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

Daniel J. Benjamin*
Cornell University and NBER

August 31, 2013

Abstract

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. (94 words)

JEL classification: D63, J33, J41, M52, D64

Keywords: distributional preferences, fairness, altruism, gift exchange, rotten kid theorem

*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. E-mail: db468@cornell.edu.

1 Introduction

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 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).

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 2 illustrates and previews the main results of the paper. Section 3 lays out the more general set-up of the bilateral exchange game. Section 4 introduces the general properties that distributional preferences might satisfy. Section 5 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 6 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 7 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 8 provides two sets of sufficient conditions for the equilibrium to be efficient, each corresponding to one of the two cases. Section 9 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.

2 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 $\pi_1(a_1, a_2) = \frac{I_1 + a_2 - n(a_1)}{P_1}$, where $P_1 > 0$ is the market price of consumption faced by the child. The parent's consumption is $\pi_2(a_1, a_2) = \frac{I_2 + a_1 - a_2}{P_2}$, 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 8 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

[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 $\rho \geq \frac{1}{2}$.

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 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 8 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.

3 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* $\frac{\partial \pi_2}{\partial a_2} < 0$.

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.

[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.

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.

4 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 $(\pi_1, \pi_2) \in D_f$, $\frac{\partial U}{\partial \pi_2} > 0$; and (c) for all $(\pi_1, \pi_2) \in A_f$, $\frac{\partial U}{\partial \pi_1} > 0$.*

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 $\widetilde{\pi}_2(p; I)$ 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.*

[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).

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 implies local non-satiation but additionally requires that it is possible to find a more-preferred allocation in

[4] To be precise, at any material payoff pair where $\frac{\partial U(\pi_1,\pi_2)}{\partial \pi_1} > 0$ and $\frac{\partial U(\pi_1,\pi_2)}{\partial \pi_2} > 0$, 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 $\frac{\partial U(\pi_1,\pi_2)}{\partial \pi_1} > 0$ and $\frac{\partial U(\pi_1,\pi_2)}{\partial \pi_2} > 0$ 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 $\frac{\partial U_2(\pi_1,\pi_2)}{\partial \pi_1} > 0$ and $\frac{\partial U_2(\pi_1,\pi_2)}{\partial \pi_2} > 0$ 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.

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,

[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.

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 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 $\frac{\partial U}{\partial \pi_1} = \frac{\partial U}{\partial \pi_2} = 0$ 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*

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

TA2 would be satisfied if, as in Example 1 in Section 2, $\lim_{\pi_2 \to \infty} \frac{\partial U(\underline{\pi}_1, \pi_2)/\partial \pi_2}{\partial U(\underline{\pi}_1, \pi_2)/\partial \pi_1} = 0$ (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

[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 $\frac{\partial U}{\partial \pi_1} = \frac{\partial U}{\partial \pi_2} = 0$, 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 $\min\{x, y\} = 0$.

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.

5 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 $\overrightarrow{\pi}(a_1, a_2) \equiv (\pi_1(a_1, a_2), \pi_2(a_1, a_2))$. 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** $B(a_1) = \{\overrightarrow{\pi}(a_1, a_2)\}_{a_2 \in \mathbb{R}}$. 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 $p = p(a_1, a_2) \equiv -\left.\frac{d\pi_1}{d\pi_2}\right|_{B(a_1)}$ 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: $\overrightarrow{\pi}(\widehat{a}_1, a_2(\widehat{a}_1)) \in \text{graph}(f)$. This occurs when

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

and

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

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.

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 $\frac{\partial U_2}{\partial\pi_1}>0$ and $\frac{\partial U_2}{\partial\pi_2}>0$ 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.

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.

[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.

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 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]

6 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 $U_1\left(\overrightarrow{\pi}\left(\widehat{a}_1, \widehat{a}_2\right)\right) \geq U_1\left(\overrightarrow{\pi}\left(a_1, a_2\right)\right)$ and $U_2\left(\overrightarrow{\pi}\left(\widehat{a}_1, \widehat{a}_2\right)\right) \geq U_2\left(\overrightarrow{\pi}\left(a_1, a_2\right)\right)$, 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 $\pi_1\left(\widehat{a}_1, \widehat{a}_2\right) \geq \pi_1\left(a_1, a_2\right)$ and $\pi_2\left(\widehat{a}_1, \widehat{a}_2\right) \geq \pi_2\left(a_1, a_2\right)$, at least one inequality strict.*

[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, $\widetilde{\pi}_1\left(a_1, a_2\right)$ and $\widetilde{\pi}_2\left(a_1, a_2\right)$, defined by $\widetilde{\pi}_1\left(a_1, a_2\right) \equiv \pi_1\left(-a_1, a_2\right)$ and $\widetilde{\pi}_2\left(a_1, a_2\right) \equiv \pi_1\left(a_1, -a_2\right)$, 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: $\pi_1\left(a_1, a_2\right) = \left(k_1 - a_1\right) a_2$ and $\pi_2\left(a_1, a_2\right) = a_1 - c\left(a_2\right) - k_2$, where $c\left(\cdot\right)$ 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.

[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 $\overline{\overline{a}}_1$.

If a transaction (a_1, a_2) is MPE, then I will also refer to the resulting material payoff pair $\overrightarrow{\pi}(a_1, a_2)$ 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

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

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

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

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 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 $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$.*

[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."

[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.

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]

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). 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.

7 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.

[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.

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* $U_1\left(\overrightarrow{\pi}(a_1, a_2(a_1))\right) \geq 0$*.*

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. $\frac{dp(a_1, a_2(a_1))}{da_1} = 0$.

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.

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 $\frac{dp(a_1,a_2(a_1))}{da_1} = 0$ 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: $\frac{dp(a_1,a_2(a_1))}{da_1} = 0$ 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.

8 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.

8.1 Efficient Case I: Budget Curves Are Parallel Shifts

As discussed in Section 7, 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.

[16] In an influential paper, Bergstrom (1989) argued but did not prove that global conditional transferability, as defined in Section 8, 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.

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]

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 $U_1\left(\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\right) \geq 0$, then the unique equilibrium transaction is the efficient transaction $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$.*

Figure 5 illustrates Theorem 3.

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 2). 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.

[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.

8.2 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 2, 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 $\pi_1(\widehat{a}_1, \widehat{a}_2) = \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$, $U_2\left(\overrightarrow{\pi}(\widehat{a}_1, \widehat{a}_2)\right) = 0$, 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:* $\lim_{\overrightarrow{\pi} \to \overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right), \overrightarrow{\pi} \in D_f} \left(\frac{\partial U_2\left(\overrightarrow{\pi}\right)}{\partial \pi_2} - p(\overline{\overline{a}}_1, \widehat{a}_2) \frac{\partial U_2\left(\overrightarrow{\pi}\right)}{\partial \pi_1} \right) > 0.$

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 $U_1\left(\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\right) \geq 0$, then the unique equilibrium transaction is the efficient transaction $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$.*

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.

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.

[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.

9 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 2, 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 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

[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.

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

[1] James Andreoni and B. Douglas Bernheim. Social image and the 50-50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77(5):1607–1636, 2009.

[2] James Andreoni and John Miller. Giving according to garp: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2):737–753, March 2002.

[3] Max H. Bazerman, George F. Loewenstein, and Sally Blount White. Reversals of preference in allocation decisions: Judging an alternative versus choosing among alternatives. *Administrative Science Quarterly*, 37(2):220–240, June 1992.

[4] Gary S. Becker. A theory of social interactions. *Journal of Political Economy*, 82(6):1063–93, November/December 1974.

[5] Joyce Berg, John Dickhaut, and Kevin McCabe. Trust, reciprocity, and social history. *Games and Economic Behavior*, 10:122–142, 1995.

[6] Theodore C. Bergstrom. A fresh look at the rotten kid theorem – and other household mysteries. *Journal of Political Economy*, 97(5):1138–59, 1989.

[7] Gary E. Bolton and Axel Ockenfels. Erc: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1):166–193, March 2000.

[8] Gary E. Bolton and Axel Ockenfels. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments: Comment. *American Economic Review*, 96(5):1906–1911, 2006.

[9] Colin F. Camerer. *Behavioral Game Theory*. Princeton University Press, Princeton, NJ, 2003.

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

[11] Gary Charness and Matthew Rabin. Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117(3):817–869, August 2002.

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

[13] James C. Cox, Daniel Friedman, and Vjollca Sadiraj. Revealed altruism. *Econometrica*, 76(1):31–69, 2008.

[14] James C. Cox and Vjollca Sadiraj. Direct tests of individual preferences for efficiency and equity. *Economic Inquiry*, December 2010.

[15] Bouwe R. Dijkstra. Samaritan versus rotten kid: Another look. *Journal of Economic Behavior and Organization*, 64:91–110, 2007.

[16] Martin Dufwenberg, Paul Heidhues, Georg Kirchsteiger, Frank Riedel, and Joel Sobel. Other-regarding preferences in general equilibrium. *Review of Economic Studies*, 78:640–66, 2011.

[17] Dirk Engelmann and Martin Strobel. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review*, 94(4):857–869, September 2004.

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

[19] Ernst Fehr, Georg Kirchsteiger, and Arno Riedl. Does fairness prevent market clearing? an experimental investigation. *Quarterly Journal of Economics*, 108(2):437–459, 1993.

[20] Ernst Fehr, Georg Kirchsteiger, and Arno Riedl. Gift exchange and reciprocity in competitive experimental markets. *European Economic Review*, 42:1–34, 1998.

[21] Ernst Fehr, Alexander Klein, and Klaus M. Schmidt. Fairness and contract design. *Econometrica*, 75(1):121–154, 2007.

[22] Ernst Fehr, Michael Naef, and Klaus M. Schmidt. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments: Comment. *American Economic Review*, 96(5):1912–1917, 2006.

[23] Ernst Fehr and Klaus Schmidt. Theories of fairness and reciprocity: Evidence and economic applications. In M. Dewatripont, L.P. Hansen, and S. Turnovski, editors, *Advances in Economic Theory, Eighth World Conference of the Econometric Society, Vol. 1*, pages 208–257. Cambridge, U.K.: Cambridge University Press, 2003.

[24] Ernst Fehr and Klaus M. Schmidt. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3):817–868, August 1999.

[25] Ernst Fehr and Klaus M. Schmidt. The role of equality, efficiency, and rawlsian motives in social preferences: A reply to englemann and strobel. January 2004. University of Zurich Institute for Empirical Research in Economics Working Paper Number 179.

[26] Raymond Fisman, Shachar Kariv, and Daniel Markovits. Individual preferences for giving. *American Economic Review*, 97(5):1858–1876, 2007.

[27] Drew Fudenberg and Eric Maskin. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3):533–554, 1986.

[28] Werner Güth, R. Schmittberger, and B. Schwarze. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, 3:367–388, 1982.

[29] Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler. Fairness as a constraint on profit seeking entitlements in the market. *American Economic Review*, 76(4):728–41, 1986.

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

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

[32] Nick Netzer and Armin Schmutzler. Explaining gift-exchange – the limits of good intentions. May 2013. University of Zurich Socioeconomic Institute Working Paper No. 0919.

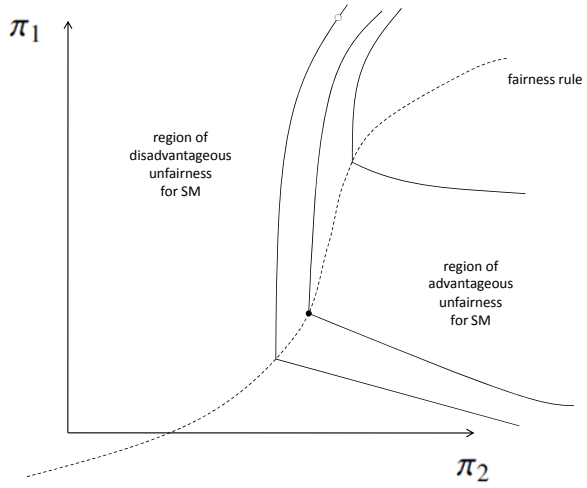
[33] Vittorio Pelligra and Luca Stanca. To give or not to give? equity, efficiency and altruistic behavior in an artefactual field experiment. *Journal of Socio-Economics*, 46:1–9, 2013.

[34] John K.-H. Quah. Supplement to 'the comparative statics of constrained optimization problems'. *Econometrica*, 75(2):401–431, 2007.

[35] Matthew Rabin. Incorporating fairness into game theory and economics. *American Economic Review*, 83(5):1281–1302, December 1993.

[36] Amartya Sen. Behaviour and the concept of preference. *Economica*, 40(159):241–259, 1973.

