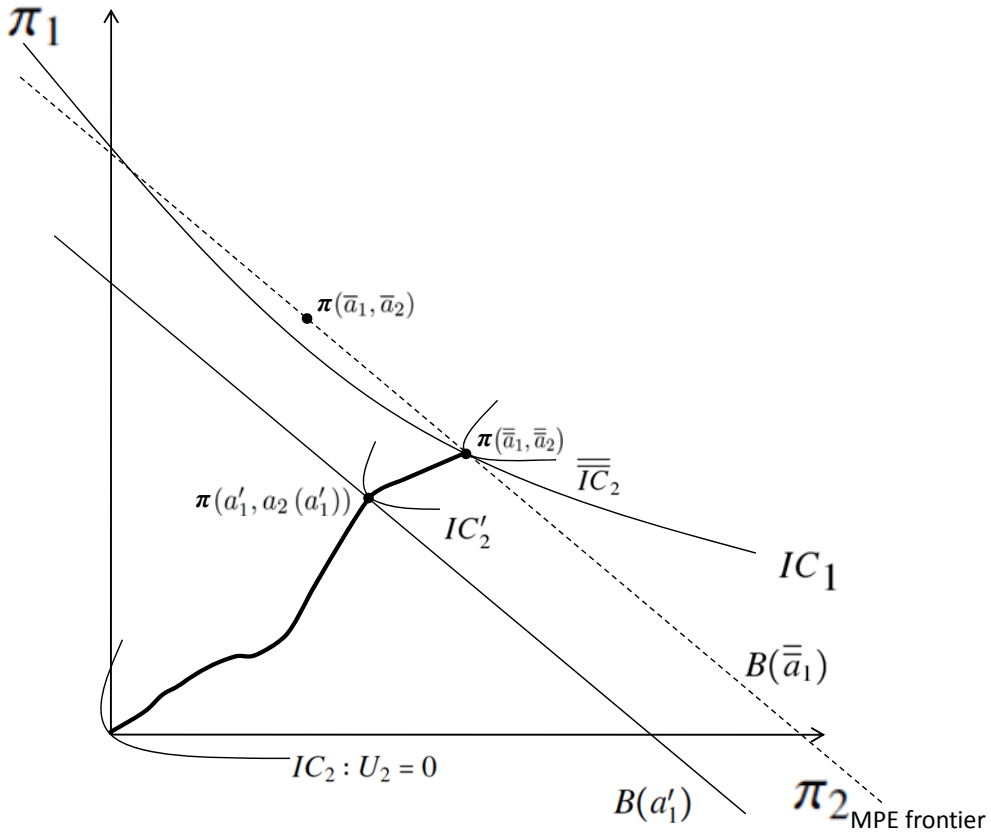


Figure 5. Case I: Budget curves are parallel shifts. Here the material payoff functions are quasi-linear in a_2, hence the budget curves are linear. The thick curve shows the path of material payoff pairs that could occur, $\boldsymbol{\pi}(a_1, a_2(a_1))$, given different possible actions by FM. As FM's action increases, the material payoffs move along the path up to SM's favorite transaction and then go back down the same path (if FM takes an "inefficiently high" level of her action). Since FM's distributional preferences are monotonic, the equilibrium occurs at SM's favorite transaction.

[6a] [6b]

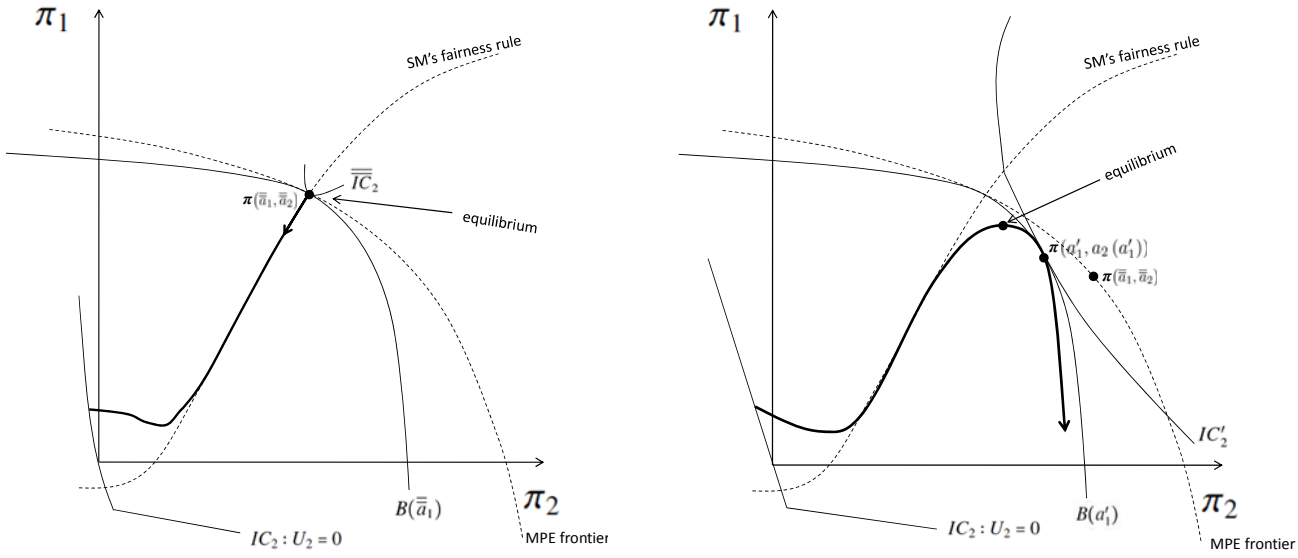


Figure 6. Case II: SM's has fairness-kinked distributional preferences. In both panels, the thick curve shows the path of material payoff pairs that could occur, $\boldsymbol{\pi}(a_1, a_2(a_1))$, given different possible actions by FM. The figures assume FM is purely self-regarding. Panel (a): SM's favorite transaction is on the fairness rule and is a global optimum for FM. Panel (b): SM's favorite transaction is not on the fairness rule, and the equilibrium is neither MPE nor UPE.