Cornell University
ILR School

Cornell University ILR School
DigitalCommons@ILR

Institute for Compensation Studies

Centers, Institutes, Programs

6-15-2010

Social Preferences and the Efficiency of Bilateral Exchange

Daniel J. Benjamin
Cornell University, daniel.benjamin@gmail.com

Follow this and additional works at: http://digitalcommons.ilr.cornell.edu/ics

This Article is brought to you for free and open access by the Centers, Institutes, Programs at DigitalCommons@ILR. It has been accepted for inclusion in Institute for Compensation Studies by an authorized administrator of DigitalCommons@ILR. For more information, please contact jdd10@cornell.edu.

Social Preferences and the Efficiency of Bilateral Exchange

Abstract
Under what conditions do social 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 necessary conditions and sufficient conditions on the players' preferences for the outcome of their interaction to be Pareto efficient. The results have implications for interpreting the rotten kid theorem, gift exchange in the laboratory, and gift exchange in the field.

Keywords
social preferences, fairness, altruism, gift exchange, rotten kid theorem

Comments
Suggested Citation
Benjamin, D. J. (2010). *Social preferences and the efficiency of bilateral exchange* (ICS 2010-001). Retrieved [insert date] from Cornell University, ILR School, Compensation Research Initiative site: http://digitalcommons.ilr.cornell.edu/ics/1

A more recent version of this paper can be found here: http://digitalcommons.ilr.cornell.edu/ics/14/

This article is available at DigitalCommons@ILR: http://digitalcommons.ilr.cornell.edu/ics/1

Social Preferences and the Efficiency of Bilateral Exchange

Daniel J. Benjamin*
Cornell University and NBER and Institute for Social Research

June 15, 2010

Abstract

Under what conditions do social 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 necessary conditions and sufficient conditions on the players' preferences for the outcome of their interaction to be Pareto efficient. The results have implications for interpreting the rotten kid theorem, gift exchange in the laboratory, and gift exchange in the field. (100 words)

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

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

*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, Stefan Penczynski, Giacomo Ponzetto, 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 Number T32-AG00186 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 sequential 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 social preferences.

I analyze theoretically a simple, two-stage bilateral exchange game initially defined in terms of "material payoffs"—the standard payoffs from one's own consumption. Each of the two players in turn chooses how much of an action to take. The action increases the other player's material payoff but at the cost of reducing one's own material payoff. I assume that gains from trade are possible: both players could get a higher material payoff than their respective outside options if both took a positive amount of their action. However, I further assume that contracting is infeasible and 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 social 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 necessary conditions and sufficient conditions on the players' preferences for the outcome of their interaction to be Pareto efficient.

While much of the literature on social preferences assumes a particular model of distributional concerns, most commonly "inequity aversion" (e.g., Fehr, Klein, & Schmidt 2007), I study how results depend on general properties of social preferences that are shared by many specific models. In particular, three general properties of social preferences play a prominent role in the analysis. The first property a player's social preferences might have is "joint-monotonicity": given any initial transaction, *there exists* an alternative transaction that gives higher material payoff to both players that is preferred. This property is a weakening of the monotonicity property assumed by altruistic preferences, which says that given any initial transaction, *any* alternative transaction that gives higher material payoff to both players is preferred. Unlike with monotonic (altruistic)

preferences, an agent with joint-monotonic preferences may dislike it if only one player's material payoff increases. As such, joint-monotonicity accommodates the possibility of fairness concerns. The second property is "normality": both players' material payoffs enter the utility function as "normal goods." That is, holding constant the rate of tradeoff between the players' material payoffs, if the pie gets larger, the actor prefers that both material payoffs increase. The third property is that the utility function is "fairness-kinked." Such social preferences capture behavior that involves following a "fairness rule," which specifies one player's material payoff as a strictly increasing function of the other player's material payoff; an example is the "50-50 split rule," which specifies that one player gets the same material payoff as the other player. An actor with fairness-kinked social preferences prefers an outcome that exactly implements the fairness rule over a range of rates of tradeoff between the players' material payoffs. Joint-monotonicity and normality are satisfied by virtually all existing models of altruism and fairness, while fairness-kinkedness is satisfied by several leading fairness models (e.g., Fehr & Schmidt 1999; Charness & Rabin 2002).

The central results of the paper describe two main cases in which social preferences generate efficiency in bilateral exchange. The first case occurs when material payoffs are "conditionally transferable": the marginal impact of SM's action on both FM's material payoff and SM's material payoff is not affected by FM's action. A leading example is when both players' material payoffs are linear in SM's action, which is a standard assumption for situations where SM's action is a monetary payment to FM. When the material payoffs are conditionally transferable, two further conditions are sufficient to guarantee efficiency: SM's social preferences satisfy normality, and FM's preferences satisfy monotonicity (and not just joint-monotonicity). Intuitively, whenever the material payoffs are conditionally transferable, FM's action determines the total amount of material surplus to be divided, and SM's action determines the distribution of that surplus. If SM's social preferences are normal, then whenever FM's action makes the pie larger, SM's best-response action will lead to a distribution of the surplus that gives greater material payoff to both players on net. Given SM's best-response function, as long as FM is purely self-regarding or altruistic, FM will prefer to take the level of her action that maximizes the total amount of material surplus.

Even if the material payoffs are not conditionally transferable, there is a second case in which social preferences generate efficiency: SM's social preferences are kinked. This case arises when SM's social preferences are sufficiently fairness-kinked that he behaves in accordance with a fairness rule. If, in addition, FM's preferences satisfy monotonicity (and not just joint-monotonicity), then the equilibrium is efficient. Intuitively, whenever SM behaves in accordance with a fairness rule,

FM maximizes both his own material payoff and SM's material payoff by choosing the action that induces the highest achievable point on this fairness rule. Like in the first case, if FM is purely self-regarding or altruistic, FM will choose the action that generates an efficient outcome.

In either of the two cases sketched above, joint-monotonicity is sufficient for SM's social preferences for not for FM's. Even though SM's behavior causes the players' material incentives to become aligned, if FM is willing to accept a lower material payoff for herself in order to come out ahead of SM, the equilibrium may not be efficient.

The remainder of this paper is organized as follows. Section 2 describes the bilateral exchange game. Section 3 introduces the properties of social preferences. Preparatory results are contained in Section 4. Section 5 presents necessary conditions for the equilibrium to be efficient. Section 6 presents sufficient conditions. Section 7 relates the results to existing literature in behavioral economics and the economics of the family and applies the results to the rotten kid game and to gift exchange. Section 8 speculates about possible extensions, such as incorporating social signaling or intentions-based reciprocity into the social preferences. All proofs are in an Online Appendix.

2 The Bilateral Exchange Game

In this section, I introduce the bilateral exchange environment that I will analyze in the rest of the paper. The first mover (FM) chooses the level of a costly action a_1 that helps the second mover (SM). SM then chooses the level of a costly action a_2 that helps FM. In the classic gift exchange example, an employer pays a wage to a worker, and then the worker exerts some amount of effort. Define a **transaction** to be any pair of real numbers (a_1, a_2). (I allow any real values to avoid dealing with boundary conditions.) Rather than taking an action, either player can choose during her or his turn not to trade, in which case both players receive an outside option payoff, the same as if the action pair had been $(0, 0)$.[1]

I assume that FM's and SM's respective preferences can be represented by utility functions $U_1(\pi_1(a_1, a_2), \pi_2(a_1, a_2))$ and $U_2(\pi_1(a_1, a_2), \pi_2(a_1, a_2))$. That is, each player's utility is a function of the same subutility functions, π_1 and π_2, called **material payoff functions**. The interpretation

[1] An alternative assumption would be that if FM takes her outside option, the payoffs are as if the action pair were $(0, 0)$, but if SM takes his outside option after FM has already chosen a_1, the payoffs are as if the action pair were $(a_1, 0)$. The assumption in the text is weakly better for SM because SM could always get the payoff from $(a_1, 0)$ after FM has chosen a_1 by taking action $a_2 = 0$. Therefore, if SM chooses to trade with the assumption in the text, then SM would also choose to trade with this alternative assumption. The text is simpler to represent diagrammatically, and since the analysis primarily focuses on situations where trade occurs, the choice of outside option assumption makes little difference.

is that these "material payoffs" represent the purely self-regarding component of payoffs, and each player's preferences may depend on the other player's material payoff in addition to his/her own.[2] This formulation of preferences is common in applied theory because it allows the analyst to make assumptions separately on the material payoff functions, which define the game under study—the mapping between actions and self-regarding outcomes—and on the $U_1(\pi_1, \pi_2)$ and $U_2(\pi_1, \pi_2)$, which capture **social preferences** over how the material payoffs are allocated. For brevity, I will sometimes use the vector notation, $\overrightarrow{\pi}(a_1, a_2) \equiv (\pi_1(a_1, a_2), \pi_2(a_1, a_2))$.

I formalize the bilateral exchange game by making the following assumptions:

A1. *$\pi_1(a_1, a_2)$ and $\pi_2(a_1, a_2)$ are twice continuously-differentiable.*

A2. *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 $\frac{\partial \pi_2}{\partial a_2} < 0$.*

A3. *FM's material payoff function is weakly concave in the "goods" $(-a_1, a_2)$, and SM's material payoff function is weakly concave in the "goods" $(a_1, -a_2)$, with at least one player's material payoff function strictly concave in at least one of the arguments. This assumption helps ensure uniqueness of equilibrium.*

A4. *The outside option material payoffs and utilities are normalized to zero: $\pi_1(0,0) = \pi_2(0,0) = U_1(0,0) = U_2(0,0) = 0$. As a tie-breaker, I assume each player chooses to trade unless his or her outside option gives strictly higher utility.*

A5. *There are (material) gains from trade: $\frac{-\partial \pi_1(0,0)/\partial a_1}{\partial \pi_1(0,0)/\partial a_2} < \frac{\partial \pi_2(0,0)/\partial a_1}{-\partial \pi_2(0,0)/\partial a_2}$. If this condition is satisfied, then 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: $\frac{\partial \pi_2(0,0)}{\partial a_1} da_1 + \frac{\partial \pi_2(0,0)}{\partial a_2} da_2 > 0$.*

A6. *Since the action spaces are unbounded, the following technical condition helps ensure existence of optimal actions: For any $\widehat{a}_1$ and $\widehat{a}_2$, each of the mappings from one agent's action to a real number defined 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.*

I postpone until the next section discussion of properties of the social preferences.

[2] Dufwenberg, Heidhues, Kirchsteiger, Riedel, & Sobel (2008) refer to this traditional approach as a model of "well-being externalities" and have a careful discussion of alternative approaches.

Two prominent examples of bilateral exchange games are the gift exchange game and the rotten kid game:

Example 1. ***Gift exchange game (Akerlof 1982; Akerlof & Yellen 1990; Fehr, Kirchsteiger, & Riedl 1993).*** *FM is an employer who pays salary a_1 to a worker. Then SM, the worker, exerts effort level a_2. The material payoff functions are $\pi_1(a_1, a_2) = a_2 - a_1$ and $\pi_2(a_1, a_2) = v(a_1) - c(a_2)$, where v is the worker's concave benefit from money satisfying $v' > 0$ and $v'' \leq 0$; and c is his convex cost of effort satisfying $c' > 0$ and $c'' \geq 0$. In addition, either $c'' > 0$, $\lim_{x\to-\infty} c'(x) = 0$, and $\lim_{x\to\infty} c'(x) = \infty$; or $v'' < 0$, $\lim_{x\to-\infty} v'(x) = \infty$, and $\lim_{x\to\infty} v'(x) = 0$.*[3]

Example 2. ***Rotten kid game (Becker 1974).*** *FM is a child who chooses how much effort a_1 to exert to earn money for the family. Then SM, the parent, transfers some amount, a_2, of family income to the child. 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 convex cost of effort function (in dollars) satisfying $n' > 0$, $n'' > 0$, $\lim_{x\to-\infty} n'(x) = 0$, 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. 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.*

While the rotten kid game is a bilateral exchange game in its essential features, the players are usually assumed not to have an outside option. Other examples include the trust game (Berg, Dickhaut, & McCabe 1995) and a sequential public goods game.