[1a]



[1b]

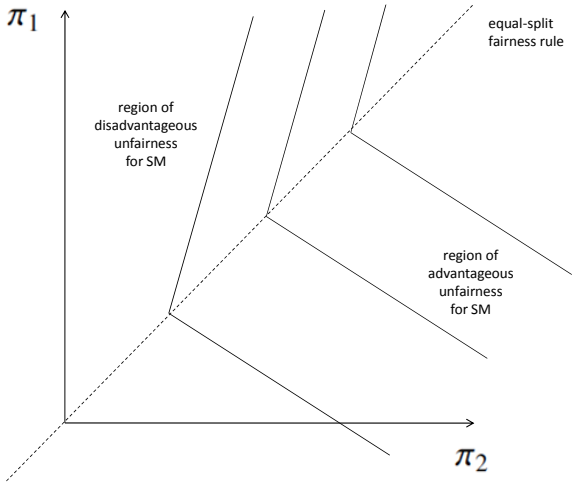
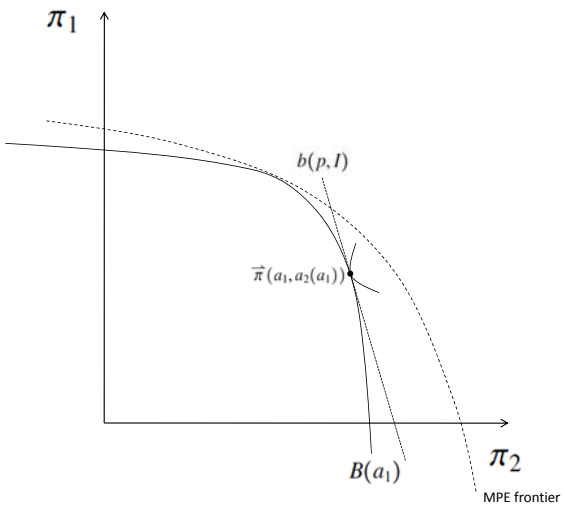
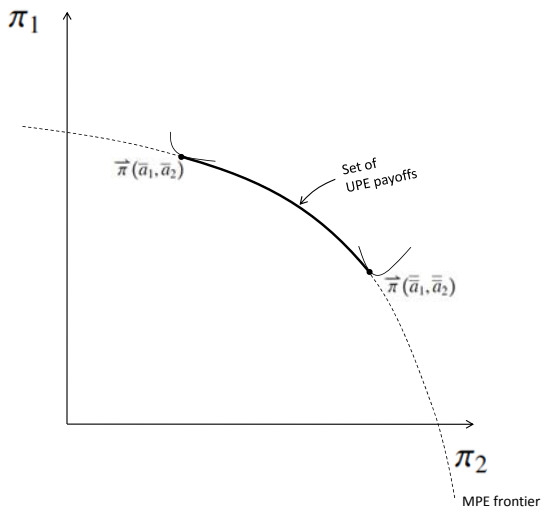


Figure 1. Interpersonal indifference curves. Panel (a): Fairness-kinked preferences. The fairness rule is an upward-sloping locus of material payoff pairs. The indifference curves are kinked at each material payoff pair on the fairness rule. These particular preferences are joint-monotonic but not monotonic; 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]



[3]



[4]

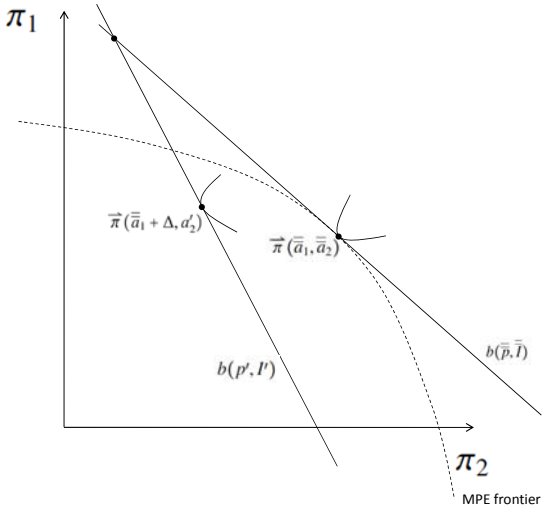
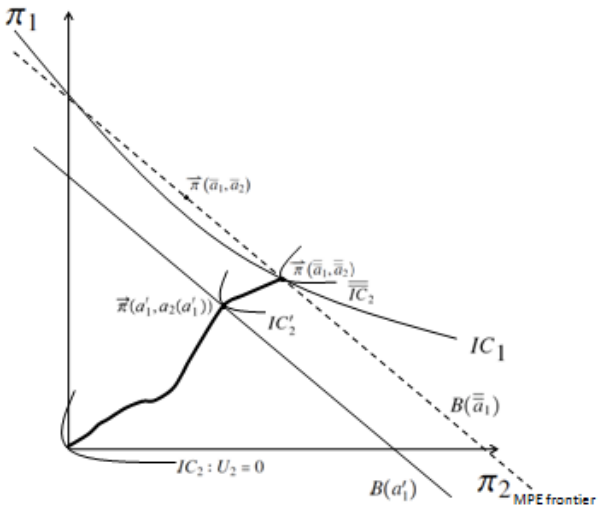


Figure 2. SM's optimal choice on the budget curve $B(a_1)$, $a_2(a_1)$, is the same as SM's optimal choice on the budget line $b(p, I)$ that first-order approximates the budget curve.

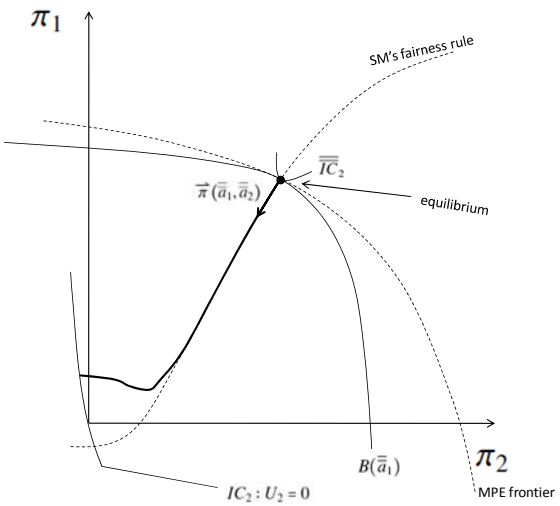
Figure 3. Relationship between utility Pareto efficiency and material Pareto efficiency. SM's distributional preferences are joint-monotonic, and FM's are monotonic.

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, $\vec{\pi}(\bar{\bar{a}}_1, \bar{\bar{a}}_2)$, the budget line is tangent is to the MPE frontier. At the new material payoff pair, $\vec{\pi}(\bar{\bar{a}}_1 + \Delta, a_2')$, there is a different budget line. The movement from $\vec{\pi}(\bar{\bar{a}}_1, \bar{\bar{a}}_2)$ to $\vec{\pi}(\bar{\bar{a}}_1 + \Delta, a_2')$ can be decomposed into an income effect and a substitution effect.

[5]



[6a]



[6b]

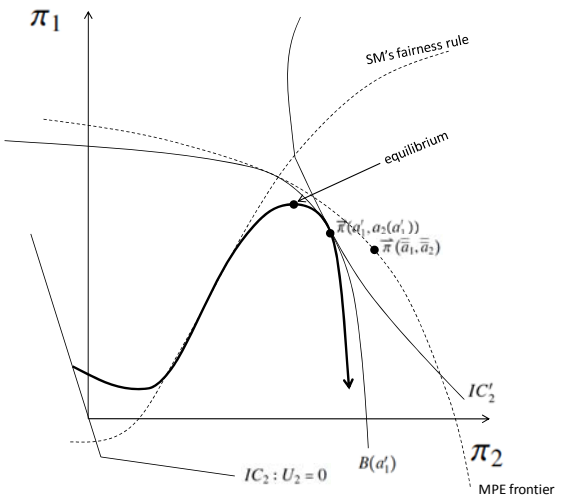


Figure 5. Case I: Budget curves are parallel shifts. Here the material payoff functions are quasi-linear in a_2, hence the budget curves are linear. The thick curve shows the path of material payoff pairs that could occur, $\overrightarrow{\pi}(a_1, a_2(a_1))$, given different possible actions by FM. 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.

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, $\overrightarrow{\pi}(a_1, a_2(a_1))$, 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.

Web Appendix A: Neither Player's Distributional Preferences Are Monotonic
(Not for publication)

In this appendix, I discuss how the conclusions of the analysis are affected if both FM's and SM's distributional preferences are joint-monotonic but not necessarily monotonic.

Theorem 1 in the main text shows that if one player's distributional preferences are joint-monotonic and the other player's are monotonic, then the set of UPE material payoff pairs is a subset of the set of MPE material payoff pairs. Theorem A1 generalizes to the case where both players' distributional preferences are joint-monotonic. In that case, there are UPE material payoff pairs that are not MPE. To state the result, let the interpersonal indifference curve of FM that goes through SM's favorite transaction $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$ be denoted $\overline{\overline{IC}}_1$, and let the indifference curve of SM that goes through FM's favorite transaction $(\overline{\pi}_1, \overline{\pi}_2)$ be denoted $\overline{IC}_2$.

Theorem A1. *Suppose U_1 and U_2 are 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 is a connected set that includes $(\overline{\pi}_1, \overline{\pi}_2)$ and $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$ and lies within the region enclosed by $\overline{\overline{IC}}_1$, $\overline{IC}_2$, and the MPE frontier.*

Figure A1 illustrates the relationship between the set of MPE material payoff pairs and the set of UPE material payoff pairs in the case where neither player has monotonic distributional preferences. The set of UPE material payoff pairs lies within the region enclosed by $\overline{\overline{IC}}_1$, $\overline{IC}_2$, and the MPE frontier because both players prefer any material payoff within that region to any feasible material payoff pair outside that region. A material payoff pair that is UPE either occurs at a tangency point between the players' indifference curves—at a point where both indifference curves are upward sloping (as shown in the figure)—or it occurs on the MPE frontier if the "relevant tangency" lies outside the set of feasible material-payoff pairs.

Theorem A1 specializes to Theorem 1 when at least one player has monotonic distributional preferences. In that case, graphically, there cannot be a tangency between the players' indifference curves because the indifference curves of the player with monotonic distributional preferences are everywhere downward sloping.

Dufwenberg, Heidhues, Kirchsteiger, Riedel, & Sobel (2011) independently prove a different result that is also more general than Theorem 1. They assume that the agents' distributional preferences satisfy a condition they call "social monotonicity," which can be defined as follows:

Definition A1. *U_1 and U_2 are **social-monotonic** if for any (π_1, π_2) and any $\varepsilon > 0$, there is some $(\widehat{\pi}_1, \widehat{\pi}_2)$ such that $0 < \widehat{\pi}_1 - \pi_1 < \varepsilon$, $0 < \widehat{\pi}_2 - \pi_2 < \varepsilon$, $U_1(\widehat{\pi}_1, \widehat{\pi}_2) > U_1(\pi_1, \pi_2)$, and $U_2(\widehat{\pi}_1, \widehat{\pi}_2) > U_2(\pi_1, \pi_2)$.*

The definition differs from joint monotonicity because it requires that for any material payoff pair, there is an arbitrarily close alternative material payoff pair giving more to both players that *both* agents strictly prefer. If the players' preferences satisfy social monotonicity, then both players' preferences are joint-monotonic, but both players' preferences can be joint-monotonic without satisfying social monotonicity. (In comparing social monotonicity with joint monotonicity, Dufwenberg et al mis-state the definition of joint monotonicity to be essentially the same as my statement of social monotonicity.)

Under the same conditions as Theorem 1, except that the players' distributional preferences are assumed to be socially monotonic, Dufwenberg et al prove that the set of UPE material payoff pairs is a subset of the set of MPE material payoff pairs. Their result is more general than Theorem 1 because if one player's distributional preferences are joint-monotonic and the other player's distributional preferences are monotonic, then the players' preferences satisfy social monotonicity.

We now turn from discussing which material payoff pairs are efficient to discussing whether the equilibrium is efficient. Theorem 2 in the main text gives necessary conditions for the equilibrium to be MPE. That theorem applies directly when both players' preferences are joint-monotonic. As discussed above, however, when both players' preferences are joint-monotonic, there may be UPE transactions that are not MPE. Theorem A2 presents necessary conditions for the equilibrium to be UPE. If the equilibrium is MPE, then it is also UPE, but there are also other cases where the equilibrium is UPE but not MPE.

Theorem A2. *Suppose U_1 and U_2 are joint-monotonic and quasi-concave, and both are either twice-continuously differentiable or fairness-kinked. If the equilibrium $(a_1, a_2(a_1))$ is UPE and not MPE, then $(a_1, a_2(a_1))$ is a fairness-rule optimum for SM. If, in addition, $(a_1, a_2(a_1))$ is a strict fairness-rule optimum for SM and any fairness rule is continuously differentiable, then at least one of the following must be true:*

1. *SM's indifference curve for disadvantageously unfair transactions is tangent to SM's fairness rule at $\overrightarrow{\pi}(a_1, a_2(a_1))$.*
2. *U_1 is fairness-kinked, $\overrightarrow{\pi}(a_1, a_2(a_1))$ is on FM's fairness rule, and the respective fairness rules f_1 and f_2 have different slopes at $\overrightarrow{\pi}(a_1, a_2(a_1))$.*

Figure A2a illustrates the Case 2 listed in the theorem, which can be interpreted as a setting where the two agents have different, self-serving ideas about what is fair. However, Figure A2b shows that even if the equilibrium occurs on both players' fairness rules, the equilibrium is not necessarily UPE. A corollary of Theorem A2 is that if both players' interpersonal indifference curves are smooth—thereby ruling out fairness-kinkedness—then the equilibrium is UPE if and only if it is MPE.

Theorems 3 and 4 provide sufficient conditions for the equilibrium to be MPE and UPE, but they assume that FM's preferences are purely self-regarding or monotonic. If FM's preferences are required only to be joint-monotonic, then the conclusions of the theorems may not hold. Even though SM's behavior aligns the *material* incentives of the two players, if FM's preferences are non-monotonic, then she may prefer not to maximize the players' material payoffs. For the case of preferences that satisfy the conditions of Theorem 3—except that FM's distributional preferences are merely joint-monotonic—Figure A3 illustrates an equilibrium that is neither MPE nor UPE.

[A1]

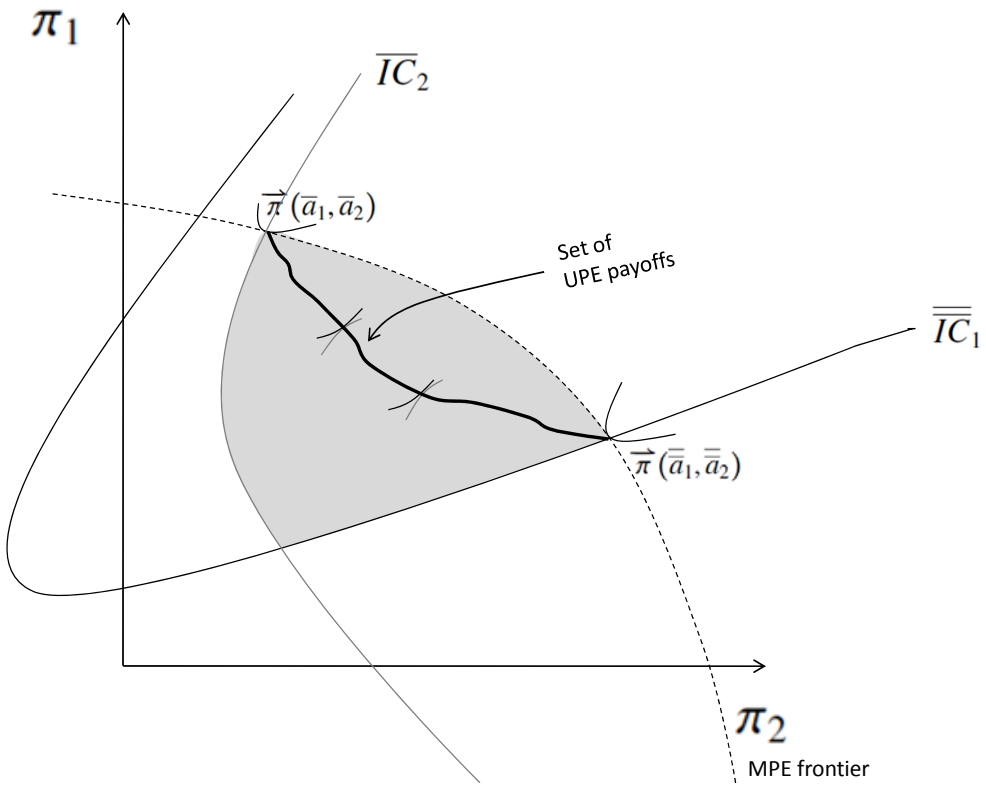


Figure A1. Relationship between utility Pareto efficiency and material Pareto-efficiency. Both players have joint-monotonic preferences. The set of UPE material payoff pairs must lie in the gray region.

[A2a] [A2b]

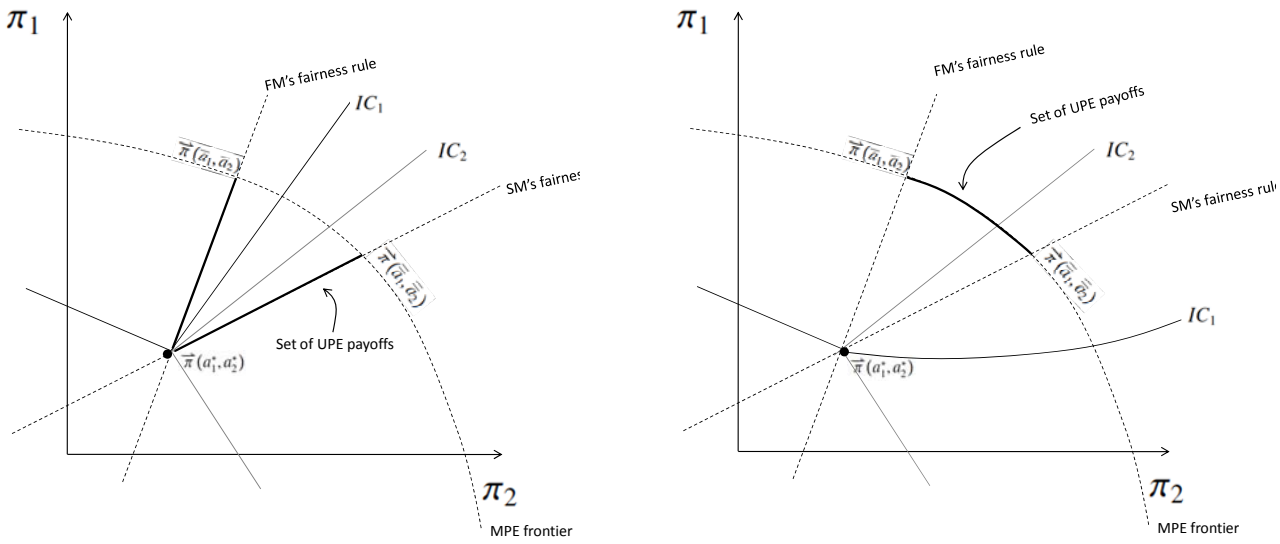


Figure A2. Panel (a): An equilibrium that is UPE but not MPE. Both players' indifference curves are fairness-kinked at the equilibrium, but FM and SM have different fairness rules. Panel (b): A similar situation, except that the equilibrium is neither UPE nor MPE.

[A3]

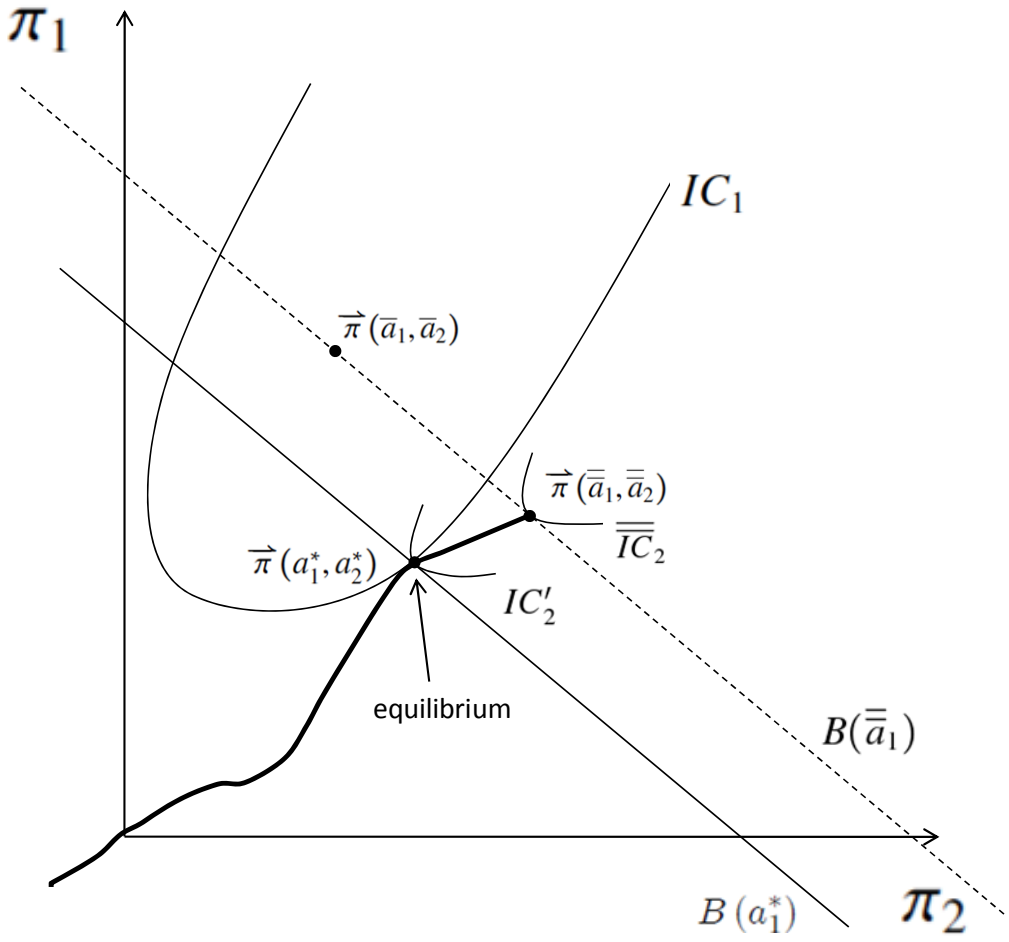


Figure A3. The conditions of Theorem 3 are satisfied, except that FM's distributional preferences are merely joint-monotonic (and not monotonic). Because FM's distributional preferences are joint-monotonic, the equilibrium does not occur at SM's favorite transaction. The dark line shows the path of material payoff pairs that could occur, $\overrightarrow{\pi}(a_1, a_2(a_1))$, given different possible actions by FM. This path increases up to SM's favorite transaction and then goes back down the same path (if FM takes an "inefficiently high" level of her action). The equilibrium material payoff pair occurs at FM's most-preferred point along this path.

Web Appendix B: Proofs

(Not for publication)

Before proving the results in the text, we establish a technical lemma.

Technical Lemma. *Suppose* U_1 and U_2 *are joint-monotonic and quasi-concave. Then:*

1. *The set of individually-rational transactions*

$$T \equiv \{(a_1, a_2) \mid U_1(\overrightarrow{\pi}(a_1, a_2)) \geq 0, U_2(\overrightarrow{\pi}(a_1, a_2)) \geq 0\}$$

is non-empty and compact, as is the set of payoff pairs $T_\pi \equiv \{\overrightarrow{\pi}(a_1, a_2) \mid (a_1, a_2) \in T\}$.

2. *Along any graph of the form* $(g(\pi_2), \pi_2)$*, where* g *is a continuous, decreasing, weakly concave function,* U_i *has a unique maximum* π_2^* *and strictly decreases as* π_2 *moves away from this maximum, for* $i = 1, 2$*. Moreover, the MPE frontier and each budget curve* $B(a_1)$ *is such a graph.*

Proof of part 1: The transaction $(a_1, a_2) = (0, 0)$ gives material payoffs $\overrightarrow{\pi}(0, 0) = (0, 0)$ and utilities $U_1(\overrightarrow{\pi}(0, 0)) = U_2(\overrightarrow{\pi}(0, 0)) = 0$, so both sets are non-empty. By TA2, T necessarily lies to the north and east (respectively) of two lines $\pi_1 = \underline{\underline{\pi}}_1 \leq \underline{\pi}_1$ and $\pi_2 = \underline{\underline{\pi}}_2 \leq \underline{\pi}_2$, i.e., $T \subseteq \{(a_1, a_2) \mid \pi_1(a_1, a_2) \geq \underline{\underline{\pi}}_1, \pi_2(a_1, a_2) \geq \underline{\underline{\pi}}_2\}$. Hence T_π is closed and bounded and therefore compact. It follows from A4 that T is also closed and bounded and therefore compact.

Proof of part 2: WLOG, consider U_2. We first show that for any real number k, the set $\{\pi_2 \mid U_2(g(\pi_2), \pi_2) \geq k\}$ is an interval (possibly unbounded). Let $\pi_2' < \pi_2''$ be two values in this set. By construction, $U_2 \geq k$ at $(g(\pi_2'), \pi_2')$ and $(g(\pi_2''), \pi_2'')$. It follows that $U_2 \geq k$ at $(g(\pi_2'), \pi_2'')$. (To see this, let $\overline{y} = \max\{y \in [g(\pi_2''), g(\pi_2')] \mid U_2(y, \pi_2'') \geq k\}$ (the maximum exists by continuity). If $\bar{y} = g(\pi_2')$ then we are done, so assume $\bar{y} < g(\pi_2')$. By joint-monotonicity, we can choose $\hat{y}, \hat{x}$ with $\bar{y} < \hat{y} < g(\pi_2')$ and $\hat{x} > \pi_2''$ so that $U_2(\hat{y}, \hat{x}) > U_2(\bar{y}, \pi_2'') \geq k$. The line segment connecting $(g(\pi_2'), \pi_2')$ and $(\hat{y}, \hat{x})$ meets the line $x = \pi_2''$ at a point with some y-coordinate strictly between $\bar{y}$ and $g(\pi_2')$. By quasi-concavity, the value of U_2 at this point is $\geq k$. This contradicts the maximality of $\bar{y}$.) Now, for any $\pi_2' < \pi_2 < \pi_2''$, the point $(g(\pi_2), \pi_2)$ lies weakly inside the triangle defined by these three points since g is weakly concave. Since U_2 is quasi-concave, $U_2(g(\pi_2), \pi_2) \geq k$ also.