There are also a few properties that material payoff functions could have that will be useful to define for the analysis and discussion later. An important property is **local conditional transferability at action pair** (a_1, a_2): in a neighborhood of (a_1, 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 $k > 0$. If the material payoff functions are locally conditionally transferable at some action pair, then conditional on FM's action, SM's action is locally a linear transfer of material payoff from SM to FM.

[3] Following Fehr, Kirchsteiger, & Riedl (1993), in order to rule out negative payoff values, most laboratory gift-exchange experiments use as material payoff functions: $\pi_1(a_1, a_2) = (k_1 - a_1)a_2$ and $\pi_2(a_1, a_2) = a_1 - c(a_2) - k_2$, where $k_1 > 0$ and k_2 are constants, and $a_1 \leq k_1$ and $a_2 \geq 0$ have restricted domain. As long as the equilibrium is in the interior of the domains of a_1 and a_2, all of the analysis in this paper will apply.

Definition 1. *The material payoff functions are* ***globally conditionally transferable*** *if there exist functions G, H, and Z and constant $k > 0$ such that*

$$\begin{aligned} \pi_1(a_1, a_2) &= -G(a_1) + Z(a_1, a_2) \\ \pi_2(a_1, a_2) &= H(a_1) - kZ(a_1, a_2). \end{aligned}$$

A special case of global conditional transferability is ***conditional quasi-linearity****, where $Z(a_1, a_2) = A(a_1) a_2$ for some function A. A special case of conditional quasi-linearity is* ***quasi-linearity in SM's action****, where $A(a_1) = 1$.*

If the material payoff functions are globally conditionally transferable, then they satisfy conditional transferability at (a_1, a_2) for all action pairs (a_1, a_2). Material payoffs that are quasi-linear in SM's action, $\pi_1(a_1, a_2) = -G(a_1) + a_2$ and $\pi_2(a_1, a_2) = H(a_1) - ka_2$, are often used to model situations where SM's action is a monetary payment. The material payoff functions in Example 2 satisfy quasi-linearity in SM's action. The material payoff functions in Example 1 satisfy quasi-linearity in SM's action if $c'' = 0$ but are not locally conditional transferability at any action pair if $c'' > 0$.

Another property that will come up later is (a_1, a_2)**-additive-separability**, the usual kind of additive separability that is satisfed by the material payoff functions in both Examples 1 and 2. However, an alternative property that also plays a role is an unusual form of additive-separability.

Definition 2. *The material payoff functions are (a_1, π_1)****-additively-separable*** *if there exist functions F, H, and Z such that*

$$\begin{aligned} \pi_1(a_1, a_2) &= Z(a_1, a_2) \\ \pi_2(a_1, a_2) &= H(a_1) - F(Z(a_1, a_2)), \end{aligned}$$

where $F' > 0$.

In this case, SM's material payoff function can be written as an additively-separable function of FM's action and FM's material payoff: $\pi_2 = H(a_1) - F(\pi_1)$. The material payoffs functions do not satisfy this property in either Example 1 or Example 2. However, they could describe a setting where FM's action both increases SM's material payoff directly, and generates a valuable asset that SM's action allocates. For example, an investor (FM) might invest an amount of money a_1 and pay 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. Material payoff functions

that are (a_1, π_1)-additively-separable are also globally conditionally transferable if $F'' = 0$, but are not if $F'' \neq 0$.

2.1 Equilibrium

The solution concept is subgame-perfect Nash equilibrium.

In a typical optimal contracting problem, an employer (FM) can make the wage a function of output, which is a noisy function of the worker's effort. The worker (SM) has an incentive to exert effort because the wage partly depends on effort. In contrast, in the bilateral exchange game, FM's action is a number, not a function of any variable (like output) that depends on SM's action. Therefore SM has no extrinsic incentive to choose a high level of his action. Clearly, if SM were purely selfish, with utility function $U_2(\pi_1, \pi_2) = \pi_2$, then regardless of FM's action, SM would take minimal action (here actually, negative infinity, since the actions are unbounded). There would be no exchange because FM would prefer her outside option. Hence, any exchange that occurs in equilibrium is a consequence of SM's social preferences. That is why this stark setting of no contracting and no repetition makes the implications of social preferences as clear as possible.

2.2 Efficiency

Recall that a transaction is defined to be Pareto efficient if there is no alternative transaction that could have made one party better off without making the other worse off. Here, there are two potentially relevant interpretations of Pareto efficiency, depending on whether the players' welfare is measured by material payoffs or by utilities.

Definition 3. *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 $U_1(\overrightarrow{\pi}(\widehat{a}_1, \widehat{a}_2)) \geq U_1(\overrightarrow{\pi}(a_1, a_2))$ and $U_2(\overrightarrow{\pi}(\widehat{a}_1, \widehat{a}_2)) \geq U_2(\overrightarrow{\pi}(a_1, a_2))$, at least one inequality strict.*

Definition 4. *A transaction (a_1, a_2) is **materially Pareto efficient (MPE)** if there is no other transaction $(\widehat{a}_1, \widehat{a}_2)$ such that $\pi_1(\widehat{a}_1, \widehat{a}_2) \geq \pi_1(a_1, a_2)$ and $\pi_2(\widehat{a}_1, \widehat{a}_2) \geq \pi_2(a_1, a_2)$, at least one inequality strict.*

If a transaction (a_1, a_2) is MPE, then I will also refer to the resulting material payoff pair $\overrightarrow{\pi}(a_1, a_2)$ as MPE; analogously for UPE. A transaction (a_1, a_2) is MPE if and only if at that transaction, the material-payoff marginal rates of substitution are equal: $\frac{\partial \pi_1(a_1,a_2)/\partial a_1}{\partial \pi_1(a_1,a_2)/\partial a_2} = \frac{\partial \pi_2(a_1,a_2)/\partial a_1}{\partial \pi_2(a_1,a_2)/\partial a_2}$. If the material payoff functions are globally conditionally transferable or (a_1, π_1)-additively-separable,

then there is a unique materially-efficient action for FM[4] (the converse also holds locally; see the discussion of Proposition 1 below, and see Dijkstra 2007). However, more generally, the level of a_1 that corresponds to an MPE transaction may depend on a_2.

If both players were purely self-regarding, then these two efficiency notions would coincide, but in general, they do not. It is sometimes argued that individuals obey social norms that do not maximize their material payoffs, even though the individuals' material payoffs describe their personal welfare (e.g., Sen 1973; Köszegi & Rabin 2008).[5] To the extent that individuals' social preferences reflect adherence to social norms, a social planner might be interested in promoting material Pareto efficiency rather than utility Pareto efficiency. On the other hand, if as usually assumed utility represents both behavior and welfare, then utility Pareto efficiency is the appropriate concept of social welfare.

This paper asks: What social preferences and material payoff functions for the players lead to a MPE equilibrium? A UPE equilibrium?

3 Social Preferences

3.1 Existing Models

Each player is assumed to maximize a continuous utility function defined over both players' material payoffs. The following three prominent fairness models and one prominent altruism model help motivate the analysis that follows, which is general enough to accommodate all of them. For concreteness, I state these models with respect to SM's social preferences.

- Fehr & Schmidt's (1999) "inequity-averse preferences" have the form

$$U_2(\pi_1, \pi_2) = \pi_2 - \alpha \max\{\pi_1 - \pi_2, 0\} - \beta \max\{\pi_2 - \pi_1, 0\}, \tag{1}$$

  where $\alpha \geq 0$ is SM's aversion to "disadvantageous unfairness" (FM earning more than SM), and $\beta \geq 0$ is his aversion to "advantageous unfairness" (SM earning more than FM).

[4] That is, the set of MPE transactions is $\{(\widehat{a}_1, a_2)\}_{a_2 \in \mathbb{R}}$, for a constant $\widehat{a}_1$. In principal-agent problems, it is standard to assume conditional quasi-linearity because having a unique efficient action for the agent simplifies the analysis (Grossman & Hart 1983).

[5] For example, Sen (1973, pp.253-254) writes: "In economic analysis individual preferences seem to enter in two different roles: preferences come in as determinants of behaviour and they also come in as the basis of welfare judgements...[However] mores and rules of behaviour [will] drive a wedge between behaviour and welfare. People's behaviour may still correspond to some consistent *as if* preference but a numerical representation of the *as if* preference cannot be interpreted as individual welfare. In particular, basing normative criteria, e.g., Pareto optimality, on these *as if* preferences poses immense difficulties."

- Bolton & Ockenfels's (2000) "Equity, Reciprocity, and Competition (ERC) preferences," written here in additively-separable form, are:

$$U_2(\pi_1, \pi_2) = \pi_2 - \omega \left(\frac{\pi_2}{\pi_2 + \pi_1} - \frac{1}{2} \right)^2, \tag{2}$$

  where $\omega \geq 0$ weights a quadratic loss in deviation from an equal split. These preferences are well defined as long as $\pi_1, \pi_2 > 0$. When applying their model to a gift-exchange game, Bolton & Ockenfels (2000, pp.183-187) instead use

$$U_2(\pi_1, \pi_2) = -|\pi_1 - \pi_2|. \tag{3}$$

- Charness & Rabin (2002) propose "social welfare preferences,"

$$U_2(\pi_1, \pi_2) = \pi_2 + \gamma \pi_1 + \delta \min\{\pi_1, \pi_2\}, \tag{4}$$

  where $\gamma \geq 0$ is SM's positive regard for FM, and $\delta \geq 0$ is his additional concern for whoever gains least from the transaction.

- Becker's (1974) model of altruism within the family assumes that $U_2(\pi_1, \pi_2)$ is twice-continuously differentiable, monotonically increasing in both arguments, quasi-concave, and normal (i.e., π_1 and π_2 enter U_2 as normal goods).

Applied economic analysis involving social preferences generally proceeds by studying the implications of one of the above models. Instead, I will study the behavioral implications of properties of social preferences in order to understand what features of these functional forms drive results, and which implications follow from which classes of models.

3.2 Properties of Social Preferences

Following convention, I describe the relevant monotonicity and concavity notions before getting to the key properties of normality and fairness-kinkedness.

A primary difference between altruistic preferences and fairness preferences is that altruistic preferences $U(\pi_1, \pi_2)$ are assumed to be monotonically increasing in both arguments, like consumption preferences over goods. In contrast, fairness preferences such as (1), (2), and (3) allow for a type of non-monotonicity: an individual may prefer to reduce the payoff of a person who is ahead.

Experimental economists disagree about the extent to which individuals are willing to reduce a player's material payoff in order to ensure a more equal allocation. In hypothetical choices, Bazerman, Loewenstein, & White (1992) found that 25% of experimental participants preferred receiving \$500 for themselves and \$500 for a friendly neighbor rather than receiving \$600 for themselves and \$800 for the neighbor. When the choice was between \$600 for each versus \$600 for themselves and \$800 for the neighbor, 68% chose the fair but inefficient outcome. In 3-player allocation problems with real money at stake, Fehr, Naef, & Schmidt (2006) found similar patterns. However, other researchers found that fewer participants make such materially Pareto inefficient choices (Charness & Rabin 2002; Engelmann & Strobel 2004). Relatedly, models of positional (or status) preferences also predict a willingness to sacrifice one's own material payoff to reduce others'.

It is important to allow for this empirically relevant kind of non-monotonicity—where individuals prefer to reduce a player's material payoff to reach a fairer allocation—in order to determine whether and how it matters. At the same time, it is useful to rule out too much spitefulness or self-hating, in which case an equilibrium might not exist. Joint-monotonicity is a new condition that appropriately weakens monotonicity.[6]

Definition 5. *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 a particular direction, a direction which jointly increases both players' material payoffs. Hence joint-monotonicity limits the extent to which an agent can be spiteful or self-hating, while permitting the possibility that at some transactions, increasing only one player's material payoff might reduce utility. Figures 1a and 1c show interpersonal indifference curves from social preferences satisfying joint-monotonicity, while Figure 1b illustrates preferences that violate it. All of the above models of social preferences satisfy joint-monotonicity.

The second condition, quasi-concavity, is familiar from consumer theory and social choice.

[6] In studying other-regarding preferences in a general equilibrium environment, Dufwenberg, Heidhues, Kirchsteiger, Riedel, & Sobel (2008) 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. Formally, U_1 and U_2 satisfy "social monotonicity" if for any $\varepsilon > 0$, there is some $(\widehat{\pi}_1, \widehat{\pi}_2)$ with $0 < \widehat{\pi}_1 - \pi_1 < \varepsilon$, $0 < \widehat{\pi}_2 - \pi_2 < \varepsilon$, and $U_i(\widehat{\pi}_1, \widehat{\pi}_2) > U_i(\pi_1, \pi_2)$ for $i = 1, 2$.