This shows that there cannot be three values $\pi_2' < \pi_2 < \pi_2''$ with $U_2(g(\pi_2'), \pi_2') > U_2(g(\pi_2), \pi_2) < U_2(g(\pi_2''), \pi_2'')$. It follows that on the graph $(g(\pi_2), \pi_2)$, U_2 is either weakly monotonic everywhere, or weakly increasing on $(-\infty, \widetilde{\pi}_2)$ and weakly decreasing on $(\widetilde{\pi}_2, \infty)$ for some $\widetilde{\pi}_2$.

We now show that U_2 cannot be constant on any interval along the graph. Suppose U_2 assumes the constant value k on the interval $[\pi_2', \pi_2'']$. Quasi-concavity implies that U_2 is $\geq k$ at the point $(y_0, x_0) = \left(\frac{g(\pi_2') + g(\pi_2'')}{2}, \frac{\pi_2' + \pi_2''}{2} \right)$. For sufficiently small $\epsilon > 0$, the box $[y_0, y_0 + \epsilon] \times [x_0, x_0 + \epsilon]$ lies entirely below and to the left of the curve $C = \{(g(\pi_2), \pi_2) \mid \pi_2' < \pi_2 < \pi_2''\}$. Joint-monotonicity ensures that U_2 assumes a value $k' > k$ at some point (y', x') inside this box. Now, let $S = \{(y, x) \mid y \geq y',\ x \geq x',\ U_2(y, x) \geq U_2(y', x')\}$. We know that S does not intersect C because $U_2 \geq k'$ on S, whereas U_2 takes on the constant value k on C, by assumption. S is closed and convex, and must then be bounded (by the lines $y = y', x = x'$, as well as by the curve C since $(y', x') \in S$), so it is compact. Hence we can choose a point $(y, x) \in S$ with $x + y$ maximal. But by joint-monotonicity there exists $y'' > y', x'' > x'$ with $U_2(y'', x'') > U_2(y', x') \geq k'$, contradicting maximality. It follows that U_2 cannot be constant on $[\pi_2', \pi_2'']$ after all.

Next, we rule out that U_2 is monotonic along the entire graph; in particular, we show that for any $(g(\pi_2), \pi_2)$, there are $\pi_2' < \pi_2 < \pi_2''$ such that $U_2(g(\pi_2'), \pi_2') < U_2(g(\pi_2), \pi_2) > U_2(g(\pi_2''), \pi_2'')$. Since the graph is weakly concave, the indifference curve going through $(g(\pi_2), \pi_2)$ is either tangent to the budget curve or by TA2 intersects it at $(g(\pi_2), \pi_2)$ and at some other point $(g(\pi_2'''), \pi_2''')$. In either cases, the claim follows immediately.

We complete the proof by showing that each budget curve has a graph of the form $(g(\pi_2), \pi_2)$, where g is a continuous, decreasing, weakly concave function; we omit the proof of the same for the MPE frontier, for which the argument is analogous (and is a standard result about the "utility possibility frontier" when utility is purely self-regarding). Fix action $\overline{a}_1$. Let the budget curve $B(\overline{a}_1) \equiv \{\overrightarrow{\pi}(\overline{a}_1, a_2)\}_{a_2 \in \mathbb{R}}$ be parameterized by $\pi_1(\overline{a}_1, a_2) \equiv g(\pi_2(\overline{a}_1, a_2))$; clearly, g is not only continuous but also continuously differentiable with $\left.\frac{d\pi_1}{d\pi_2}\right|_{B(\overline{a}_1)} = \frac{dg}{d\pi_2}$. Differentiating $\pi_1(\overline{a}_1, a_2) \equiv g(\pi_2(\overline{a}_1, a_2))$ with respect to a_2 yields $\frac{\partial \pi_1}{\partial a_2} = \frac{dg}{d\pi_2}\frac{\partial \pi_2}{\partial a_2}$, and therefore $\left.\frac{d\pi_1}{d\pi_2}\right|_{B(\overline{a}_1)} = \frac{dg}{d\pi_2} = \frac{\partial \pi_1/\partial a_2}{\partial \pi_2/\partial a_2} < 0$. Hence g is decreasing. By the chain rule, $\frac{\partial}{\partial a_2}\left(\frac{dg}{d\pi_2}\right) = \frac{d^2 g}{d(\pi_2)^2}\frac{\partial \pi_2}{\partial a_2}$. Rearranging, $\frac{d^2 g}{d(\pi_2)^2} = \frac{\frac{\partial}{\partial a_2}\left(\frac{dg}{d\pi_2}\right)}{\partial \pi_2/\partial a_2} = \frac{\frac{\partial}{\partial a_2}\left(\frac{\partial \pi_1/\partial a_2}{\partial \pi_2/\partial a_2}\right)}{\partial \pi_2/\partial a_2}$. A3 implies that $\frac{\partial}{\partial a_2}\left(\frac{\partial \pi_1/\partial a_2}{\partial \pi_2/\partial a_2}\right) \geq 0$, and A1 implies that $\frac{\partial \pi_2}{\partial a_2} < 0$, so $\frac{d^2 g}{d(\pi_2)^2} \leq 0$. Hence g is weakly concave.

□

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.*

Proof: Follows directly from Theorem 3 below, and here we merely check that the assumptions can be verified or appropriately modified. A1-A4 from Section 3 clearly hold, with A2 following from

$n'(0) < 1$. TA1 from Section 4 and joint-monotonicity of U_2 are satisfied due to the assumption that $U_2(\pi_1, \pi_2)$ is monotonically increasing in both π_1 and π_2. TA2 from Section 4 is implied by the assumption 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 $\lim_{\pi_1 \to \infty} \frac{\partial U_2(\pi_1, \underline{\pi}_2)/\partial \pi_2}{\partial U_2(\pi_1, \underline{\pi}_2)/\partial \pi_1} = \infty$. We have directly assumed that U_2 is quasi-concave and normal, and U_1 is purely self-regarding. The material payoff functions being quasi-linear implies that they are globally conditionally transferable. There is no assumption that $U_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right) \geq 0$ because neither player has an outside option.

□

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.*

Proof: Follows directly from Theorem 4 below, and here we merely check that the assumptions can be verified or appropriately modified. A1-A4 from Section 3 clearly hold, with A2 following from $c'(0) < 1$. TA1 from Section 4 clearly holds. TA2 from Section 4 and the quasi-concavity of U_2 are not needed because the piecewise-linear functional form for U_2, combined with the assumptions regarding the material payoff functions, ensure that an optimal action for the worker exists in response to any a_1. The functional form for U_2 satisfies joint-monotonicity and fairness-kinkedness, and we have directly assumed that U_1 is purely self-regarding. S2 from Section 8.2 clearly holds. S3 can be replaced in the proof of Theorem 4 by the assumption of the piecewise-linear functional form for U_2. S4 and S5 hold but can be dropped as sufficient conditions because FM is purely self-regarding. S1 is satisfied as long as $(1-\sigma) - c'(\widehat{a}_2)\sigma > 0$; or rearranging, $\sigma < \frac{1}{1+c'(\widehat{a}_2)}$. In the next paragraph, we will show that the assumption that $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ is a strict fairness-rule optimum is satisfied as long as $\sigma < \frac{1}{2}$ and $\rho > \frac{1}{2}$. The conclusion then follows from setting $\overline{\sigma} = \min\left\{\frac{1}{2}, \frac{1}{1+c'(\widehat{a}_2)}\right\}$ and noting (as explained below) that here $\rho > \frac{1}{2}$ can be weakened to $\rho \geq \frac{1}{2}$.

We now show that $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$, defined implicitly as the solution to $\pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right) = \pi_2\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ and $c'\left(\overline{\overline{a}}_2\right) = 1$, satisfies $U_1\left(\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\right) = \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right) > 0$ and is a fairness-rule optimum. (Given the assumptions on the material payoff functions and the $c(\cdot)$ function, the solution to these equations exists and is unique.) The conditions on $c(\cdot)$ ensure that the solutions to these equations indeed satisfy $\pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right) > 0$. Translated to this gift-exchange game, the two conditions for $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ to be a strict fairness-rule optimum (as defined in Section 5) are $(1-\sigma) - c'\left(\overline{\overline{a}}_2\right)\sigma > 0$ and $(1-\rho) - c'\left(\overline{\overline{a}}_2\right)\rho < 0$. Substituting $c'\left(\overline{\overline{a}}_2\right) = 1$, these two conditions are satisfied as long as $\sigma < \frac{1}{2}$

and $\rho > \frac{1}{2}$, respectively. The latter condition can be weakened to $\rho \geq \frac{1}{2}$ because if $\rho = \frac{1}{2}$, $\overline{\overline{a}}_1$ remains a local optimum for FM since $(1-\rho) - c'(a_2)\rho < 0$ continues to hold for all $a_2 < \overline{\overline{a}}_2$.

□

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 $\frac{\partial U_2}{\partial \pi_1} > 0$ and $\frac{\partial U_2}{\partial \pi_2} > 0$ at $(\widehat{a}_1, a_2(\widehat{a}_1))$.*

Proof: Technical Lemma immediately gives existence and uniqueness of an optimal action $a_2(a_1)$. The Maximum Theorem (e.g., Sundaram 1996, p.235) can now be applied (where we can ignore the compactness requirement on the budget curve since we have already proved existence of an optimal action) to show that $a_2(a_1)$ is an upper-hemicontinuous correspondence. Since $a_2(a_1)$ is single-valued, it is a continuous function.

Since U_2 is continuously differentiable at $\overrightarrow{\pi}(\widehat{a}_1, a_2(\widehat{a}_1))$, SM's unique optimum is characterized by the first-order condition, $\frac{\partial U_2}{\partial a_2}(\overrightarrow{\pi}(\widehat{a}_1, a_2)) = 0$, which after rearranging is $\frac{\partial U_2}{\partial \pi_2} - p(\widehat{a}_1, a_2)\frac{\partial U_2}{\partial \pi_1} = 0$. Joint-monotonicity rules out that both partial derivatives $\frac{\partial U_2}{\partial \pi_1}$ and $\frac{\partial U_2}{\partial \pi_2}$ are negative, and TA1 rules out that they both equal 0. Therefore, the first-order condition implies that both are positive.

□

Proposition 3.

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.*

Proof of part 1: By definition of $(\widehat{a}_1, a_2(\widehat{a}_1))$ being a strict fairness-rule optimum:

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

and

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

Since these inequalities are strict and since $a_2(a_1)$ is a continuous function of a_1 (by Lemma 1), it follows immediately that these inequalities hold for all a_1 in a neighborhood of $\widehat{a}_1$, and thus $(a_1, a_2(a_1))$ is a strict fairness-rule optimum for all a_1 in a neighborhood of $\widehat{a}_1$. Thus, for any a_1 in a neighborhood of $\widehat{a}_1$, SM will choose action $a_2(a_1)$ such that $\overrightarrow{\pi}(a_1, a_2(a_1)) \in \text{graph}(f)$. The fact that the fairness rule is a strictly upward-sloping locus of material payoff pairs, together with A1, implies that $a_2(a_1)$ is increasing in a_1 at $\widehat{a}_1$. Because the above inequalities are strict, they also imply that for any a_1 in a neighborhood of $\widehat{a}_1$,

$$\lim_{\overrightarrow{\pi} \to \overrightarrow{\pi}(a_1, a_2(a_1)), \overrightarrow{\pi} \in D_f} \left(\frac{\partial U_2(\overrightarrow{\pi})}{\partial \pi_2} - p(\widehat{a}_1, a_2(\widehat{a}_1)) \frac{\partial U_2(\overrightarrow{\pi})}{\partial \pi_1} \right) > 0$$

and

$$\lim_{\overrightarrow{\pi} \to \overrightarrow{\pi}(a_1, a_2(a_1)), \overrightarrow{\pi} \in A_f} \left(\frac{\partial U_2(\overrightarrow{\pi})}{\partial \pi_2} - p(\widehat{a}_1, a_2(\widehat{a}_1)) \frac{\partial U_2(\overrightarrow{\pi})}{\partial \pi_1} \right) < 0$$

(where note that we are now holding the price fixed at $p(\widehat{a}_1, a_2(\widehat{a}_1))$ as a_1 varies). It follows that U_2 is locally normal in π_1 and π_2 at $(p(\widehat{a}_1, a_2(\widehat{a}_1)); I(\widehat{a}_1, a_2(\widehat{a}_1)))$.

Proof of part 2: Consider a small increase in FM's action $\widehat{a}_1' > \widehat{a}_1$. Assume (for contradiction) that SM weakly decreases his action, so that SM's material payoff rises while FM's falls. Call A the allocation $\overrightarrow{\pi}(\widehat{a}_1, a_2(\widehat{a}_1))$ and B the allocation $\overrightarrow{\pi}(\widehat{a}_1', a_2(\widehat{a}_1'))$. In the (π_2, π_1) plane, A is northwest of B. Now draw two downward-sloping lines with slopes $-p(\widehat{a}_1, a_2(\widehat{a}_1))$ and $-p(\widehat{a}_1', a_2(\widehat{a}_1')) \leq -p(\widehat{a}_1, a_2(\widehat{a}_1))$ going through A and B, respectively; this inequality is implied by $\frac{\partial p}{\partial a_2} \leq 0$ (by Part 2 of the Technical Lemma) and $-\frac{\partial}{\partial a_1}\left(\frac{\partial \pi_1 / \partial a_2}{\partial \pi_2 / \partial a_2}\right) = \frac{\partial p}{\partial a_1} \geq 0$ (by hypothesis). If these two slopes are equal, then weak local normality is contradicted. We can therefore assume that $-p(\widehat{a}_1', a_2(\widehat{a}_1')) < -p(\widehat{a}_1, a_2(\widehat{a}_1))$, so that the slope of the line through B is steeper than the slope of the line through A.

The two lines will intersect at some generic point, say C. There are three cases. Case 1 is that C is strictly southeast of both A and B, and Case 2 is that C is strictly southeast of A and northwest of B. The proof in these two cases proceeds identically: The change from A to B can be decomposed into a substitution effect and an income effect. The substitution effect causes a move from A to a point A' weakly northwest of A. Because of weak normality in π_1, the income effect then makes us move from A' to B, where B needs to be weakly north of A'—and therefore weakly north of A. But B actually lies strictly south of A, a contradiction.

Case 3 is that C is strictly northwest of both A and B. The change from A to B can again be decomposed into a substitution effect and an income effect, where the substitution effect causes a move from A to a point A' weakly northwest of A. Because of weak normality in π_2, the income

effect then makes us move from A' to B, where B needs to be weakly west of A'—and therefore weakly west of A. But B actually lies strictly east of A, a contradiction.

□

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 $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$.*

Proof: We will prove that SM's favorite transaction exists, and deduce the result for FM by symmetry. If SM's favorite transaction exists, then joint-monotonicity implies that it must lie on the MPE frontier. Technical Lemma implies that there does in fact exist a maximizing material payoff pair on the MPE frontier, and it is unique. Since this payoff pair is on the MPE frontier, there is in turn exactly one transaction $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ that achieves these payoffs. To see that, we will work in the (a_1, a_2) plane and study the *material* indifference curves for FM and SM. At a MPE action pair, we must have a tangency between the material indifference curves: $-\frac{\partial \pi_1/\partial a_1}{\partial \pi_1/\partial a_2} = \frac{da_2}{da_1}\Big|_{\pi_1=\overline{\overline{\pi}}_1} = \frac{da_2}{da_1}\Big|_{\pi_2=\overline{\overline{\pi}}_2} = -\frac{\partial \pi_2/\partial a_1}{\partial \pi_2/\partial a_2}$. By A3, $\frac{d^2 a_2}{d(a_1)^2}\Big|_{\pi_1=\overline{\overline{\pi}}_1} \geq 0$ and $\frac{d^2 a_2}{d(a_1)^2}\Big|_{\pi_2=\overline{\overline{\pi}}_2} \leq 0$ with at least one of these equalities strict. It follows that SM's favorite transaction is unique.

FM's favorite material payoff pair $(\overline{\pi}_1, \overline{\pi}_2)$ is UPE because there is no alternative feasible material payoff pair that FM prefers. Analogously, SM's favorite material payoffs pair $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$ is UPE because there is no alternative feasible material payoff pair that SM prefers.

Note that no material payoff pair (π'_1, π'_2) that is strictly within the materially-feasible set can be UPE; by joint-monotonicity of U_2, there is some feasible material payoff pair $(\pi''_1, \pi''_2) \gg (\pi'_1, \pi'_2)$ that SM prefers, and FM also prefers (π''_1, π''_2) by monotonicity.

Finally, any material payoff pair (π'_1, π'_2) on the MPE frontier between $(\overline{\pi}_1, \overline{\pi}_2)$ and $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$ is UPE. For contradiction, suppose (π'_1, π'_2) is not UPE. Then there exists another material payoff pair (π''_1, π''_2) giving at least equally high utility to both players. We may assume (π''_1, π''_2) to be MPE; if not, then by joint-monotonicity, there exists an MPE material payoff pair giving yet higher utility to both players that we can use instead. Suppose $(\overline{\pi}_1, \overline{\pi}_2)$ is northwest of $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$ on the MPE frontier; the argument is analogous if the positioning is reversed. If (π''_1, π''_2) is northwest of (π'_1, π'_2) on the MPE frontier, then $(\pi''_1, \pi''_2), (\pi'_1, \pi'_2), \left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$ lie in that order along the MPE frontier, and $U_2(\pi''_1, \pi''_2) \geq U_2(\pi'_1, \pi'_2) < U_2\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$; but this contradicts the Technical Lemma. On the other

hand, if (π_1'', π_2'') is southeast of (π_1', π_2') on the MPE frontier, then $(\overline{\pi}_1, \overline{\pi}_2), (\pi_1', \pi_2'), (\pi_1'', \pi_2'')$ lie in that order along the MPE frontier, and $U_1(\overline{\pi}_1, \overline{\pi}_2) > U_1(\pi_1', \pi_2') \leq U_1(\pi_1'', \pi_2'')$; but this also contradicts the Technical Lemma.

□

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.*

Proof: We will prove that given any action $\widehat{a}_1$, the transaction $(\widehat{a}_1, a_2(\widehat{a}_1))$ resulting from the unique best-response $a_2(\widehat{a}_1)$ is MPE if and only if $(\widehat{a}_1, a_2(\widehat{a}_1))$ is SM's favorite transaction. The "if" direction follows immediately from the fact that SM's favorite transaction is MPE (Theorem 1), so we focus on the "only if" direction. Suppose $(\widehat{a}_1, a_2(\widehat{a}_1))$ is MPE but is not SM's favorite transaction $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. Every point on the MPE frontier $\overrightarrow{\pi}(a_1, a_2)$ touches exactly one budget curve, $B(a_1)$; the transaction (a_1, a_2) satisfies the MPE condition $\frac{\partial \pi_1/\partial a_1}{\partial \pi_1/\partial a_2} = \frac{\partial \pi_2/\partial a_1}{\partial \pi_2/\partial a_2}$, which implies $\left.\frac{d\pi_1}{d\pi_2}\right|_{MPE} = \frac{\partial \pi_1/\partial a_1}{\partial \pi_2/\partial a_1} = \frac{\partial \pi_1/\partial a_2}{\partial \pi_2/\partial a_2} = \left.\frac{d\pi_1}{d\pi_2}\right|_{B(a_1)}$, and therefore the budget curve is tangent to the MPE frontier at $\overrightarrow{\pi}(a_1, a_2)$. Hence SM's indifference curve passing through $\overrightarrow{\pi}(\widehat{a}_1, a_2(\widehat{a}_1))$ is tangent to the MPE frontier at $\overrightarrow{\pi}(\widehat{a}_1, a_2(\widehat{a}_1))$. So there is some $\overrightarrow{\pi}(a_1', a_2')$ on the MPE frontier between $\overrightarrow{\pi}(\widehat{a}_1, a_2(\widehat{a}_1))$ and $\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$, sufficiently close to $\overrightarrow{\pi}(\widehat{a}_1, a_2(\widehat{a}_1))$, such that $U_2(\overrightarrow{\pi}(a_1', a_2')) < U_2(\overrightarrow{\pi}(\widehat{a}_1, a_2(\widehat{a}_1)))$. But this contradicts the fact that U_2 is strictly decreasing as we move away from $\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ along the MPE frontier (as stated in Technical Lemma).

Finally, Theorem 1 states that SM's favorite transaction is UPE.

□

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.*

Proof: From Lemma 2, if FM chooses action $\overline{\overline{a}}_1$, SM will choose action $\overline{\overline{a}}_2$. The facts that $U_2(\overrightarrow{\pi}(0,0)) = 0$ and $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ is SM's favorite transaction imply that $U_2\left(\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\right) \geq 0$. Since some action other than $\overline{\overline{a}}_1$ may give FM an even higher utility than $U_1\left(\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\right) \geq 0$, 0 is a lower bound on FM's equilibrium utility. From Technical Lemma, the set of individually-rational transactions T is compact. Since $U_1(\overrightarrow{\pi}(a_1, a_2(a_1)))$ is continuous, there exists an optimal action a_1 in T. The result follows.

□

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 $U_1(\overrightarrow{\pi}(a_1, a_2(a_1))) \geq 0$.*

Proof: The fact that $(a_1, a_2(a_1))$ is SM's favorite transaction follows directly from Lemma 2. Suppose that $U_1(\overrightarrow{\pi}(a_1, a_2(a_1))) < 0$. Then FM would choose her outside option rather than taking action a_1, so $(a_1, a_2(a_1))$ is not an equilibrium. But this is a contradiction.

□

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. *$\frac{dp(a_1, a_2(a_1))}{da_1} = 0$.*
3. *U_2 is fairness-kinked, and $(a_1, a_2(a_1))$ is a fairness-rule optimum.*

Proof: SM's best-response function $a_2(a_1)$ solves the problem of choosing SM's most-preferred material payoff pair along the budget curve $B(a_1)$:

$$(\pi_1^*(a_1, a_2(a_1)), \pi_2^*(a_1, a_2(a_1))) = \arg\max_{\overrightarrow{\pi}} U_2(\overrightarrow{\pi}) \text{ subject to } \overrightarrow{\pi} \in B(a_1). \tag{1}$$

As described in the text and illustrated in Figure 2, the solution to this problem, $\overrightarrow{\pi}$, is the same as the solution to the standard consumer optimization where the budget line is the linear approximation to the budget curve at the solution $\overrightarrow{\pi}^*(a_1, a_2(a_1))$ to the problem (1):

$$(\widetilde{\pi}_1(p, I), \widetilde{\pi}_2(p, I)) = \arg\max_{\overrightarrow{\pi}} U_2(\overrightarrow{\pi}) \text{ subject to } \pi_1 + p\pi_2 = I, \tag{2}$$

where $p = p(a_1, a_2(a_1)) = -\left.\frac{d\pi_1}{d\pi_2}\right|_{B(a_1)}$ and $I = \pi_1^*(a_1, a_2(a_1)) + p(a_1, a_2(a_1))\pi_2^*(a_1, a_2(a_1))$. Since U_2 is either twice-continuously differentiable or fairness-kinked, $p(a_1, a_2(a_1))$, $I(a_1, a_2(a_1))$, $\widetilde{\pi}_1(p, I)$, and $\widetilde{\pi}_2(p, I)$ are all continuously differentiable functions. Now, there are two possible cases, depending on whether the change in FM's action leads to a change in p.

Case 1: $\frac{dp(a_1, a_2(a_1))}{da_1} \neq 0$. The Slutsky equation can be applied to find the effects on $\widetilde{\pi}_1$ and $\widetilde{\pi}_2$:

$$\begin{aligned}
\frac{d}{da_1}\widetilde{\pi}_1(p, I) &= \frac{d\widetilde{\pi}_1(p, \widetilde{\pi}_1 + p\widetilde{\pi}_2)}{dp} + \frac{\partial\widetilde{\pi}_1(p, I)}{\partial I}(\omega_1 - \pi_1^*) \\
\frac{d}{da_1}\widetilde{\pi}_2(p, I) &= \underbrace{\frac{d\widetilde{\pi}_2(p, \widetilde{\pi}_1 + p\widetilde{\pi}_2)}{dp}}_{\text{substitution effect}} + \underbrace{\frac{\partial\widetilde{\pi}_2(p, I)}{\partial I}(\omega_2 - \pi_2^*)}_{\text{income effect}}
\end{aligned}$$

where π_1^* and π_2^* are the solutions from (1), (ω_1, ω_2) is the material payoff pair where the original budget line intersects with the new budget line (in standard consumer theory, this intersection point would be interpreted as the endowment consumption bundle), and we omit writing the dependence of p and I on $(a_1, a_2(a_1))$ to avoid cluttering notation.