Definition 6. *U is **quasi-concave** if for any $(\pi_1, \pi_2), (\widehat{\pi}_1, \widehat{\pi}_2)$ such that $U(\pi_1, \pi_2) \leq U(\widehat{\pi}_1, \widehat{\pi}_2)$, $U(\pi_1, \pi_2) \leq U(\lambda\pi_1 + (1-\lambda)\widehat{\pi}_1, \lambda\pi_2 + (1-\lambda)\widehat{\pi}_2)$ for any $\lambda \in [0,1]$.*

For social preferences, quasi-concavity means that along an interpersonal indifference curve, the higher FM's material payoff, the less of SM's material payoff the decision-maker is willing to give up to increase the FM's material payoff (and vice-versa). Equivalently, it means that the upper level sets of U are convex, which is a helpful regularity condition. All of the above models of social preferences satisfy quasi-concavity.

A third potential assumption about U is that if the pie is larger, holding constant the rate of tradeoff in material payoffs, it is preferred that both players earn a higher material payoff. Since this thought experiment involves considering a linear tradeoff in material payoffs, it corresponds to the familiar assumption of "normal goods."

Definition 7. *Suppose $\widetilde{\pi}_1(p; I)$ and $\widetilde{\pi}_2(p; I)$, defined by*

$$(\widetilde{\pi}_1, \widetilde{\pi}_2) = \arg \max_{\{(\pi_1,\pi_2):\pi_1 + p\pi_2 = I\}} U(\pi_1, \pi_2),$$

*are finite, real-valued functions. 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 $I \in \mathbb{R}$ and $p > 0$. U is **(weakly) normal** if U is (weakly) normal in both π_1 and π_2.*

(The functions $\widetilde{\pi}_1(p; I)$ and $\widetilde{\pi}_2(p; I)$ will in fact be well defined given other assumptions on U.) Quah (2007, Theorem S1 and Proposition S1) shows that the statement about behavior "U is locally normal in π_i" is essentially equivalent to the following statement about indifference curves: for a small increase in π_i, holding constant the other player's material payoff, the marginal rate of substitution between the players' material payoffs becomes less favorable toward player i.[7] Becker's (1974) altruism model explicitly assumes normality, and all of the above fairness functional forms—(1), (2), (3), and (4)—also satisfy normality or weak normality.[8] Nonetheless, existing work has not recognized that normality is a strong and central assumption in generating fair-minded behavior because it has been assumed implicitly as a byproduct of specific functional forms.

[7] To be more precise, these statements are equivalent when $\frac{\partial U(\pi_1,\pi_2)}{\partial \pi_1} > 0$ and $\frac{\partial U(\pi_1,\pi_2)}{\partial \pi_2} > 0$ (see discussion in Quah 2007, p.6). These partial derivatives may not be everywhere positive when U is joint-monotonic 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.

[8] When actions are bounded, piecewise-linear functional forms like (1) and (4) satisfy only weak normality. However, requiring normality to be strict only matters for ensuring that the second-mover's action is strictly increasing in the first-mover's action (Lemma 2) and that the equilibrium is unique (Theorems 3 and 4).

A fourth property that will also turn out to be important is conformance to rules of fair behavior. For example, the "50-50 split rule" has been documented in many laboratory experiments. The simplest such experiment is a "dictator game," where one player allocates a given amount of money between himself and another player. Typically 20-30% of participants give exactly half of the money to the other player (Camerer 2003). Conformance to a rule of fair behavior can be modeled with kinked indifference curves.[9] Indeed, a kink around equal material payoffs is the feature of fairness models (1), (3), and (4) that allows them to explain the preponderance of equal splits. I will give additional examples after formally defining and explaining this property of social preferences.

Definition 8. *U is **potentially fairness-kinked** if it can be expressed as $U = \min\left\{U^A, U^B\right\}$, where U^A, U^B are twice-continuously differentiable utility functions satisfying:*

- *There exists some (π_1, π_2) at which $U^A(\pi_1, \pi_2) = U^B(\pi_1, \pi_2)$.*
- *If $U^A(\pi_1, \pi_2) \leq U^B(\pi_1, \pi_2)$, then $U^A(\widehat{\pi}_1, \pi_2) < U^B(\widehat{\pi}_1, \pi_2)$ for all $\widehat{\pi}_1 > \pi_1$.*
- *If $U^A(\pi_1, \pi_2) \geq U^B(\pi_1, \pi_2)$, then $U^A(\pi_1, \widehat{\pi}_2) > U^B(\pi_1, \widehat{\pi}_2)$ for all $\widehat{\pi}_2 > \pi_2$.*

*Moreover, U is **fairness-kinked at** (π_1, π_2) if U is potentially fairness-rule adherent and a kink occurs at (π_1, π_2): i.e., $U^A(\pi_1, \pi_2) = U^B(\pi_1, \pi_2)$ and $\frac{\partial U^A(\pi_1,\pi_2)/\partial \pi_2}{\partial U^A(\pi_1,\pi_2)/\partial \pi_1} > \frac{\partial U^B(\pi_1,\pi_2)/\partial \pi_2}{\partial U^B(\pi_1,\pi_2)/\partial \pi_1}$.*

(Note that if U^A and U^B are joint-monotonic and quasi-concave, then so is U.) Preferences that are potentially fairness-kinked have indifference curves that are everywhere smooth (to ensure that marginal analysis is applicable), except possibly at material payoff pairs along the graph of the "fairness rule."

Definition 9. *For a potentially fairness-kinked U, the **fairness rule** is the function $f(\pi_2)$ that, given a material payoff for the second-mover π_2, assigns the first-mover a material payoff according to $U^A(f(\pi_2), \pi_2) = U^B(f(\pi_2), \pi_2)$.*

Transactions that exactly satisfy the fairness rule are called **fair** transactions. Fairness-kinked preferences can be interpreted as social preferences that penalize deviations from the fairness rule.

[9] 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 utility function can explain equal-split behavior in both cases. See Andreoni & Bernheim (2009) for an alternative model of 50-50 split behavior based on signaling. As I discuss in more detail in Section 8, I conjecture that as long as the players' behavior can be represented *as if* it were generated by distributional social preferences, then the predictions about behavior will be essentially the same.

To see this, note that $U = \min\left\{U^A, U^B\right\}$ can be equivalently expressed as $U = \frac{U^B+U^A}{2} - \left|\frac{U^B-U^A}{2}\right|$. The first term can be thought of as a "standard" smooth utility function, while the second term represents disutility from not adhering to the fairness rule.

The single-crossing properties in the definition of "potentially fairness-kinked" have two useful implications. First, for SM, $U = U^A$ in 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 $U = U^B$ in the region of **advantageously unfair** transactions for SM. (Analogously, for FM, U^A represents preferences at advantageously unfair transactions, and U^B represents preferences at disadvantageously unfair transactions.) Second and more substantively, the fairness rule f is a strictly increasing function: when the pie increases, the fairness rule assigns a larger piece of pie to both players. Hence, fairness-kinked preferences are locally normal when the agent's optimum occurs at a fair transaction.[10] Figure 1c shows social preferences that are fairness-kinked.

If indifference curves are kinked, then the agent's optimum will occur at a kink point for a range of budget constraints. The standard consumer theory example of "perfect complements" is an extreme case where one segment of each indifference curve is vertical, the other segment is horizontal, and the kinks occur along the identity line. Because the indifference segments are vertical and horizontal, the optimum occurs at the kink for any downward-sloping, weakly concave budget curve passing through the kink point, regardless of its slope. Because the kink points occur along the identity line, the agent always prefers 1 unit of the first good for each unit of the second good. Fairness-kinked preferences generalize perfect complements. The indifference curve segments can be downward-sloping instead of vertical or horizontal, meaning that the optimum occurs at the kink point for a range of slopes, but if one of the goods is sufficiently inexpensive relative to the other, then the agent's optimum will occur away from the kink. Also with fairness-kinked preferences, the path of kink points—the fairness rule—can be any upward-sloping curve, not just the identity line.

Fairness-kinked preferences provide a simple model of behavior of an agent who follows a salient norm of fair behavior, as long as doing so is not too costly for one of the players. For example:

- *Splitting of money payoffs.* Besides laboratory experiments, the equal-split fairness rule has also been documented in a variety of field contexts, such as negotiations, asymmetric joint

[10] To be precise, suppose U is fairness-kinked, and the material payoff pair $(\widetilde{\pi}_1, \widetilde{\pi}_2)$ that maximizes U on budget line $\pi_1 + p\pi_2 = I$ occurs at a kink point: $U^A = U^B$, $\frac{\partial U^A}{\partial \pi_2} - p\frac{\partial U^A}{\partial \pi_1} < 0$, and $\frac{\partial U^B}{\partial \pi_2} - p\frac{\partial U^B}{\partial \pi_1} > 0$ evaluated at $(\widetilde{\pi}_1, \widetilde{\pi}_2)$. Then U is locally normal at $(I; p)$.

ventures among corporations, share tenancy in agriculture, and bequests to children (Andreoni & Bernheim 2009). However, financial contracts often apportion profit according to unequal percentages that are standard in the industry. Suppose a_1 and a_2 are the actions of the two parties, $m_1(a_1, a_2)$ and $m_2(a_1, a_2)$ are the resulting monetary profits, and the norm is that $m_1 = km_2$, with $k > 0$ possibly equal to 1. Suppose material payoff functions are strictly increasing functions of the amount of money earned, $\pi_1 = q_1(m_1)$ and $\pi_2 = q_2(m_2)$. Fairness-kinked preferences with the fairness rule $\pi_1 = q_1\left(kq_2^{-1}(\pi_2)\right)$ would represent behavior that splits money according to this rule.

- *Posted prices.* If a farmer leaves apples by the side of the road with a sign saying "\$1 per apple," then a motorist passing by may feel obliged to leave \$1 for each apple he takes (Dawes & Thaler 1988). Let a denote the number of apples taken by the motorist, let m denote the dollar amount of money he leaves, let $\pi_1(a, m)$ denote the farmer's material payoff, and let $\pi_2(a, m)$ denote the motorist's material payoff, where $\frac{\partial \pi_1}{\partial a} < 0$, $\frac{\partial \pi_1}{\partial m} > 0$, $\frac{\partial \pi_2}{\partial a} > 0$, and $\frac{\partial \pi_2}{\partial m} < 0$. The fairness rule is an implicit function characterized by $\{\overrightarrow{\pi}(a, m)\}_{a=m}$.

It is possible that FM and SM each have fairness-kinked preferences but with different fairness rules. This might occur if one or both parties has a self-serving perception of what is fair (as in Babcock & Loewenstein 1997).

Finally, I note two technical assumptions (TAs) that I will assume hold throughout the analysis. The first assumption is made necessary by the weakening of monotonicity. What matters for behavior is whether the decision-maker's *indifference curves* are kinked or smooth. 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.[11] TA1 ensures that the indifference curves are kinked if and only if U is kinked.

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).*

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

Second, I assume that at some sufficiently low material payoff for SM and high payoff for FM, the interpersonal indifference curves become vertical or upward-sloping, so that the player puts weakly negative weight on FM's material payoff. Analogously, at some sufficiently low material payoff for FM and high payoff for SM, the interpersonal indifference curves become horizontal or upward-sloping. This assumption ensure the existence of optimal actions for the players by helping to make the set of individually-rational transactions compact, but does not otherwise play a substantive role. Of course, this assumption would be violated when U is purely self-regarding, but I will not need it in those cases.

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

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

4 Some Preliminaries

4.1 Characterizing UPE Transactions

As is standard, call a transaction (a_1, a_2) **individually-rational** if both players earn at least their outside option: $U_1(\overrightarrow{\pi}(a_1, a_2)) \geq 0$ and $U_2(\overrightarrow{\pi}(a_1, a_2)) \geq 0$. Let

$$(\overline{a}_1, \overline{a}_2) \equiv \arg \max_{\left\{(a_1, a_2) | U_2\left(\overrightarrow{\pi}(a_1, a_2)\right) \geq 0\right\}} U_1(\overrightarrow{\pi}(a_1, a_2))$$

be called FM's **favorite transaction**, her most-preferred transaction among the individually-rational 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

$$\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right) \equiv \arg \max_{\left\{(a_1, a_2) | U_1\left(\overrightarrow{\pi}(a_1, a_2)\right) \geq 0\right\}} U_2(\overrightarrow{\pi}(a_1, a_2))$$

be called SM's **favorite transaction**, his most-preferred transaction among the individually-rational transactions, with corresponding material payoff pair $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$. 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 1. *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. In addition, if U_1 or U_2 (or both) is monotonic, then the set of UPE material payoff pairs coincides exactly with the set of material payoff pairs on the MPE frontier between $(\overline{\pi}_1, \overline{\pi}_2)$ and $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$.

Figure 2a 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 social 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 material payoff pair outside that region. A material payoff pair that is UPE either occurs at a tangency point between the players' indifference curves, or it occurs on the MPE frontier if the "relevant tangency" lies outside the set of feasible material payoff pairs.

Figure 2b illustrates the case where FM has monotonic social preferences. In this case, the set of UPE material payoff pairs is exactly the subset of the MPE frontier between $(\overline{\pi}_1, \overline{\pi}_2)$ and $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$. Graphically, there cannot be a tangency between the players' indifference curves in the interior of the feasible set because both players' indifference curves are downward-sloping. Intuitively, given any interior material payoff pair, there is an alternative material payoff pair to the northeast that SM strictly prefers because his preferences are joint-monotonic, and FM also prefers this alternative because his preferences are monotonic.[12]

If one or both of the players is purely self-regarding, then Theorem 1 does not technically apply. 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, $(\overline{\pi}_1, \overline{\pi}_2) = (\infty, -\infty)$, $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right) = (-\infty, \infty)$, or both.

4.2 SM's Behavior

Given FM's action a_1, SM can be thought of as selecting a pair of material payoffs on the **(material payoff) budget curve** $B(a_1) = \{\overrightarrow{\pi}(a_1, a_2)\}_{a_2 \in \mathbb{R}}$ by his choice of action a_2. For any value of a_1, the slope of the budget curve—the "price" of π_1 in terms of π_2—is denoted $p(a_1, a_2) \equiv -\left.\frac{d\pi_1}{d\pi_2}\right|_{B(a_1)} = -\frac{\partial \pi_1 / \partial a_2}{\partial \pi_2 / \partial a_2} > 0$. Given FM's action a_1, SM maximizes his preferences subject to this budget curve.

It will sometimes be useful in this section and later to apply the standard tools of consumer

[12] Dufwenberg, Heidhues, Kirchsteiger, Riedel, & Sobel (2008) independently prove a related result; their Theorem 3 implies that when at least one player has monotonic social preferences, material Pareto efficiency is a necessary condition for utility Pareto efficiency.

theory to analyze SM's behavior. To facilitate doing so, it will be helpful to approximate SM's budget curve with the budget *line* that is tangent to the budget curve at his optimal material payoff pair. Since SM's interpersonal indifference curves are convex, his optimum with respect to his budget curve will also be an optimum with respect to this budget line. At a transaction (a_1, a_2) that identifies a point $\overrightarrow{\pi}(a_1, a_2)$ on the budget curve $B(a_1)$, the equation for the budget line is $\pi_1 + p\pi_2 = I$, where $p = p(a_1, a_2)$ is the slope of the budget curve at that point, 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 whatever level of "income" just suffices for SM to afford his optimal material payoff pair. Figure 3 depicts a budget curve $B(a_1)$ and the approximating budget line $b(p, I)$.

Lemma 1 establishes results about SM's behavior that are helpful for backward-inducting the equilibrium.

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 social preferences are non-monotonic in general, they will be monotonic on the margin at his optimal action, as long as his optimum occurs at a smooth region of his indifference curves. Graphically, if his optimal action occurs on a smooth region of his indifference curves, then his optimal action must occur at a tangency point between an indifference curve and the budget curve. Since the budget curve is always downward-sloping in the space of material payoffs, the tangency point must occur at a downward-sloping region of the indifference curve (see Figure 3). 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.

Lemma 1 suggests 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. If SM's interpersonal indifference curves are everywhere smooth, then this conclusion applies directly. Even if SM's social preferences are fairness-kinked, there are only two possibilities. 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. Hence the lemma explains why conclusions that

hold when both players have monotonic preferences will generalize to the case where SM instead has joint-monotonic preferences.

The irrelevance of non-monotonicities in SM's preferences may seem surprising, given that the willingness of a second mover to reduce players' material payoffs is a prominent component of the evidence that motivates the models of fairness listed at the beginning of Section 3. The logic of the lemma applies in the bilateral exchange game because the budget curve is both downward-sloping (in the space of material payoffs) *and* continuous. If the budget curve were upward-sloping, SM's optimum could occur at a tangency between the budget set and an upward-sloping portion of his indifference curve, violating the conclusion of the lemma. If the budget set were discrete, SM's optimum could occur at an upward-sloping part of his indifference curve since there is no tangency condition for the optimum. In the prototypical laboratory payoff-allocation decision problem, there are two options available, with the "unfair" option giving higher material payoffs to both players than the "fair" option does. Since the budget set is discrete and "upward-sloping," Lemma 1 does not apply; an individual with joint-monotonic social preferences might prefer the materially-dominated "fair" option, even though an individual with monotonic social preferences never would.

Relatedly, Lemma 1 implies that SM, if he had the option, would never choose to "punish" FM for taking a low action by choosing a material payoff pair that is materially-dominated by some point on the budget curve. In contrast, in a two-player sequential-move game with *discrete* action spaces—such as the much-studied "ultimatum game" (Güth, Schmittberger, & Schwarze 1982)—a SM with the very same joint-monotonic social preferences might well choose to punish FM by choosing a material payoff pair that is worse for both players (see Rabin 1997).

The following proposition gives conditions under which SM's optimal action is increasing in FM's action. A straightforward observation is that if FM is purely self-regarding, a higher transfer by FM leads to a higher transfer by SM at any equilibrium. If it did not, FM could obtain higher material payoff (and hence higher utility) by reducing her action. However, the proposition is applicable away from equilibrium and even if FM is not purely self-regarding.

Proposition 1. *Suppose U_2 is joint-monotonic and quasi-concave. Suppose $\frac{\partial}{\partial a_1}\left(\frac{\partial \pi_1/\partial a_2}{\partial \pi_2/\partial a_2}\right) \geq 0$. If U_2 is (weakly) locally normal in π_1 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 (weakly) increasing in a_1 at $\widehat{a}_1$. Hence if U_2 is (weakly) normal in π_1, then $a_2(a_1)$ is (weakly) increasing in a_1 at all $\widehat{a}_1$.*

The assumption that $\frac{\partial}{\partial a_1}\left(\frac{\partial \pi_1/\partial a_2}{\partial \pi_2/\partial a_2}\right) \geq 0$ states that, holding SM's action constant, an increase

in FM's action makes FM's material payoff weakly cheaper relative to SM's material payoff. It is satisfied when the actions enter the material payoff functions as complements in the sense that FM's material payoff function is weakly supermodular in the "goods" $(-a_1, a_2)$ and SM's material payoff function is weakly supermodular in the "goods" $(a_1, -a_2)$—as in Examples 1 and 2.[13] Whenever $\frac{\partial}{\partial a_1}\left(\frac{\partial \pi_1/\partial a_2}{\partial \pi_2/\partial a_2}\right) \geq 0$, normality of U_2 in π_1 ensures that $a_2(a_1)$ is increasing in a_1. To understand why, consider an increase in FM's action. If the price of FM's material payoff remained the same as before, normality of U_2 in π_1 would imply that SM prefers to increase his action. To the extent that FM's material payoff becomes cheaper, the incentive for SM to increase his action is reinforced.

A reciprocity motive—roughly speaking, a preference to behave kindly toward individuals who behave kindly—is built in to some fairness models (e.g., Rabin 1993; Cox, Friedman, & Sadiraj 2008), but *not* the ones listed at the beginning of Section 3. Reciprocity cannot be fully captured in models where utility depends only on the players' material payoffs, as I assume here. An influential defense of using models without a built-in reciprocity motive is that they are simpler to analyze, while nonetheless generating similar behavior (e.g., Fehr & Schmidt 2003). Proposition 1 shows that, for the usual payoff specifications of the bilateral exchange game, normality of U is the critical implicit assumption that causes SM to behave in a way that looks like reciprocity in commonly-used functional forms for social preferences defined only over material payoffs.

4.3 FM's Behavior

FM chooses her action a_1 anticipating the reaction $a_2(a_1)$ of SM. Lemma 2 summarizes key observations about FM's behavior.

Lemma 2. *Suppose U_2 is joint-monotonic and quasi-concave. Then:*

1. *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 $(\overline{\overline{a}}_1, \overline{\overline{a}}_2)$, and it is UPE.*

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

The first part of Lemma 2 is the surprising statement that SM's favorite transaction is the *only* MPE transaction that is possible for FM to induce. To understand why, note that there is exactly

[13] The assumption that $\frac{\partial}{\partial a_1}\left(\frac{\partial \pi_1/\partial a_2}{\partial \pi_2/\partial a_2}\right) \geq 0$ is also satisfied by the material payoff functions typically used in gift-exchange experiments, mentioned in footnote 3, which are weakly supermodular on the restricted domain of the actions.

one budget curve (corresponding to a particular value of a_1) passing through each MPE material payoff pair. At that MPE material payoff pair, the budget curve is tangent to the MPE frontier. Therefore, if and only if a MPE material payoff pair is maximal with respect to the budget curve—i.e., an optimal choice for SM—it is also maximal with respect to the MPE frontier—i.e., a favorite transaction for SM. Since SM's favorite transaction exists, is unique, and is UPE, the result follows. Because of this result, I will sometimes refer to SM's favorite transaction as "the" efficient transaction, even though technically there are many other MPE transactions.

This first part of the lemma has an immediate corollary: SM's favorite transaction is the only candidate for an *equilibrium* that is MPE. In turn, that observation has two important ramifications. First, because it is his favorite transaction, SM never prefers to deviate. Therefore, a necessary condition for an equilibrium to be MPE is $U_1\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right) \geq 0$: *FM* does not prefer to deviate from SM's favorite transaction. For this reason, subsequent sections will focus on whether FM has an incentive to take the action that induces the efficient transaction.

Second, even when the equilibrium of the bilateral exchange game is efficient, FM would *always* be better off with a contract determined by Coase/Nash bargaining, as long as writing and enforcing a contract is not too costly. The Nash bargaining solution 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.

Part 2 addresses existence of equilibrium. This equilibrium will involve FM choosing her outside option if SM's optimal response to any possible a_1 resulted in lower utility for FM than her outside payoff. 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 Part 1) there exists an action for FM that induces SM's favorite transaction.

Hence $U_1\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right) \geq 0$ is both a sufficient condition for trade to occur and a necessary condition for the equilibrium of the game to be MPE. Intuitively, this condition means that holding constant FM's social preferences, SM is neither too selfish nor too altruistic. If SM were too selfish, then $\overline{\overline{\pi}}_1$ would be so small that FM would prefer her outside option to $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$. If SM were too altruistic, then $\overline{\overline{\pi}}_2$ would be so small that FM would prefer her outside option to $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$.

5 Necessary Conditions for the Equilibrium to be Efficient

This section gives necessary conditions for the equilibrium of the bilateral exchange game to be efficient.

Theorem 2. *Suppose U_1 and U_2 are joint-monotonic, quasi-concave, and potentially fairness-kinked. If the equilibrium $(a_1, a_2(a_1))$ is MPE, then $(a_1, a_2(a_1))$ is SM's favorite transaction, and $U_1(\overrightarrow{\pi}(a_1, a_2(a_1))) \geq 0$. Furthermore, at least one of the following must be true:*

1. *FM's favorite transaction is the same as SM's favorite transaction, i.e., $(\overline{a}_1, \overline{a}_2) = (\overline{\overline{a}}_1, \overline{\overline{a}}_2)$.*
2. $\frac{dp(a_1, a_2(a_1))}{da_1} = 0$.
3. *U_2 is fairness-kinked at $\overrightarrow{\pi}(a_1, a_2(a_1))$.*

Two of the necessary conditions for the equilibrium to be MPE were discussed previously, in the context of Lemma 2, but are stated formally here: if the equilibrium transaction is MPE, then it must be SM's favorite transaction $(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2)$, and it must be true that $U_1(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2) \geq 0$. The third necessary condition is that at least one of three things must be true. 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 MPE (as well as UPE). Possibility (2) is that FM's action does not affect the slope of the budget curve at the equilibrium transaction, and possibility (3) is that SM's indifference curve is kinked at the equilibrium transaction.