To calculate the income effect, we begin by finding (ω_1, ω_2). We suppress dependence on $a_2(a_1)$ by writing the equation for the budget line as $\pi_1(a_1) = I(a_1) - p(a_1)\pi_2(a_1)$. Since (ω_1, ω_2) is the intersection of the old budget line and the new budget line, it satisfies $\omega_1 = I(a_1) - p(a_1)\omega_2$ and $\omega_1 = I(a_1 + \Delta a_1) - p(a_1 + \Delta a_1)\omega_2$. Solving these two equations simultaneously gives $\omega_2 = \frac{I(a_1+\Delta a_1)-I(a_1)}{p(a_1+\Delta a_1)-p(a_1)} = \frac{\frac{I(a_1+\Delta a_1)-I(a_1)}{\Delta a_1}}{\frac{p(a_1+\Delta a_1)-p(a_1)}{\Delta a_1}}$, so for small Δa_1,

$$\omega_2 = \frac{dI(a_1)/da_1}{dp(a_1)/da_1} \text{ and } \omega_1 = I(a_1) - p(a_1)\omega_2.$$

We now calculate $(\omega_1 - \pi_1^*)$ and $(\omega_2 - \pi_2^*)$. Using the definition of I, $\frac{dI(a_1)}{da_1} = \frac{dp(a_1)}{da_1}\pi_2^*(a_1) + p(a_1)\frac{d\pi_2^*(a_1)}{da_1} + \frac{d\pi_1^*(a_1)}{da_1}$. Substituting and simplifying gives

$$\begin{aligned}(\omega_2 - \pi_2^*) &= \frac{p(a_1, a_2(a_1))\frac{d\pi_2^*(a_1,a_2(a_1))}{da_1} + \frac{d\pi_1^*(a_1,a_2(a_1))}{da_1}}{\frac{dp(a_1,a_2(a_1))}{da_1}} \\ &= \frac{p(a_1, a_2(a_1))\frac{\partial\pi_2^*(a_1,a_2(a_1))}{\partial a_1} + \frac{\partial\pi_1^*(a_1,a_2(a_1))}{\partial a_1}}{\frac{dp(a_1,a_2(a_1))}{da_1}} = 0.\end{aligned}$$

The second equality can be intepreted as an envelope condition: the indirect effect through a_2, $p(a_1, a_2(a_1))\frac{\partial\pi_2^*(a_1,a_2(a_1))}{\partial a_2} + \frac{\partial\pi_1^*(a_1,a_2(a_1))}{\partial a_2} = 0$, equals zero because, at a fixed $p = p(a_1, a_2(a_1))$, SM has maximized "income" by choosing the material payoff pair on the MPE frontier. The third equality follows from $p \equiv -\left.\frac{d\pi_1}{d\pi_2}\right|_{B(a_1)} = -\frac{\partial\pi_1(a_1,a_2)/\partial a_2}{\partial\pi_2(a_1,a_2)/\partial a_2}$ and the MPE condition, $\frac{\partial\pi_1^*(a_1,a_2)/\partial a_2}{\partial\pi_2^*(a_1,a_2)/\partial a_2} = \frac{\partial\pi_1^*(a_1,a_2)/\partial a_1}{\partial\pi_2^*(a_1,a_2)/\partial a_1}$. Now, substituting $\omega_2 = \pi_2^*$ into the equation for ω_1 gives $\omega_1 = I(a_1) - p(a_1)\pi_2^*$, but since this expression equals π_1^*, $(\omega_1 - \pi_1^*) = 0$. Therefore, starting from an MPE transaction, the income effect from a change in FM's action equals zero.

To calculate the substitution effect, we define $\widetilde{I}(p) = p\widetilde{\pi}_2 + \widetilde{\pi}_1$ and use the implicit function theorem on the first-order condition for problem (2), $\frac{\partial U_2(\widetilde{I}-p\widetilde{\pi}_2,\widetilde{\pi}_2)}{\partial\pi_2} - p\frac{\partial U_2(\widetilde{I}-p\widetilde{\pi}_2,\widetilde{\pi}_2)}{\partial\pi_1} = 0$:

$$\begin{aligned}\frac{d\widetilde{\pi}_2(p,\, \widetilde{\pi}_1 + p\widetilde{\pi}_2)}{dp} &= -\frac{\frac{\partial^2 U_2}{\partial\pi_1\partial\pi_2}\left(\frac{d\widetilde{I}(p)}{dp} - \widetilde{\pi}_2\right) - \frac{\partial U_2}{\partial\pi_1} - p\frac{\partial^2 U_2}{\partial(\pi_1)^2}\left(\frac{d\widetilde{I}(p)}{dp} - \widetilde{\pi}_2\right)}{\frac{\partial^2 U_2}{\partial(\pi_2)^2} - 2p\frac{\partial^2 U_2}{\partial\pi_1\partial\pi_2} + p^2\frac{\partial^2 U_2}{\partial(\pi_1)^2}} \\ &= -\frac{-\left(\frac{\partial U_2}{\partial\pi_1}\right)^3}{\frac{\partial^2 U_2}{\partial(\pi_2)^2}\left(\frac{\partial U_2}{\partial\pi_1}\right)^2 - 2\frac{\partial U_2}{\partial\pi_1}\frac{\partial U_2}{\partial\pi_2}\frac{\partial^2 U_2}{\partial\pi_1\partial\pi_2} + \left(\frac{\partial U_2}{\partial\pi_2}\right)^2\frac{\partial^2 U_2}{\partial(\pi_1)^2}} = \frac{1}{-\left.\frac{d^2\pi_2}{d(\pi_1)^2}\right|_{U_2(\pi_1^*,\pi_2^*)}},\end{aligned}$$

where the second equality follows from $\frac{d\widetilde{I}(p)}{dp} = \widetilde{\pi}_2$ and substituting SM's first-order condition for problem (2). A similar calculation yields $\frac{d\widetilde{\pi}_1(p,\, p\widetilde{\pi}_1+\widetilde{\pi}_2)}{dp} = \frac{p}{d^2\pi_2/d(\pi_1)^2\big|_{U_2\left(\pi_1^*,\pi_2^*\right)}}$.

An interior equilibrium transaction satisfies FM's first-order condition, which can be written in terms of the budget lines: $\frac{d}{da_1}U_1\left(\widetilde{\pi}_1\left(p,\, I\right),\, \widetilde{\pi}_2\left(p,\, I\right)\right) = 0$. (A4 combined with joint-monotonicity of U_2 ensures that SM's favorite transaction is indeed interior.) Using the income and substitution effects derived above,

$$\begin{aligned}\frac{d}{da_1}U_1\left(\widetilde{\pi}_1\left(p,\, I\right),\, \widetilde{\pi}_2\left(p,\, I\right)\right) &= \frac{\partial U_1}{\partial \pi_1}\frac{d}{da_1}\widetilde{\pi}_1\left(p,\, I\right) + \frac{\partial U_1}{\partial \pi_2}\frac{d}{da_1}\widetilde{\pi}_2\left(p,\, I\right) \\ &= \left(\frac{\partial U_1}{\partial \pi_2} - p\frac{\partial U_1}{\partial \pi_1}\right)\cdot \frac{1}{-\frac{d^2\pi_2}{d(\pi_1)^2}\Big|_{U_2\left(\pi_1^*,\pi_2^*\right)}}.\end{aligned}$$

Recall that $\frac{\partial U_1}{\partial \pi_1} > 0$ and $\frac{\partial U_1}{\partial \pi_2} > 0$ (from Lemma 1). Hence FM's first-order condition is satisfied only if (A) SM's indifference curve is kinked at (π_1^*, π_2^*), i.e., $d^2\pi_2/d\left(\pi_1\right)^2\Big|_{U_2\left(\pi_1^*,\pi_2^*\right)} = -\infty$; or (B) FM's favorite transaction is (π_1^*, π_2^*), i.e., $\frac{\partial U_1/\partial \pi_2}{\partial U_1/\partial \pi_1} = p$ at (π_1^*, π_2^*), which is also SM's favorite transaction.

Case 2: $\frac{dp(a_1, a_2(a_1))}{da_1} = 0$. Since there is no substitution effect, the new and old budget lines do not intersect at an "endowment" (ω_1, ω_2). In this case, the Slutsky equation is:

$$\begin{aligned}\frac{d}{da_1}\widetilde{\pi}_1\left(p,\, I\right) &= \frac{\partial \widetilde{\pi}_1\left(p,\, I\right)}{\partial I}\frac{dI\left(a_1, a_2\left(a_1\right)\right)}{da_1} \\ \frac{d}{da_1}\widetilde{\pi}_2\left(p,\, I\right) &= \underbrace{\frac{\partial \widetilde{\pi}_2\left(p,\, I\right)}{\partial I}\frac{dI\left(a_1, a_2\left(a_1\right)\right)}{da_1}}_{\text{income effect}}.\end{aligned}$$

Differentiating $I\left(a_1, a_2\left(a_1\right)\right) = p\left(a_1, a_2\left(a_1\right)\right)\pi_2\left(a_1, a_2\left(a_1\right)\right) + \pi_1\left(a_1, a_2\left(a_1\right)\right)$ gives

$$\begin{aligned}\frac{dI\left(a_1, a_2\left(a_1\right)\right)}{da_1} &= \left(\frac{\partial \pi_1}{\partial a_1} + p\frac{\partial \pi_2}{\partial a_1}\right) + \left(\frac{\partial \pi_1}{\partial a_2} + p\frac{\partial \pi_2}{\partial a_2}\right)\frac{da_2\left(a_1\right)}{da_1} + \frac{dp\left(a_1, a_2\left(a_1\right)\right)}{da_1}\pi_2 \\ &= \frac{\partial \pi_1}{\partial a_1} + p\frac{\partial \pi_2}{\partial a_1} = 0\end{aligned}$$

In the first line, the third term is zero by hypothesis, and the second term is zero using the envelope theorem as above. The third equality follows from an analogous envelope observation: for fixed $p = p\left(a_1, a_2\left(a_1\right)\right)$, FM's action a_1 maximizes income since $\left(a_1, a_2\left(a_1\right)\right)$ is MPE. Since the income effect is zero, FM's first-order condition is clearly satisfied: $\frac{d}{da_1}U_1\left(\widetilde{\pi}_1\left(p,\, I\right),\, \widetilde{\pi}_2\left(p,\, I\right)\right) = \frac{\partial U_1}{\partial \pi_1}\frac{d}{da_1}\widetilde{\pi}_1\left(p,\, I\right) + \frac{\partial U_1}{\partial \pi_2}\frac{d}{da_1}\widetilde{\pi}_2\left(p,\, I\right) = 0$.

□

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 $U_1\left(\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\right) \geq 0$*, then the unique equilibrium transaction is the efficient transaction* $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$.

Proof: Since the material payoff functions are globally conditionally transferable, the budget curves are all parallel lines with slope $-p \equiv -\left.\frac{d\pi_1}{d\pi_2}\right|_{B(a_1)} = \frac{\partial\pi_2(a_1,a_2)/\partial a_2}{\partial\pi_1(a_1,a_2)/\partial a_2} = -k$ for some $k > 0$. Because U_2 is normal, SM's best-response function $a_2(a_1)$ ensures that π_1 and π_2 are both strictly increasing in $I(a_1)$. Since U_1 is monotonic or purely self-regarding, FM maximizes her utility by taking the action $\widetilde{a}_1$ that maximizes $I(a_1)$. This is the action $\widetilde{a}_1 = \overline{\overline{a}}_1$ that induces SM's favorite transaction because that is the unique action that induces an MPE transaction (by Lemma 2). Since $U_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right) \geq 0$, this action gives FM at least as high utility as her outside option and is therefore the unique equilibrium.

□

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* $U_1\left(\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\right) \geq 0$*, then the unique equilibrium transaction is the efficient transaction* $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$.

Proof: We first show that $(\widehat{a}_1, \widehat{a}_2)$ exists and is the unique transaction satisfying $\pi_1(\widehat{a}_1, \widehat{a}_2) = \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$, $U_2\left(\overrightarrow{\pi}(\widehat{a}_1, \widehat{a}_2)\right) = 0$, and $\widehat{a}_1 < \overline{\overline{a}}_1$. Given A1, A3, and A4, clearly there is a unique material payoff pair on SM's $U_2 = 0$ indifference curve such that $\pi_1 = \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$, so $(\widehat{a}_1, \widehat{a}_2)$ exists. Call that material payoff pair $\left(\overline{\overline{\pi}}_1, \widehat{\pi}_2\right)$. Clearly $\widehat{\pi}_2 < \overline{\overline{\pi}}_2$ (since all feasible material payoff pairs $(\pi_1, \pi_2) \neq \left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$ with $\pi_1 = \overline{\overline{\pi}}_1$ have $\pi_2 < \overline{\overline{\pi}}_2$). In the remainder of this paragraph, we show that $(\widehat{a}_1, \widehat{a}_2)$ is unique and satisfies $(\widehat{a}_1, \widehat{a}_2) \ll \left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. Define $\widetilde{a}_2(a_1)$ implicitly by $\pi_1(a_1, \widetilde{a}_2(a_1)) = \overline{\overline{\pi}}_1$, which is a continuously differentiable, strictly increasing function (by A1): $\frac{d\widetilde{a}_2(a_1)}{da_1} = -\frac{\partial\pi_1/\partial a_1}{\partial\pi_1/\partial a_2} > 0$. It is also weakly convex:

$$\begin{aligned}
\frac{d^2\widetilde{a}_2(a_1)}{d(a_1)^2} &= \frac{-\frac{\partial^2\pi_1(a_1,\widetilde{a}_2(a_1))}{\partial(a_1)^2}\frac{\partial\pi_1}{\partial a_2} - \frac{\partial^2\pi_1(a_1,\widetilde{a}_2(a_1))}{\partial a_1\partial a_2}\frac{d\widetilde{a}_2(a_1)}{da_1}\frac{\partial\pi_1}{\partial a_2} + \frac{\partial\pi_1}{\partial a_1}\left(\frac{\partial^2\pi_1(a_1,\widetilde{a}_2(a_1))}{\partial a_1\partial a_2} + \frac{\partial^2\pi_1}{\partial(a_2)^2}\frac{d\widetilde{a}_2(a_1)}{da_1}\right)}{\left(\frac{\partial\pi_1}{\partial a_2}\right)^2} \\
&= \frac{-\frac{\partial^2\pi_1}{\partial(a_1)^2}\frac{\partial\pi_1}{\partial a_2} + \frac{\partial\pi_1}{\partial a_1}\frac{\partial^2\pi_1}{\partial(a_2)^2}\frac{d\widetilde{a}_2(a_1)}{da_1}}{\left(\frac{\partial\pi_1}{\partial a_2}\right)^2} \geq 0,
\end{aligned}$$