To understand the economic intuition for possibilities (2) and (3), there are four steps. The first step is to notice that when FM's action deviates from the equilibrium $\overline{\overline{a}}_1$ to some other level $\overline{\overline{a}}_1 + \Delta$, the resulting change in the material payoffs can be decomposed into a "substitution effect" and an "income effect." The original material payoff pair $(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2)$ is SM's most-preferred point on the original budget curve $B(\overline{\overline{a}}_1)$, and some new material payoff pair (π'_1, π'_2) is SM's most-preferred point on the new budget curve $B(\overline{\overline{a}}_1 + \Delta)$. SM's original optimization problem can be reconceptualized as choosing the most-preferred point on the budget *line* that first-order approximates the original budget curve because SM's optimal choice, $(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2)$, is the same in both optimization problems. Similarly, (π'_1, π'_2) is not only SM's most-preferred point on the budget curve $B(\overline{\overline{a}}_1 + \Delta)$, but also SM's most-preferred point on the budget line that approximates it. Figure 4 draws the movement from $(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2)$ to (π'_1, π'_2) but shows only the budget lines rather than the budget curves. Since SM is now formally analogous to a consumer who is deciding how

much of the two "goods" $\widetilde{\pi}_1(p; I)$ and $\widetilde{\pi}_2(p; I)$ to "consume" as a function of price and income, his response to a change in the budget line can be characterized by the Slutsky decomposition into an income effect and a substitution effect.

The second step is to recognize that at an MPE equilibrium, the income effect is second order. Notice that the budget line at the MPE material payoff pair $\overrightarrow{\pi}\left(\overline{\overline{a}}_1, a_2\left(\overline{\overline{a}}_1\right)\right)$ is tangent to the MPE frontier. Therefore, for the relative price $p\left(\overline{\overline{a}}_1, a_2\left(\overline{\overline{a}}_1\right)\right)$, the material payoff pair $\overrightarrow{\pi}\left(\overline{\overline{a}}_1, a_2\left(\overline{\overline{a}}_1\right)\right)$ maximizes income. By an envelope argument, the income effect is second order.

The third step is to realize that at an MPE equilibrium, the substitution effect must equal zero. If the substitution effect were not zero, then by marginally deviating from the equilibrium action $\overline{\overline{a}}_1$ to some other action $\overline{\overline{a}}_1 + \Delta$, FM could cause SM to choose a material payoff pair that is either—depending on whether FM chooses $\Delta > 0$ or $\Delta < 0$—slightly northwest or slightly southeast of $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$. Moreover, since the income effect is second order, the northwest and southeast material payoff pairs that FM could induce by deviating would, to a first-order approximation, also be points on the MPE frontier. Since FM's favorite transaction does not coincide with SM's favorite transaction, FM would prefer over $\left(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2\right)$ whichever one of these is closer to her own favorite transaction. But then $\overline{\overline{a}}_1$ would not be optimal for FM. Hence if $\overline{\overline{a}}_1$ is in fact FM's optimal action, then the substitution effect must equal zero.

The fourth and final step is to notice that possibilities (2) and (3) correspond to the two possible ways that the substitution effect can equal zero. The budget lines may be parallel shifts, in which case there is no change in relative price; that is (2). Alternatively, optimal consumption may occur at a kink in the consumer's indifference curves, in which case the consumer's optimal consumption bundle does not change in response to a Slutsky-compensated change in price; that is (3). In both possibilities, since the substitution effect is zero, the income effect dominates despite being second order.[14]

The intuition for Theorem 2 in terms of a substitution effect and an income effect explains why

[14] The intuition for Theorem 2 in terms of income and substitution effects is a contribution relative to existing work about the closely-related rotten kid game. Bergstrom (1989) and Dijkstra (2007) identified the $\frac{dp(a_1, a_2(a_1))}{da_1} = 0$ condition as crucial for the rotten kid theorem to be true and provided a partial intuition. Both authors point out that when the condition fails, FM can influence the slope of the budget curve faced by SM by marginally adjusting a_1. Bergstrom (1989, p.1145) writes that FM can "'twist' the [budget curve] in a way that is favorable to her...[making it] the 'cheaper' to supply π_1 and the more 'expensive' to supply π_2. Consequently...the more π_1 [SM] will buy." Dijkstra (2007, pp.99-100) elaborates on this argument by presenting a diagram illustrating how FM might choose an action leading to a non-MPE transaction when $\frac{dp(a_1, a_2(a_1))}{da_1} \neq 0$. In terms of the discussion above, this argument describes the substitution effect but misses the income effect and the significance of the original consumption bundle being MPE, which ensures that the income effect is second order. The intuition about the income effect also makes clear that when FM is purely self-regarding, normality of SM's social preferences is a necessary condition for an MPE equilibrium, as discussed next.

normality of SM's social preferences will play an important role in possibility (2). For FM's actions in a neighborhood of $\overline{\overline{a}}_1$, there is no substitution effect, only an income effect. Therefore, if FM is purely self-regarding, local normality of U_2 in π_1 is another necessary condition for $\left(\overline{\overline{a}}_1, a_2\left(\overline{\overline{a}}_1\right)\right)$ to be an equilibrium.

The remaining question about Theorem 2 is for what material payoff functions possibility (2) in fact holds. Dijkstra (2007, his Lemma 1) answered this question: $\frac{dp(a_1,a_2(a_1))}{da_1} = 0$ at SM's favorite transaction $\left(\overline{\overline{a}}_1, a_2\left(\overline{\overline{a}}_1\right)\right)$ if and only if the material payoff functions are locally conditionally transferable at $\left(\overline{\overline{a}}_1, a_2\left(\overline{\overline{a}}_1\right)\right)$. Furthermore, Dijkstra (2007) showed that the material payoff functions are locally conditionally transferable at an MPE transaction $(\widehat{a}_1, \widehat{a}_2)$ if and only if, in a neighborhood of $(\widehat{a}_1, \widehat{a}_2)$, the material payoff functions can be approximated as globally conditionally transferable or (a_1, π_1)-additively-separable.[15]

The discussion has been about conditions for the equilibrium to be MPE, which will also be UPE (by Lemma 2). Theorem 6 in the Online Appendix characterizes the necessary conditions for the equilibrium to be UPE. There is one potentially empirically-relevant case where an equilibrium may be UPE but not MPE: both players' social preferences are fairness-kinked at the equilibrium, and the two players have different fairness rules at the equilibrium. This situation might occur if the two agents have different, self-serving ideas about what is fair. An implication of Theorem 6 is that if one or both players' interpersonal indifference curves are smooth, then the equilibrium is UPE if and only if it is MPE, in which case Theorem 2 is a complete characterization of the necessary conditions for efficiency.

6 Sufficient Conditions for the Equilibrium to be Efficient

The previous section showed that there are two interesting cases in which the equilibrium could be MPE: (1) the budget lines that approximate the budget curves are parallel shifts, and (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 both MPE and UPE.

The intuition for 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

[15] In an influential paper about the closely-related rotten kid theorem, Bergstrom (1989) had argued but did not prove that global conditional transferability is necessary for possibility (2). Dijkstra's (2007) result shows that possibility (2) actually holds under a wider class of material payoff functions that includes (a_1, π_1)-additively-separable material payoff functions, as well as any other material payoff functions that happen to be locally conditionally transferable at $\left(\overline{\overline{a}}_1, a_2\left(\overline{\overline{a}}_1\right)\right)$.

FM varies her action. FM will choose the action that maximizes both players' material payoffs if FM is purely self-regarding or altruistic, leading to an MPE equilibrium. Since this transaction is SM's favorite transaction, it will also be UPE.

6.1 Efficient Case I: Budget Curves Are Parallel Shifts

Recall from Section 5 that when FM varies her action, there is just an income effect with no substitution effect. If FM is purely self-regarding or altruistic, the (global) normality of U_2 is sufficient to ensure that the equilibrium is unique, MPE, and UPE.[16]

Theorem 3. *Suppose U_1 is quasi-concave, and 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 $U_1\left(\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\right) \geq 0$ at SM's favorite transaction $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$, then the unique equilibrium transaction is $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$, which is MPE and UPE.*

Since the material payoff functions are globally conditionally transferable, FM's action affects the level of "income" but not the "price" faced by SM. Because U_2 is normal, SM will choose his action such that both players' material payoffs are increasing in the total surplus to be divided. Hence, FM maximizes both players' material payoffs by choosing the unique action that leads to an MPE outcome. Figure 5a illustrates Theorem 3.

Theorem 3 assumes that the material payoff functions are globally conditionally transferable, even though Proposition 1 showed that (a_1, π_1)-additive-separability is also consistent with $\frac{dp(a_1, a_2(a_1))}{da_1} = 0$ at $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. The reason is that global conditional transferability implies that $\frac{dp(a_1, a_2(a_1))}{da_1} = 0$ at *all* a_1, while (a_1, π_1)-additive-separability only implies it in a neighborhood of $\overline{\overline{a}}_1$. Therefore, global conditional transferability but not (a_1, π_1)-additive-separability ensures that SM's behavior will align the players' material incentives at all a_1, in turn ensuring that $\overline{\overline{a}}_1$ is FM's global optimum, not just a local optimum.[17]

Theorem 3 could be paraphrased: Fixing the material payoff functions and SM's social preferences satisfying the assumptions of the theorem, then—depending on FM's social preferences—either the equilibrium will be no trade, or the equilibrium transaction will be MPE and UPE. To understand why, notice that if $U_1\left(\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\right) \ngeq 0$, then FM prefers her outside option to the

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

[17] Dijkstra's (2007) Proposition 2 gives the impression that the rotten kid theorem holds if the material payoff functions are globally conditionally transferable *or* (a_1, π_1)-additive-separable. As written, the proposition is true, but it assumes that "all agents' second order conditions are satisfied" without giving sufficient conditions for FM's local optimum to be a global optimum.

action $\overline{\overline{a}}_1$ that maximizes her utility conditional on trading. In that case, the equilibrium will be no trade. On the other hand, if $U_1\left(\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\right) \geq 0$ so that the equilibrium involves trade, then it will occur at $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$.

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 action such that the budget curve coincides with the MPE frontier. For that reason, loosely speaking (since there is no uncertainty in the model), FM would choose the efficient action even if she were uncertain about SM's social preferences and hence uncertain about the action SM will choose.[18]

Theorem 3 holds regardless of whether SM's social preferences are monotonic or not, as long as they are joint-monotonic. However, if FM's social preferences are joint-monotonic but not monotonic, then the conclusion of Theorem 3 may not hold. SM's behavior aligns the material incentives of the two players, but if FM has non-monotonic social preferences, then she may prefer not to maximize the players' material payoffs. Figure 5b illustrates an equilibrium that is neither MPE not UPE.

6.2 Efficient Case II: SM's Social Preferences Are Fairness-Kinked

If SM's best response $a_2\left(\widehat{a}_1\right)$ to some $\widehat{a}_1$ occurs on the fairness rule at a point where SM's social preferences are strictly kinked, then SM behaves in accordance with the fairness rule in a neighborhood of $\widehat{a}_1$. That is, SM obeys his fairness rule locally. This observation, combined with the fact that the fairness rule is upward-sloping, means that SM's social preferences are locally normal when his optimum occurs at a kink point. It follows that if SM's indifference curve is (even very slightly kinked) at his favorite transaction, then the action that induces SM's favorite transaction will be a *local* optimum for FM, if FM is purely self-regarding or altruistic. The central intuition for how fairness-kinked social preferences can lead to an MPE equilibrium comes from this logic: The fairness rule is characterized by the players' material payoffs increasing or decreasing in tandem. As long as SM's favorite transaction occurs at a kink, SM responds to small changes in FM's action by adjusting his own action to ensure that the fairness rule remains satisfied. This behavior means that both players' material payoffs are increasing in the size of the pie, which gives FM an incentive to maximize the size of the pie, as long as FM is purely self-regarding or altruistic.

[18] Becker (1974) made this same observation in the context of the rotten kid theorem.

Theorem 4 provides sufficient conditions for SM's favorite transaction to be the unique equilibrium: FM is purely self-regarding or altruistic; the material payoff functions $\pi_1(a_1, a_2)$ and $\pi_2(a_1, a_2)$ are (a_1, a_2)-additively-separable;[19] FM's favorite transaction gives higher material payoff to FM than SM's favorite transaction does;[20] and SM's interpersonal preferences are normal and *sufficiently* kinked at his favorite transaction.

Theorem 4. *Suppose U_1 is quasi-concave, and suppose U_2 is fairness-kinked, with U_2^A and U_2^B being joint-monotonic, quasi-concave, normal, and continuously twice-differentiable. Suppose U_1 is either purely self-regarding, or strictly monotonic with FM's favorite transaction $(\overline{a}_1, \overline{a}_2)$ giving higher material payoff to FM than SM's favorite transaction $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. Suppose the players' material payoff function are (a_1, a_2)-additively-separable. Let $(\widehat{a}_1, \widehat{a}_2)$ denote the (necessarily unique) transaction with $\widehat{a}_1 < \overline{\overline{a}}_1$ such that $\pi_1(\widehat{a}_1, \widehat{a}_2) = \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ and $U_2(\widehat{a}_1, \widehat{a}_2) = 0$. If $U_1\left(\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\right) \geq 0$, and if, at $\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$,*

$$U_2^A = U_2^B \tag{5}$$

$$\frac{\partial U_2^A}{\partial \pi_2} - p(\overline{\overline{a}}_1, \widehat{a}_2)\frac{\partial U_2^A}{\partial \pi_1} > 0 \tag{6}$$

$$\frac{\partial U_2^B}{\partial \pi_2} - p\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\frac{\partial U_2^B}{\partial \pi_1} < 0, \tag{7}$$

then the unique equilibrium transaction is $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$, which is MPE and UPE.

Equality (5) says that the SM's favorite transaction is a material payoff pair on the fairness rule. Inequality (7) combined with $\frac{\partial U_2^A}{\partial \pi_1} - p\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\frac{\partial U_2^A}{\partial \pi_2} > 0$ imply that SM's favorite transaction in fact occurs at a strict kink in SM's indifference curves. Inequality (6) is more stringent than $\frac{\partial U_2^A}{\partial \pi_1} - p\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\frac{\partial U_2^A}{\partial \pi_2} > 0$, ruling out that SM is not too altruistic at a particular, disadvantageously unfair transaction, $(\widehat{a}_1, \widehat{a}_2)$. If SM puts negative weight on FM's material payoff at disadvantageously unfair transactions, such as with "inequity averse" social preferences (1), then (6) is automatically satisfied.

Normality of U_2 plays a different role in Theorem 4 than in Theorem 3. In Theorem 3, it ensured that SM's best-response function $a_2(a_1)$ caused the players' material payoffs to increase or decrease

[19] While additive-separability is helpful in the proof of Theorem 4, it is clearly not a necessary condition for the equilibrium to occur at SM's favorite transaction. Unfortunately, I do not know if a less restrictive assumption will suffice.

[20] This assumption is more realistic than the opposite assumption. Together with the assumptions regarding $(\widehat{a}_1, \widehat{a}_2)$, its role is to rule out the possibility that some other action for FM gives FM higher material payoff than SM's favorite transaction does. With the opposite assumption instead, analogous alternative sufficient conditions could be formulated that rule out the possibility that some other action for FM gives FM *lower* material payoff than SM's favorite transaction does.

in tandem as a_1 varies. However, in Theorem 4, the fairness rule plays that role. In Theorem 4, normality in combination with (a_1, a_2)-additive-separability allows the aversion to disadvantageous unfairness local to $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ from (6) to imply a sufficient aversion to disadvantageous unfairness on the outside option indifference curve. Without normality, even if inducing SM's favorite transaction were a local optimum for FM, FM might earn a still greater material payoff by taking a much smaller action. Figure 6a illustrates the MPE/UPE equilibrium when SM has fairness-kinked social preferences, and Figure 6b shows a way the equilibrium could fail to be efficient if the assumptions of Theorem 4 are not satisfied.

Like Theorem 3, Theorem 4 suggests that, given the theorem's assumptions about material payoff functions and SM's social preferences, then—depending on FM's social preferences—either the equilibrium will be no trade, or the equilibrium transaction will be MPE and UPE. Put another way, these results roughly state that if SM is behaving according to a fairness rule, and if FM is purely self-regarding or altruistic, then any trade that occurs will be efficient. 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 a fairness rule that assigns larger material payoff to both players as the pie increases, any fairness rule can lead to an efficient equilibrium, even if it is non-linear or self-serving.

Also like Theorem 3, the conclusions of Theorem 4 do not depend on whether SM's social preferences are monotonic or joint-monotonic, but the conclusions may not hold if *FM*'s social preferences are joint-monotonic. In Figure 6a, if FM had joint-monotonic preferences, his most-preferred point on SM's fairness rule—and therefore the equilibrium—could occur at an inefficient material payoff pair.

In the informal discussion about Theorem 3, it was suggested that the theorem would hold even if FM were uncertain about exactly what SM's social preferences are. However, Theorem 4 requires that FM know what fairness rule SM is following. Otherwise, FM would not know which action would induce SM's favorite transaction. Loosely speaking, there is "social value" in having SM's fairness rule be common knowledge. Social norms like 50-50 sharing may have the consequence of providing common knowledge fairness rules.

7 Applications

While the previous sections developed results about the bilateral exchange game, this section discusses how those results may inform research on three topics: the rotten kid theorem, gift exchange in laboratory experiments, and gift exchange in field settings.

7.1 The Rotten Kid Theorem

Recall the rotten kid game from Example 2: A child (FM) chooses how much to work for pay and suffers convex cost of effort $n(a_1)$. Then the parent (SM) transfers some amount, a_2, of family income to the child. The material payoffs equal the respective private incomes divided by the respective prices of consumption: $\pi_1(a_1, a_2) = \frac{I_1+a_2-n(a_1)}{P_1}$ and $\pi_2(a_1, a_2) = \frac{I_2+a_1-a_2}{P_2}$. Family income is the sum of the child's income and the parent's income, $I_1 + I_2 + a_1 - n(a_1)$. An outcome of this game is MPE if and only if it maximizes family income.

The parent is assumed to be altruistic toward the child, with $U_2(\pi_1, \pi_2)$ monotonic and normal in both material payoffs, but the child may either be altruistic toward the parent or purely self-regarding (a "rotten kid"): $U_1(\pi_1, \pi_2) = \pi_1$. In Becker's (1974, p.1080) original description of the rotten kid theorem: "The major, and somewhat unexpected, conclusion is that if a [household] head exists, *other members are also motivated to maximize family income and consumption, even if their welfare depends on their own consumption alone.*"

Theorem 5 (Rotten Kid Theorem). *Suppose U_1 is quasi-concave, and either purely self-regarding or monotonic. Suppose U_2 is monotonic, quasi-concave, and normal. In the rotten kid game (Example 2), the unique equilibrium transaction is SM's favorite transaction $(\overline{\overline{a}}_1, \overline{\overline{a}}_2)$, which maximizes family income.*

In subsequent work, Bergstrom (1989) and Dijkstra (2007) proved that the material payoff functions in Example 2 could be generalized; Dijkstra's (2007) analysis suggested that global conditional transferability is in fact sufficient for the theorem to hold.

The theorem has been interpreted as applying to family environments but *not* market environments because researchers have emphasized the role of altruism, which is assumed to be relevant primarily within the family. One contribution of this paper is to point out that the rotten kid theorem logic applies in settings where the relevant social preference is a concern for fairness, which appears to be widespread in interactions between unrelated individuals (e.g., Kahneman, Knetsch,

& Thaler 1986; Bewley 1999). One way that Theorem 3 generalizes Theorem 5 is in allowing SM to have social preferences that are joint-monotonic but not monotonic. It follows fairly directly from Lemma 1 that this relaxation does not matter for equilibrium behavior. Although only a minor mathematical generalization, the relaxation of monotonicity is a major economic generalization because it means that the surprising efficiency conclusion may be applicable in a far wider range of settings.

Relative to existing work on the rotten kid theorem, the analysis in this paper provides three new insights about the result and what drives it. First, intuitions have been incomplete about why local (or global) conditional transferability of the material payoff functions is important. The analysis of substitution and income effects in Section 5 shows that having a substitution effect equal to zero is what is crucial. Second, this paper suggests that normality has been underemphasized in existing interpretations. Typical brief descriptions of the theorem focus on altruism and do not mention normality.[21] However, while monotonicity of SM's social preferences can be relaxed, normality of SM's social preferences is crucial. Finally, Theorem 3 shows that the equilibrium is UPE, in addition to MPE. Hence the efficiency result is even stronger than has been recognized.

7.2 Gift Exchange in the Field

In many bilateral exchange environments in the field, one player is an individual—a worker or customer—who may be concerned with distributional fairness, while the other player is a firm, who is plausibly profit-maximizing. If the purely self-regarding firm is modeled as FM, and the fairness-minded person as SM, then Theorems 3 and 4 may apply.

Theorem 3 is applicable when the material payoff functions satisfy local conditional transferability. This broad but restrictive class includes quasi-linearity in a_2 as a special case, which is typically considered a good model when *SM's action* is a transfer of money to FM. This is true when FM is a seller who provides a service, and SM is a (fair-minded) customer who decides how much to pay for the service. It is *not* true when 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, loosely speaking, the analysis in this paper suggests that an efficient outcome might be more likely in the former case than the latter.

[21] For example, Becker (1974, p.1076) summarizes the theorem: "If one member [of the family], call him the 'head,' cares sufficiently about all other members to transfer general resources to them, redistribution of income among members would not affect the consumption of any member, as long as the head continues to contribute to all." Bergstrom (1989, p.1139) writes that "[the rotten kid theorem] tells us that a sufficiently benevolent household head would *automatically* internalize all the external effects that family members have on each other."

Theorem 4 applies when SM has "sufficiently fairness-kinked" social preferences at his favorite transaction (whether or not the material payoff functions satisfy local conditional transferability). This may be a reasonable model when there is a salient fairness rule that applies to the specified transaction, e.g., an agreed-upon rate of exchange between amount of work that SM will provide and money FM will pay.

The results of the paper would appear to raise a puzzle: if social preferences alone will generate an efficient equilibrium in a variety of settings (as predicted by the theory), why do instances of exchange based solely on social preferences appear to be much less common in the field than exchange based on contracts (as judged by casual observation)? I believe that part of the answer is that, if contracting is possible, FM can generally do better at an efficient contract than at an efficient equilibrium of the bilateral exchange game, which necessarily occurs at the transaction on the Pareto frontier *most preferred by SM*. I suspect another part of the answer is that in many field setting, SM is too selfish. The theory implies the equilibrium can be efficient *if* the equilibrium is better for FM than not trading, but that condition will fail if SM's most-preferred transaction assigns too little material payoff to FM.

7.3 Gift Exchange in the Lab

Laboratory experiments have generated evidence of gift exchange and usually report moderate levels of efficiency (see Falk & Fehr, 2010, for a review).[22] The analysis in this paper suggests two reasons why the extent of efficiency measured from gift exchange in the lab should not be expected to generalize to field settings.

First, the fairness rule that is usually salient in the lab—a 50-50 split of monetary payoffs—is often not the relevant fairness rule in field settings. In the lab, where the monetary payoffs functions are common knowledge, the fairness rule that equates monetary payoffs is salient and easy to implement. In a field setting such as manager-worker gift exchange, since disutility of effort is not easily measured in monetary units, this fairness rule would likely not be relevant to a worker. Some other fairness rule—such as a usual rate of exchange between salary and output—might instead be salient. Since individuals may feel more compelled to follow one of these fairness rules

[22] Note that when efficiency is measured in the lab, it is material Pareto efficiency, not utility Pareto efficiency. In the lab, *monetary* payoffs are specified as a function of the players' actions, $m_1(a_1, a_2)$ and $m_2(a_1, a_2)$. "Efficiency" is operationalized as what might be called "money Pareto efficiency," defined by the condition $\frac{\partial m_1(a_1,a_2)/\partial a_1}{\partial m_1(a_1,a_2)/\partial a_2} = \frac{\partial m_2(a_1,a_2)/\partial a_1}{\partial m_2(a_1,a_2)/\partial a_2}$. Under the plausible assumption that players' material payoffs are increasing functions of the amount of money earned—i.e., $\pi_1(a_1, a_2) = q_1(m_1(a_1, a_2))$ and $\pi_2(a_1, a_2) = q_2(m_2(a_1, a_2))$, with $q_1', q_2' > 0$—a transaction (a_1, a_2) is money Pareto efficient if and only if it is MPE.

than the other, it is problematic to generalize the amount of efficiency measured in the lab due to the equal-split fairness rule to the amount of efficiency expected in the field due to a different fairness rule.