where the second equality follows from substituting $\frac{d\widetilde{a}_2(a_1)}{da_1} = -\frac{\partial\pi_1/\partial a_1}{\partial\pi_1/\partial a_2}$, and the inequality follows from A1, $\frac{\partial^2\pi_1}{\partial(a_1)^2} \leq 0$ (due to A3), and $\frac{d\widetilde{a}_2(a_1)}{da_1} > 0$. Define $\widetilde{\widetilde{a}}_2(a_1, \pi_2)$ implicitly by $\pi_2(a_1, \widetilde{\widetilde{a}}_2) = \pi_2$, which is a continuously differentiable function, strictly increasing in a_1, strictly decreasing in π_2,

and (due to A3) weakly concave in a_1. By A3, we also know that $\widetilde{a}_2(a_1)$ is strictly convex or $\widetilde{\widetilde{a}}_2(a_1, \pi_2)$ is strictly concave in a_1 (or both). From Theorem 1, we know there exists a unique a_1 such that $\widetilde{a}_2(a_1) = \widetilde{\widetilde{a}}_2\left(a_1, \overline{\overline{\pi}}_2\right)$, which is $\overline{\overline{a}}_1$. In the (a_1, a_2) plane, draw the graph of $\widetilde{a}_2(a_1)$ as an increasing, convex curve and the graph of $\widetilde{\widetilde{a}}_2\left(a_1, \overline{\overline{\pi}}_2\right)$ as an increasing, concave curve. These curves are tangent at $\overline{\overline{a}}_1$. Since $\widehat{\pi}_2 < \overline{\overline{\pi}}_2$ and $\widetilde{\widetilde{a}}_2(a_1, \pi_2)$ is decreasing in π_2, draw the graph of $\widetilde{\widetilde{a}}_2(a_1, \widehat{\pi}_2)$ as an upward shift of the graph of $\widetilde{\widetilde{a}}_2\left(a_1, \overline{\overline{\pi}}_2\right)$. There are two intersections of the graphs of $\widetilde{a}_2(a_1)$ and $\widetilde{\widetilde{a}}_2(a_1, \widehat{\pi}_2)$, one with $(a_1, a_2) \gg \left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ and one with $(a_1, a_2) \ll \left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. The latter is $(\widehat{a}_1, \widehat{a}_2)$.

We next show that $\overline{\overline{a}}_1$ is a local optimum for FM. By hypothesis, $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ is a strict fairness-rule optimum. Therefore, Part 1 of Proposition 3 implies that $(a_1, a_2(a_1))$ is a strict fairness-rule optimum for all a_1 in a neighborhood of $\overline{\overline{a}}_1$. It follows that $\overline{\overline{a}}_1$ is a *local* optimum for FM, regardless of whether her distributional preferences are purely self-regarding or monotonic.

In the remainder of the proof, we show that $\overline{\overline{a}}_1$ is a *global* optimum for FM. To avoid cluttering notation with limits, we define the function U_2^D, which fully characterizes SM's preferences in the region of disadvantageous unfairness but is everywhere twice-continuously differentiable: $U_2^D\left(\overrightarrow{\pi}\right) \equiv U_2\left(\overrightarrow{\pi}\right)$ for all $\overrightarrow{\pi} \in D_f$, and $\left(\frac{\partial U_2^D\left(\overrightarrow{\pi}\right)}{\partial \pi_1}, \frac{\partial U_2^D\left(\overrightarrow{\pi}\right)}{\partial \pi_2}\right) = \lim_{\overrightarrow{\pi}' \to \overrightarrow{\pi}, \overrightarrow{\pi}' \in D_f} \left(\frac{\partial U_2\left(\overrightarrow{\pi}'\right)}{\partial \pi_1}, \frac{\partial U_2\left(\overrightarrow{\pi}'\right)}{\partial \pi_2}\right)$ for all $\overrightarrow{\pi} \in \text{graph}(f)$. (We do not constrain U_2^D in the region of advantageous unfairness because we will not use it there.) Now, it will be helpful in what follows to prove a preparatory claim.

Preparatory claim: We claim that

$$\frac{\partial U_2^D}{\partial \pi_2} - p(\overline{\overline{a}}_1, \widehat{a}_2)\frac{\partial U_2^D}{\partial \pi_1} > 0$$

at *all* individually-rational transactions (a_1, a_2) such that $\pi_1(a_1, a_2) > \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$.

To prove it, suppose to the contrary there were some $\overrightarrow{\pi}(a_1, a_2)$ at which $\frac{\partial U_2^D}{\partial \pi_2} - p(\overline{\overline{a}}_1, \widehat{a}_2)\frac{\partial U_2^D}{\partial \pi_1} \leq 0$; we will first show that $\frac{\partial U_2^D}{\partial \pi_1} \geq 0$. There are two cases. When $\frac{\partial U_2^D}{\partial \pi_2} - p(\overline{\overline{a}}_1, \widehat{a}_2)\frac{\partial U_2^D}{\partial \pi_1} = 0$, $\frac{\partial U_2^D}{\partial \pi_1}$ and $\frac{\partial U_2^D}{\partial \pi_2}$ have the same sign since $p(\overline{\overline{a}}_1, \widehat{a}_2) > 0$, and so $\frac{\partial U_2^D}{\partial \pi_1} \geq 0$ there (else joint-monotonicity is violated). And when $\frac{\partial U_2^D}{\partial \pi_2} - p(\overline{\overline{a}}_1, \widehat{a}_2)\frac{\partial U_2^D}{\partial \pi_1} < 0$, we must again have $\frac{\partial U_2^D}{\partial \pi_1} \geq 0$ (else $\frac{\partial U_2^D}{\partial \pi_2} < 0$, violating joint-monotonicity). Now that we have established that $\frac{\partial U_2^D}{\partial \pi_1} \geq 0$, we know that by choosing a value k slightly larger than $p(\overline{\overline{a}}_1, \widehat{a}_2)$, we must have

$$\frac{\partial U_2^D}{\partial \pi_2} - k\frac{\partial U_2^D}{\partial \pi_1} < 0$$

at $\overrightarrow{\pi}(a_1, a_2)$. Since k is very close to $p(\overline{\overline{a}}_1, \widehat{a}_2)$, using S1, we also know that

$$\frac{\partial U_2^D}{\partial \pi_2} - k\frac{\partial U_2^D}{\partial \pi_1} > 0$$

at $\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. Drawing budget lines l, l' each with slope $-k$ passing through the two points $\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ and $\overrightarrow{\pi}(a_1, a_2)$, respectively, the above inequalities imply that SM's most-preferred point on l is below $\pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ and his most-preferred point on l' is above $\pi_1(a_1, a_2)$. By assumption, $\pi_1(a_1, a_2) > \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. Since $\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ lies on the MPE frontier, which is downward sloping and concave, l is to the right of l'. So S3 (the normality assumption) is violated; a contradiction. This proves the Preparatory Claim.

We will prove that $\overline{\overline{a}}_1$ is the global optimum for FM in two cases, but before proceeding, we note three useful facts.

First, at any individually-rational transaction such that $\pi_1(a_1, a_2) = \pi_1(\widehat{a}_1, \widehat{a}_2)$ and $\pi_2(a_1, a_2) > \pi_2(\widehat{a}_1, \widehat{a}_2)$ we must have $(a_1, a_2) \gg (\widehat{a}_1, \widehat{a}_2)$. Suppose not. In that case, since A1 rules out $a_1 \geq \widehat{a}_1$ and $a_2 \leq \widehat{a}_2$ or vice-versa, it must be that $(a_1, a_2) \ll (\widehat{a}_1, \widehat{a}_2)$. Assuming for now that $\frac{da_2}{da_1}\Big|_{\pi_1=\pi_1(\widehat{a}_1,\widehat{a}_2)} < \frac{da_2}{da_1}\Big|_{\pi_2=\pi_2(\widehat{a}_1,\widehat{a}_2)}$, then by A1 and weak concavity of π_2 (from A3), $\pi_2(a_1, a_2) < \pi_2(\widehat{a}_1, \widehat{a}_2)$; a contradiction. We now show that $\frac{da_2}{da_1}\Big|_{\pi_1=\pi_1(\widehat{a}_1,\widehat{a}_2)} < \frac{da_2}{da_1}\Big|_{\pi_2=\pi_2(\widehat{a}_1,\widehat{a}_2)}$. Recall from the argument in the first paragraph of this proof that $(\widehat{a}_1, \widehat{a}_2)$ is the unique intersection of the graphs of $\widetilde{a}_2(a_1)$ and $\widetilde{\widetilde{a}}_2(a_1, \widehat{\pi}_2)$ such that $(\widehat{a}_1, \widehat{a}_2) \ll \left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. Since the graphs of $\widetilde{a}_2(a_1)$ and $\widetilde{\widetilde{a}}_2\left(a_1, \overline{\overline{\pi}}_2\right)$ are tangent at $\overline{\overline{a}}_1$, $\frac{d\widetilde{a}_2\left(\overline{\overline{a}}_1\right)}{da_1} = \frac{\partial\widetilde{\widetilde{a}}_2\left(\overline{\overline{a}}_1, \overline{\overline{\pi}}_2\right)}{\partial a_1}$. Since $\widetilde{a}_2(a_1)$ is increasing and convex, $\frac{d\widetilde{a}_2(\widehat{a}_1)}{da_1} < \frac{d\widetilde{a}_2\left(\overline{\overline{a}}_1\right)}{da_1}$. Due to S2 (in particular, SM's material payoff function being additively separable), $\widetilde{\widetilde{a}}_2(a_1, \pi_2)$ is additively separable, and since it is also increasing and concave in a_1, $\frac{\partial\widetilde{\widetilde{a}}_2(\widehat{a}_1,\pi_2(\widehat{a}_1,\widehat{a}_2))}{\partial a_1} = \frac{\partial\widetilde{\widetilde{a}}_2\left(\widehat{a}_1,\overline{\overline{\pi}}_2\right)}{\partial a_1} > \frac{\partial\widetilde{\widetilde{a}}_2\left(\overline{\overline{a}}_1,\overline{\overline{\pi}}_2\right)}{\partial a_1}$. Combining these observations and noting that $\frac{d\widetilde{a}_2(\widehat{a}_1)}{da_1} = \frac{da_2}{da_1}\Big|_{\pi_1=\pi_1(\widehat{a}_1,\widehat{a}_2)}$ and $\frac{\partial\widetilde{\widetilde{a}}_2(\widehat{a}_1,\pi_2(\widehat{a}_1,\widehat{a}_2))}{\partial a_1} = \frac{da_2}{da_1}\Big|_{\pi_2=\pi_2(\widehat{a}_1,\widehat{a}_2)}$, the needed inequality follows.

Second, due to S2 (the material payoff functions being additively separable), the slope of any budget curve, $p(a_1, a_2)$, does not depend on a_1. Therefore, $\frac{\partial p(a_1,a_2)}{\partial a_1} = 0$, and it thus follows from the Technical Lemma that $\frac{\partial p(a_1,a_2)}{\partial a_2} \leq 0$.

Third, since $\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right) \in \text{graph}(f)$, and since SM's fairness rule is strictly increasing, any (π_1, π_2) with $\pi_1 \geq \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ is in the region of disadvantageous unfairness, and thus we can use U_2^D.

Case 1: FM is purely self-regarding. To prove that $\overline{\overline{a}}_1$ is the unique global optimum for FM, it is sufficient to show that there does not exist any individually-rational transaction $(a_1, a_2(a_1)) \neq \left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ that satisfies $\pi_1(a_1, a_2(a_1)) \geq \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. Suppose to the contrary that there exists an individually-rational transaction $(a_1', a_2(a_1')) \neq \left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ such that $\pi_1(a_1', a_2(a_1')) \geq \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$.

We first show that without loss of generality, we can assume that $(a_1', a_2(a_1')) \gg \left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. By A1, the only other possibility is $(a_1', a_2(a_1')) \ll \left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. But in that case, there exists $(a_1'', a_2'') \gg$

$\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ such that $\overrightarrow{\pi}\left(a_1'', a_2''\right) = \overrightarrow{\pi}(a_1', a_2\left(a_1'\right))$ (this follows from an argument similar to that in the first paragraph of this proof). Since $a_2'' > a_2\left(a_1'\right)$ and $\frac{\partial p(a_1,a_2)}{\partial a_2} \leq 0$, the change from the budget line through $(a_1', a_2\left(a_1'\right))$ to the budget line through (a_1'', a_2'') is a Slutsky-compensated decrease in the price of FM's material payoff, and thus SM chooses a higher material payoff for FM: $\pi_1(a_1'', a_2\left(a_1''\right)) \geq \pi_1(a_1'', a_2'')$. Hence in the remainder of the proof, we can simply use $(a_1'', a_2\left(a_1''\right))$ in place of $(a_1', a_2\left(a_1'\right))$ and relabel it as $(a_1', a_2\left(a_1'\right))$.