Second, as mentioned above, in many field settings, it is plausible that only one of the players cares about fairness, while in the lab, both players are drawn from a common pool of fair-minded experimental participants. Theorems 3 and 4 apply when FM has purely self-regarding or monotonic social preferences, but not necessarily when FM has joint-monotonic preferences that are not monotonic. Some FMs in the lab may prefer a higher relative material payoff even if it means a lower absolute payoff. To the extent that they can achieve this by choosing an inefficiently low action, the amount of efficiency measured in the lab will *understate* what can be expected in analogous field settings where FM would be purely self-regarding.

8 Discussion

The analysis in this paper could be extended in several directions. First, while this paper has focused on bilateral exchange, there are of course important multi-player field settings. How conclusions from the bilateral exchange environment generalize will depend on the nature of the multilateral social preferences. For example, suppose a profit-maximizing firm simultaneously offers a wage to each of several workers, and then each worker sequentially chooses his effort level. If each worker cares about the fairness of his own bilateral transaction with the firm, then the analysis in this paper applies immediately to each bilateral transaction. Consistent with that possibility, Maximiano, Sloof, & Sonnemans (2007) find in a laboratory labor market that behavior in the presence of other workers is nearly identical to behavior when there is only a single worker. However, if each worker has social preferences over all players' material payoffs, then the analysis becomes more complex.

Second, the analysis could be extended to allow each player to be uncertain about the other player's material payoff function and social preferences (see, e.g., Fehr, Klein, & Schmidt 2007). In some cases, uncertainty about social preferences will make little difference—such as when a purely self-regarding FM believes that a sufficiently large proportion of second-mover types will adhere to a 50-50 sharing rule. In that case, it is still optimal for FM to take the action that induces a 50-50-rule-following SM to respond efficiently. However, in many cases, it seems likely that uncertainty will reduce efficiency. For example, even if all second-mover types behave in accordance with a fairness rule, if each adheres to a *different* fairness rule, then FM's uncertainty

will make it optimal for him to play against a "representative SM" who has smooth preferences. As a result, the equilibrium may be materially Pareto inefficient.

Third, while I have assumed that other-regarding behavior is driven exclusively by preferences over material payoffs, the analysis could be extended to incorporate more complex mechanisms that are known to influence social behavior. One such mechanism is signaling: A purely self-regarding agent may appear to be other-regarding because he is signaling to others (or to himself) that he is a fair-minded type (e.g., Andreoni & Bernheim 2009). I conjecture that the equilibrium of the bilateral exchange game will be very similar if the players' other-regarding behavior is driven by signaling instead of preferences over the distribution of material payoffs. If the players' behavior can be represented *as if* it were generated by distributional social preferences, then the predictions about equilibrium behavior are the same whether the behavior is in fact the result of distributional preferences or signaling.[23]

Finally, another mechanism that can generate other-regarding behavior is intentions-based reciprocity: a player's distributional preferences may depend on his beliefs about the other player's distributional preferences. In particular, he may prefer to treat more kindly someone whom he thinks would treat him kindly (Rabin 1993). In work that is closely related to the present paper, Netzer & Schmutzler (2009) study a bilateral exchange game where 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. With intentions-based reciprocity, it is not clear how to define "utility Pareto efficiency" because preferences depend on endogenous beliefs, so Netzer & Schmutzler focus on material Pareto efficiency. They argue that when FM is purely self-regarding and SM's behavior is driven by intentions-based reciprocity, the equilibrium is generically materially Pareto inefficient because SM is unwilling to reciprocate high actions by FM, which are interpreted as attempts at material-payoff maximization instead of as kind motivation. However, Charness & Rabin (2002) argue that intentions-based reciprocity becomes important only in response to a first-mover's unkind behavior, while the distributional social preferences I study here drive a second-mover's behavior when FM has behaved kindly. If so, the analysis in this paper is more relevant than the implications of intentions-based reciprocity for gift-exchange settings where both parties are gaining from the transaction. For the more general case where the players' behavior is influenced both by intentions-based reciprocity and distributional social preferences, the equilibrium is not yet known.

[23] A difference is that if the players' underlying preferences are *actually* purely self-regarding, then the set of UPE transactions is identical with the set of MPE transactions. However, if the equilibrium occurs at SM's favorite transaction, then it will still be MPE (and therefore UPE).

References

[1] George A. Akerlof. Labor contracts as partial gift exchange. *Quarterly Journal of Economics*, 97(4):543–569, November 1982.

[2] George A. Akerlof and Janet L. Yellen. The fair wage-effort hypothesis and unemployment. *Quarterly Journal of Economics*, 105(2):255–283, May 1990.

[3] 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.

[4] 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.

[5] Linda Babcock and George Loewenstein. Explaining bargaining impasse: The role of self-serving biases. *Journal of Economic Perspectives*, 11(1):109–126, 1997.

[6] 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.

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

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

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

[10] Truman F. Bewley. *Why Wages Don't Fall During a Recession.* Harvard University Press, Cambridge, MA, 1999.

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

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

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

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

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

[16] Robyn M. Dawes and Richard H. Thaler. Cooperation. *Journal of Economic Perspectives*, 2(3):187–197, 1988.

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

[18] Martin Dufwenberg, Paul Heidhues, Georg Kirchsteiger, Frank Riedel, and Joel Sobel. Other-regarding preferences in general equilibrium. September 2008. University of California San Diego Working Paper.

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

[20] Armin Falk and Ernst Fehr. Reciprocity in experimental markets. In Charles Plott and Vernon Smith, editors, *Handbook of Experimental Economics Results: Volume 1*, chapter 38, pages 325–334. North-Holland, 2010.

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

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

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

[24] 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.

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

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

[27] Sanford J. Grossman and Oliver D. Hart. An analysis of the principal-agent problem. *Econometrica*, 51(1):7–46, January 1983.

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

[31] Sandra Maximiano, Randolph Sloof, and Joep Sonnemans. Gift exchange in a multi-worker firm. *Economic Journal*, 117:1025–1050, 2007.

[32] Nick Netzer and Armin Schmutzler. Rotten kids with bad intentions. December 2009. University of Zurich Socioeconomic Institute Working Paper No. 0919.

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

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

[35] Matthew Rabin. Bargaining structure, fairness, and efficiency. *Berkeley Department of Economics Working Paper*, February 1997.

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

Online Appendix: Proofs

(Not for publication)

We begin with a technical lemma before proving the results in the text.

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 $U_1(\overrightarrow{\pi}(0, 0)) = U_2(\overrightarrow{\pi}(0, 0)) = 0$ (by A4), 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 A1 and A6 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 $g(\pi_2'') < \hat{x}$ and $\bar{y} < \hat{y} < g(\pi_2')$ 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 (x', y') 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 $x = x', y = y'$, as well as by the curve C since $(x', y') \in S$), so it is compact. Hence we can choose a point $(x, y) \in S$ 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 and the MPE frontier have graphs of the form $(g(\pi_2), \pi_2)$, where g is a continuous, decreasing, weakly concave function. We start with the budget curve. Fix action a_1. Since $B(a_1) \equiv \{\overrightarrow{\pi}(a_1, a_2)\}_{a_2 \in \mathbb{R}}$, $\left.\frac{d\pi_1}{d\pi_2}\right|_{B(a_1)} = \frac{\partial \pi_1/\partial a_2}{\partial \pi_2/\partial a_2}$. We can represent the graph of the budget curve as $(g(\pi_2), \pi_2)$, where $\frac{dg}{d\pi_2} = \left.\frac{d\pi_1}{d\pi_2}\right|_{B(a_1)} < 0$, hence g is continuous and decreasing. Now, we can parameterize the graph of the budget curve as $(g(\pi_2(a_2)), \pi_2(a_2))$. Note that $\pi_2(a_2'') < \pi_2(a_2')$ if $a_2'' > a_2'$ (from A2). Weak quasi-concavity of the material payoff functions (which follows from A3) implies that $\frac{\partial}{\partial a_2}\left(\frac{\partial \pi_1/\partial a_2}{\partial \pi_2/\partial a_2}\right) = \frac{\partial}{\partial a_2}\left(\frac{dg}{d\pi_2}\right) > 0$. Hence g is strictly concave. A1-A3 imply that the MPE frontier has a graph of the form $(g(\pi_2), \pi_2)$, where g is a continuous, decreasing, weakly concave function (a standard result about the "utility possibility frontier" when utility is purely self-regarding).

□

Theorem 1. *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 $(\overline{\overline{a}}_1, \overline{\overline{a}}_2)$, exist and are unique. The set of UPE material payoff pairs is a connected set that includes $(\overline{\pi}_1, \overline{\pi}_2)$ and $(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2)$ and lies within the region enclosed by $\overline{\overline{IC}}_1$, $\overline{IC}_2$, and the MPE frontier. In addition, if U_1 or U_2 (or both) is monotonic, then 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. Since the space of individually-rational payoffs is compact (by Technical Lemma), there must be some point that maximizes U_2. Joint-monotonicity implies that a maximizing material payoff pair must lie on the MPE frontier, and Technical Lemma implies that there must be a *unique* material payoff pair that achieves the maximum utility. 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: $-\frac{\partial \pi_1/\partial a_1}{\partial \pi_1/\partial a_2} = \frac{da_2}{da_1}\Big|_{\pi_1=\overline{\overline{\pi}}_1} = \frac{da_2}{da_1}\Big|_{\pi_2=\overline{\overline{\pi}}_2} = -\frac{\partial \pi_2/\partial a_1}{\partial \pi_2/\partial a_2}$. Since the material indifference curves are strictly convex in $(-a_1, a_2)$ and $(a_1, -a_2)$, respectively (by A3), $\frac{d^2 a_2}{d(a_1)^2}\Big|_{\pi_1=\overline{\overline{\pi}}_1} > 0$ and $\frac{d^2 a_2}{d(a_1)^2}\Big|_{\pi_2=\overline{\overline{\pi}}_2} < 0$, 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. 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.

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.

For the additional result, suppose WLOG that U_1 is monotonic. 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 $(\pi_1'', \pi_2''), (\pi_1', \pi_2'), (\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2)$ lie in that order along the MPE frontier, and $U_2(\pi_1'', \pi_2'') \geq U_2(\pi_1', \pi_2') < U_2(\overline{\overline{\pi}}_1, \overline{\overline{\pi}}_2)$; 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 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 $(\widehat{a}_1, a_2(\widehat{a}_1))$.*

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, $\frac{\partial U_2}{\partial a_2}(\overrightarrow{\pi}(\widehat{a}_1, a_2)) = \frac{\partial U_2}{\partial \pi_2} - p(\widehat{a}_1, a_2)\frac{\partial U_2}{\partial \pi_1} = 0$. 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 (TA) rules out that they both equal 0. Therefore, the first-order condition implies that both are positive.

□

Proposition 1. *Suppose U_2 is joint-monotonic and quasi-concave. Suppose $\frac{\partial}{\partial a_1}\left(\frac{\partial \pi_1/\partial a_2}{\partial \pi_2/\partial a_2}\right) \geq 0$. If U_2 is (weakly) locally normal in π_1 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 (weakly) increasing in a_1 at $\widehat{a}_1$. Hence if U_2 is (weakly) normal in π_1, then $a_2(a_1)$ is (weakly) increasing in a_1 at all $\widehat{a}_1$.*

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 $(\pi_2 - \pi_1)$-plane, A is northwest of B. Now draw the lines with slopes $-p(\widehat{a}_1, a_2(\widehat{a}_1))$ and $-p(\widehat{a}_1', a_2(\widehat{a}_1')) > -p(\widehat{a}_1, a_2(\widehat{a}_1))$ going through A and B respectively; this inequality is implied by A3 and $\frac{\partial}{\partial a_1}\left(\frac{\partial \pi_1/\partial a_2}{\partial \pi_2/\partial a_2}\right) \leq 0$. The lines will intersect at some generic point, say C. There are two cases: C is either strictly southeast of both A and B, or C is strictly southeast of A and northwest of B. The argument is similar in

both cases, so suppose the latter. The change from A to B can be decomposed into a substitution effect and an income effect (where C is the "endowment" consumption bundle). The substitution effect causes a move from A to a point A' weakly northwest of A. Because of (weak) normality, the income effect then makes us move from A' to B, where B needs to be (weakly) northeast of A'—and therefore (weakly) north of A.But B actually lies (strictly) south of A, a contradiction.