Because $\overrightarrow{\pi}(a_1', a_2\left(a_1'\right))$ lies strictly in the interior of the region of disadvantageous unfairness and $a_2\left(a_1'\right)$ is a best response, $\frac{\partial U_2^D}{\partial \pi_2} - p(a_1', a_2\left(a_1'\right))\frac{\partial U_2^D}{\partial \pi_1} = 0$ at $\overrightarrow{\pi}(a_1', a_2\left(a_1'\right))$. We now show that $\pi_1(a_1', a_2\left(a_1'\right)) = \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ leads to a contradiction (and then turn in the next paragraph to the case $\pi_1(a_1', a_2\left(a_1'\right)) > \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$). Clearly $\pi_2(a_1', a_2\left(a_1'\right)) < \pi_2\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. Now consider $(a_1''', a_2''') \gg (a_1', a_2\left(a_1'\right))$ such that $\pi_1\left(a_1''', a_2'''\right) = \pi_1(a_1', a_2\left(a_1'\right))$, $\pi_2(a_1', a_2\left(a_1'\right)) < \pi_2\left(a_1''', a_2'''\right) < \pi_2\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ (the existence of such a transaction follows from an argument similar to that in the first paragraph of this proof). We know that

$$\begin{aligned}
& \frac{\partial U_2^D\left(\overrightarrow{\pi}(a_1', a_2\left(a_1'\right))\right)}{\partial \pi_2} - p(a_1', a_2\left(a_1'\right))\frac{\partial U_2^D\left(\overrightarrow{\pi}(a_1', a_2\left(a_1'\right))\right)}{\partial \pi_1} \\
> & \frac{\partial U_2^D\left(\overrightarrow{\pi}\left(a_1''', a_2'''\right)\right)}{\partial \pi_2} - p(a_1', a_2\left(a_1'\right))\frac{\partial U_2^D\left(\overrightarrow{\pi}\left(a_1''', a_2'''\right)\right)}{\partial \pi_1} \\
\geq & \frac{\partial U_2^D\left(\overrightarrow{\pi}\left(a_1''', a_2'''\right)\right)}{\partial \pi_2} - p(\overline{\overline{a}}_1, \widehat{a}_2)\frac{\partial U_2^D\left(\overrightarrow{\pi}\left(a_1''', a_2'''\right)\right)}{\partial \pi_1},
\end{aligned}$$

where the first inequality follows from S3 (the normality of U_2), and the second inequality follows from $a_2\left(a_1'\right) > \widehat{a}_2$, $\frac{\partial p(a_1,a_2)}{\partial a_2} \leq 0$, and $\frac{\partial p(a_1,a_2)}{\partial a_1} = 0$. Therefore,

$$\frac{\partial U_2^D\left(\overrightarrow{\pi}\left(a_1''', a_2'''\right)\right)}{\partial \pi_2} - p(\overline{\overline{a}}_1, \widehat{a}_2)\frac{\partial U_2^D\left(\overrightarrow{\pi}\left(a_1''', a_2'''\right)\right)}{\partial \pi_1} < 0,$$

but this is a contradiction because the Preparatory Claim implies that the left-hand side is ≥ 0.

We now show that $\pi_1(a_1', a_2\left(a_1'\right)) > \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ also leads to a contradiction. By A1, there is a unique transaction a_2' that satisfies $\pi_1(a_1', a_2') = \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$, where $a_2' < a_2\left(a_1'\right)$. Draw budget lines m, m' with respective slopes $-p(a_1', a_2')$ and $-p(a_1', a_2\left(a_1'\right))$ passing through the two points $\overrightarrow{\pi}(a_1', a_2')$ and $\overrightarrow{\pi}(a_1', a_2\left(a_1'\right))$. We know that at $\overrightarrow{\pi}(a_1', a_2')$, $\frac{\partial U_2^D}{\partial \pi_2} - p(a_1', a_2')\frac{\partial U_2^D}{\partial \pi_1} \geq \frac{\partial U_2^D}{\partial \pi_2} - p(\overline{\overline{a}}_1, \widehat{a}_2)\frac{\partial U_2^D}{\partial \pi_1} \geq 0$, where the first inequality follows from $a_2' > \widehat{a}_2$, $\frac{\partial p(a_1,a_2)}{\partial a_2} \leq 0$, and $\frac{\partial p(a_1,a_2)}{\partial a_1} = 0$, and the second inequality follows from the Preparatory Claim. Therefore, SM's most-preferred point on line m yields a material payoff for FM that is weakly smaller than $\pi_1(a_1', a_2')$. By construction, SM's most-preferred point on line m' is $\overrightarrow{\pi}(a_1', a_2\left(a_1'\right))$. Now, draw a third line m'' with slope $-p(a_1', a_2') \leq -p(a_1', a_2\left(a_1'\right))$ going through $\overrightarrow{\pi}(a_1', a_2\left(a_1'\right))$. Since moving from m' to m'' can be

thought of as a Slutsky-compensated price change, SM's most-preferred point on line m'' must yield a material payoff for FM that is at least as large as $\pi_1(a_1', a_2')$. But comparing FM's material payoff when moving from m'' to m reveals a violation of U_2 being normal in π_1 (the assumption S3); a contradiction.

Case 2: FM's distributional preferences are strictly monotonic, and S4 and S5 hold. We claim that there is no $a_1' \neq \overline{\overline{a}}_1$ such that $U_1(a_1', a_2(a_1')) \geq U_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. We showed in Case 1 that there is no $a_1' \neq \overline{\overline{a}}_1$ such that $\pi_1(a_1', a_2(a_1')) \geq \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. The result then follows from the observation that, since U_1 is monotonic, $\pi_1(\overline{a}_1, \overline{a}_2) > \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ (from S4), and U_1 is weakly quasi-concave (from S5), the region enclosed by the upper-contour set of FM's $U_1 = U_1\left(\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\right)$ indifference curve and MPE frontier contains only material payoff pairs satisfying $\pi_1(a_1, a_2) > \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. This completes the proof.

□

Theorem A1. *Suppose U_1 and U_2 are 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 is a connected set that includes $(\overline{\pi}_1, \overline{\pi}_2)$ and $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$ and lies within the region enclosed by $\overline{\overline{IC}}_1$, $\overline{IC}_2$, and the MPE frontier.*

Proof: The proofs that $(\overline{a}_1, \overline{a}_2)$ and $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ exist, are unique, and are UPE are the same as in the proof of Theorem 1.

To see that the set of UPE material pairs is a connected set, consider the problem $\overrightarrow{\pi}\left(\overline{U}_2\right) \in \arg\max_{\left\{\overrightarrow{\pi}:\overrightarrow{\pi}\in T_\pi, U_2\left(\overrightarrow{\pi}\right)=\overline{U}_2\right\}} U_1\left(\overrightarrow{\pi}\right)$. The Maximum Theorem (e.g., Sundaram 1996, p.235) implies that $\overrightarrow{\pi}\left(\overline{U}_2\right)$ is an upper-hemicontinuous correspondence. It follows that $\left\{\overrightarrow{\pi}\left(\overline{U}_2\right)\right\}_{\overline{U}_2\in\left[0, U_2\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)\right]}$ is a connected set. But $\left\{\overrightarrow{\pi}\left(\overline{U}_2\right)\right\}_{\overline{U}_2\in\left[0, U_2\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)\right]}$ is exactly the set of UPE material payoff pairs.

There does not exist a UPE material payoff pair $(\widehat{\pi}_1, \widehat{\pi}_2)$ outside of the region enclosed by $\overline{\overline{IC}}_1$, $\overline{IC}_2$, and the MPE frontier because by construction $(\widehat{\pi}_1, \widehat{\pi}_2)$ is worse than $(\overline{\pi}_1, \overline{\pi}_2)$ or $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$ for *both* FM and SM.

□

Theorem A2. *Suppose U_1 and U_2 are joint-monotonic and quasi-concave, and both are either twice-continuously differentiable or fairness-kinked. If the equilibrium $(a_1, a_2(a_1))$ is UPE and not MPE, then $(a_1, a_2(a_1))$ is a fairness-rule optimum for SM. If, in addition, $(a_1, a_2(a_1))$ is a strict fairness-rule optimum for SM and any fairness rule is continuously differentiable, then at least one of the following must be true: $\overrightarrow{\pi}(a_1', a_2(a_1'))$*

1. *SM's indifference curve for disadvantageously unfair transactions is tangent to SM's fairness rule at* $\overrightarrow{\pi}(a_1, a_2(a_1))$.

2. U_1 *is fairness-kinked,* $\overrightarrow{\pi}(a_1, a_2(a_1))$ *is on FM's fairness rule, and the respective fairness rules* f_1 *and* f_2 *have different slopes at* $\overrightarrow{\pi}(a_1, a_2(a_1))$.

Proof: We begin with the first claim: if the equilibrium $(a_1, a_2(a_1))$ is UPE and not MPE, then $(a_1, a_2(a_1))$ is a fairness-rule optimum for SM. We will prove that if the equilibrium $(a_1, a_2(a_1))$ is UPE and if U_2 is continuously differentiable at $\overrightarrow{\pi}(a_1, a_2(a_1))$, then $\overrightarrow{\pi}(a_1, a_2(a_1))$ is also MPE. Suppose not. Then $\overrightarrow{\pi}(a_1, a_2(a_1))$ is in the interior of the materially-feasible set. Lemma 1 implies that SM's distributional preferences are (locally) monotonic in a neighborhood of $\overrightarrow{\pi}(a_1, a_2(a_1))$. Since FM's distributional preferences are joint-monotonic, there is an alternative material payoff pair giving higher material payoff to both players that both players prefer. This contradicts UPE.

From now on, we assume that the equilibrium $(a_1, a_2(a_1))$ is UPE, is not MPE, and is a strict fairness-rule optimum for SM, and we assume that any fairness rule is continuously differentiable.

We next show that if U_1 is continuously differentiable at $\overrightarrow{\pi}(a_1, a_2(a_1))$, then SM's indifference curve for disadvantagously unfair transactions is tangent to SM's fairness rule at $\overrightarrow{\pi}(a_1, a_2(a_1))$. For contradiction, suppose that SM's indifference curve for disadvantagously-unfair transactions is not tangent to SM's fairness rule at $\overrightarrow{\pi}(a_1, a_2(a_1))$: $\frac{d\pi_1}{d\pi_2}\Big|_{U_2^D = U_2^D\left(\overrightarrow{\pi}(a_1, a_2(a_1))\right)} \neq f'(\pi_2)$. By a similar argument to that in the previous paragraph, SM's distributional preferences cannot be locally monotonic; therefore, SM's interpersonal indifference curve at $\overrightarrow{\pi}(a_1, a_2(a_1))$ is upward-sloping. Since the indifference curve also lies in the region of disadvantageous unfairness, it must be that $\frac{d\pi_1}{d\pi_2}\Big|_{U_2^D = U_2^D\left(\overrightarrow{\pi}(a_1, a_2(a_1))\right)} > f'(\pi_2)$. Since $\overrightarrow{\pi}(a_1, a_2(a_1))$ is not MPE, we know that it is in the interior of the materially-feasible set, and this, together with $\overrightarrow{\pi}(a_1, a_2(a_1))$ being UPE, implies that $\frac{d\pi_1}{d\pi_2}\Big|_{U_1 = U_1\left(\overrightarrow{\pi}(a_1, a_2(a_1))\right)} \geq \frac{d\pi_1}{d\pi_2}\Big|_{U_2^D = U_2^D\left(\overrightarrow{\pi}(a_1, a_2(a_1))\right)}$. Since $(a_1, a_2(a_1))$ is a strict fairness-rule optimum for SM, Part 1 of Proposition 3 implies that there exists a slight deviation for FM such that SM's optimal response would yield a material-payoff pair slightly southwest on SM's fairness rule. But the above inequalities imply that $\frac{d\pi_1}{d\pi_2}\Big|_{U_1 = U_1\left(\overrightarrow{\pi}(a_1, a_2(a_1))\right)} > f'(\pi_2)$ at $\overrightarrow{\pi}(a_1, a_2(a_1))$, meaning that FM would prefer this alternative material-payoff pair, contradicting that $(a_1, a_2(a_1))$ is an equilibrium.

Finally, we show that if U_1 is fairness-kinked and $\overrightarrow{\pi}(a_1, a_2(a_1))$ is on FM's fairness rule, then FM's and SM's respective fairness rules f_1 and f_2 have different slopes at $\overrightarrow{\pi}(a_1, a_2(a_1))$. For contradiction, suppose instead that f_1 and f_2 have the same slope at $\overrightarrow{\pi}(a_1, a_2(a_1))$. Since $(a_1, a_2(a_1))$

is a strict fairness-rule optimum for SM, Part 1 of Proposition 3 implies that there exists a slight deviation for FM such that SM's optimal response would yield a material-payoff pair slightly northeast along SM's fairness rule. We know that FM would prefer a sufficiently small northeast movement along SM's fairness rule, contradicting that $(a_1, a_2(a_1))$ is an equilibrium.

□