□

Lemma 2. *Suppose U_2 is joint-monotonic and quasi-concave. Then:*

1. *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 $(\overline{\overline{a}}_1, \overline{\overline{a}}_2)$, and it is UPE.*

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

Proof of part 1: 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 $(\overline{\overline{a}}_1, \overline{\overline{a}}_2)$. 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 $\overrightarrow{\pi}(\overline{\overline{a}}_1, \overline{\overline{a}}_2)$, 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 $\overrightarrow{\pi}(\overline{\overline{a}}_1, \overline{\overline{a}}_2)$ along the MPE frontier (as stated in Technical Lemma). Theorem 2 states that SM's favorite transaction is UPE.

Proof of part 2: From part 1, if FM chooses action $\overline{\overline{a}}_1$, SM will choose action $\overline{\overline{a}}_2$. Joint-monotonicity and A4-A5 imply that $U_2(\overline{\overline{a}}_1, \overline{\overline{a}}_2) > 0$. Since some action other than $\overline{\overline{a}}_1$ may give FM an even higher utility than $U_1(\overrightarrow{\pi}(\overline{\overline{a}}_1, \overline{\overline{a}}_2)) \geq 0$, this 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.

□

Theorem 2. *Suppose U_1 and U_2 are joint-monotonic, quasi-concave, and potentially fairness-kinked. If the equilibrium $(a_1, a_2(a_1))$ is MPE, then $(a_1, a_2(a_1))$ is SM's favorite transaction, and $U_1(\overrightarrow{\pi}(a_1, a_2(a_1))) \geq 0$. Furthermore, at least one of the following must be true:*

1. *FM's favorite transaction is the same as SM's favorite transaction, i.e., $(\overline{a}_1, \overline{a}_2) = (\overline{\overline{a}}_1, \overline{\overline{a}}_2)$.*
2. $\frac{dp(a_1, a_2(a_1))}{da_1} = 0$.
3. *U_2 is fairness-kinked at $\overrightarrow{\pi}(a_1, a_2(a_1))$.*

Proof: It follows directly from Lemma 2 (part 1) that if the equilibrium $(a_1, a_2(a_1))$ is MPE, then $(a_1, a_2(a_1))$ is SM's favorite transaction. Clearly, if $(a_1, a_2(a_1))$ is an equilibrium, then $U_1(\overrightarrow{\pi}(a_1, a_2(a_1))) \geq 0$; otherwise, FM would prefer her outside option.

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 3, 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 potentially 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 (as illustrated in Figure 4; 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: $p(a_1, a_2(a_1))\frac{\partial\pi_2^*(a_1,a_2(a_1))}{\partial a_2} + \frac{\partial\pi_1^*(a_1,a_2(a_1))}{\partial a_2} = 0$ 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 uses $p \equiv -\left.\frac{d\pi_1}{d\pi_2}\right|_{B(a_1)} = -\frac{\partial\pi_1(a_1,a_2)/\partial a_2}{\partial\pi_2(a_1,a_2)/\partial a_2}$ and the MPE condition, $\frac{\partial\pi_1^*(a_1,a_2)/\partial a_1}{\partial\pi_2^*(a_1,a_2)/\partial a_1} = \frac{\partial\pi_1^*(a_1,a_2)/\partial a_2}{\partial\pi_2^*(a_1,a_2)/\partial a_2}$. 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 $\frac{d\widetilde{\pi}_1(p,p\widetilde{\pi}_1+\widetilde{\pi}_2)}{dp} = \frac{p}{\left.d^2\pi_2/d(\pi_1)^2\right|_{U_2(\pi_1^*,\pi_2^*)}}$.

An interior equilibrium transaction satisfies FM's first-order condition, which can be written in terms of the budget lines: $\frac{d}{da_1}U_1\left(\widetilde{\pi}_1\left(p,\,I\right),\,\widetilde{\pi}_2\left(p,\,I\right)\right)=0$. (A6 combined with joint-monotonicity of U_2 ensures that SM's favorite transaction is indeed interior.) Using the income and substitution effects derived above,

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

Recall that $\frac{\partial U_1}{\partial \pi_1}>0$ and $\frac{\partial U_1}{\partial \pi_2}>0$ (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_1 is quasi-concave, and 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 $U_1\left(\overrightarrow{\pi}\left(\overline{\overline{a}}_1,\overline{\overline{a}}_2\right)\right)\geq 0$ at SM's favorite transaction $\left(\overline{\overline{a}}_1,\overline{\overline{a}}_2\right)$, then the unique equilibrium transaction is $\left(\overline{\overline{a}}_1,\overline{\overline{a}}_2\right)$, which is MPE and UPE.*

Proof: Since the material payoff functions are globally conditionally transferable, the budget curves are all parallel lines with slope $-p \equiv \left.\frac{d\pi_1}{d\pi_2}\right|_{B(a_1)} = \frac{\partial\pi_1(a_1,a_2)/\partial a_2}{\partial\pi_2(a_1,a_2)/\partial a_2} = -k$ 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, part 1). 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.

□

Lemma 3. *Suppose U_2 is potentially fairness-kinked, with U_2^A and U_2^B being joint-monotonic, quasi-concave, and continuously twice-differentiable. If $U_2 = \min\left\{U_2^A, U_2^B\right\}$ satisfies*

$$U_2^A = U_2^B$$

$$\frac{\partial U_2^A}{\partial \pi_2} - p\left(\widehat{a}_1, a_2\left(\widehat{a}_1\right)\right)\frac{\partial U_2^A}{\partial \pi_1} > 0$$

$$\frac{\partial U_2^B}{\partial \pi_2} - p\left(\widehat{a}_1, a_2\left(\widehat{a}_1\right)\right)\frac{\partial U_2^B}{\partial \pi_1} < 0$$

at $\overrightarrow{\pi}\left(\widehat{a}_1, a_2\left(\widehat{a}_1\right)\right)$, then SM's optimal strategy $a_2(a_1)$ satisfies the fairness rule for all a_1 in a neighborhood of $\widehat{a}_1$.

Proof: Since the given inequalities hold strictly at $\left(\widehat{a}_1, a_2\left(\widehat{a}_1\right)\right)$, there are some neighborhoods X, Y of $\widehat{a}_1, a_2\left(\widehat{a}_1\right)$, respectively, such that they hold for all $(a_1, a_2) \in X \times Y$. Once the neighborhood Y is chosen, we may take X to be small enough so that $a_2(a_1) \in Y$ for all $a_1 \in X$ (because $a_2(a_1)$ is a continuous function of a_1 by Lemma 1). Thus, for any a_1 in a neighborhood of $\widehat{a}_1$, SM will choose action $a_2(a_1)$ such that $U^A\left(\overrightarrow{\pi}(a_1, a_2(a_1))\right) = U^B\left(\overrightarrow{\pi}(a_1, a_2(a_1))\right)$.

□

Theorem 4. *Suppose U_1 is quasi-concave, and suppose U_2 is fairness-kinked, with U_2^A and U_2^B being joint-monotonic, quasi-concave, normal, and continuously twice-differentiable. Suppose U_1 is either purely self-regarding, or strictly monotonic with FM's favorite transaction $(\overline{a}_1, \overline{a}_2)$ giving higher material payoff to FM than SM's favorite transaction $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. Suppose the players' material payoff function are (a_1, a_2)-additively-separable. Let $(\widehat{a}_1, \widehat{a}_2)$ denote the (necessarily unique) transaction with $\widehat{a}_1 < \overline{\overline{a}}_1$ such that $\pi_1(\widehat{a}_1, \widehat{a}_2) = \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ and $U_2(\widehat{a}_1, \widehat{a}_2) = 0$. If $U_1\left(\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\right) \geq 0$,*

and if, at $\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$,

$$U_2^A = U_2^B$$
$$\frac{\partial U_2^A}{\partial \pi_2} - p(\overline{\overline{a}}_1, \widehat{a}_2)\frac{\partial U_2^A}{\partial \pi_1} > 0$$
$$\frac{\partial U_2^B}{\partial \pi_2} - p\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\frac{\partial U_2^B}{\partial \pi_1} < 0,$$

then the unique equilibrium transaction is $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$*, which is MPE and UPE.*

Proof: We first show that $(\widehat{a}_1, \widehat{a}_2)$ exists and is the unique transaction satisfying $\pi_1(\widehat{a}_1, \widehat{a}_2) = \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$, $U_2(\widehat{a}_1, \widehat{a}_2) = 0$, and $\widehat{a}_1 < \overline{\overline{a}}_1$. Given A2, A3, and A6, 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 $\left(\overline{\overline{\pi}}_1, \widehat{\pi}_2\right)$. Define $\widetilde{a}_2\left(a_1\right) \equiv \left\{a_1 : \pi_1(a_1, \widetilde{a}_2\left(a_1\right)) = \overline{\overline{\pi}}_1\right\}$, which is a continuous, strictly increasing function (by A2), and a strictly convex function (by the concavity assumption on FM's material payoff function). Define $\widetilde{\widetilde{a}}_2\left(a_1, \pi_2\right) \equiv \left\{a_1 : \pi_2(a_1, \widetilde{\widetilde{a}}_2) = \pi_2\right\}$, which is a continuous function, strictly increasing in both arguments, and strictly concave in a_1. From the proof of Theorem 1, we know there is a unique a_1 such that $\widetilde{a}_2\left(a_1\right) = \widetilde{\widetilde{a}}_2\left(a_1, \overline{\overline{\pi}}_2\right)$, which is $\overline{\overline{a}}_1$. Together with the fact that $\widehat{\pi}_2 < \overline{\overline{\pi}}_2$, these observations imply that there is a unique a_1 such that $\widetilde{a}_2\left(a_1\right) = \widetilde{\widetilde{a}}_2\left(a_1, \widehat{\pi}_2\right)$ and $a_1 < \overline{\overline{a}}_1$. We also note that $(\widehat{a}_1, \widehat{a}_2) \ll \left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$.

We next show that $\overline{\overline{a}}_1$ is a local optimum for FM. By hypothesis,

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

at $\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. This ensures $\frac{\partial U_2^A}{\partial \pi_2} \geq 0$ at $\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ (otherwise we would have to have $\frac{\partial U_2^A}{\partial \pi_1} < 0$, violating joint-monotonicity). Since $\overline{\overline{a}}_2 \geq \widehat{a}_2$ and $-\frac{\partial p}{\partial a_2} > 0$ (Technical Lemma), we have $-p\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right) > -p(\overline{\overline{a}}_1, \widehat{a}_2)$ and so

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

at $\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ also. Hence Lemma 3 implies that SM's optimal strategy $a_2\left(a_1\right)$ satisfies the fairness rule 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 social preferences are purely self-regarding or monotonic. To show that $\overline{\overline{a}}_1$ is a *global* optimum for FM, we first prove a preparatory claim.

Preparatory claim: We claim that

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

at *all* individually-rational transactions (a_1, a_2) such that $\pi_1(a_1, a_2) > \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. Suppose to the contrary there were some $\overrightarrow{\pi}(a_1, a_2)$ at which $\frac{\partial U_2^A}{\partial \pi_2} - p(\overline{\overline{a}}_1, \widehat{a}_2)\frac{\partial U_2^A}{\partial \pi_1} \leq 0$; we will show that

$\frac{\partial U_2^A}{\partial \pi_1} \geq 0$. There are two cases. When $\frac{\partial U_2^A}{\partial \pi_2} - p(\overline{\overline{a}}_1, \widehat{a}_2)\frac{\partial U_2^A}{\partial \pi_1} = 0$, $\frac{\partial U^A}{\partial \pi_1}$ and $\frac{\partial U^A}{\partial \pi_2}$ have the same sign since $p(\overline{\overline{a}}_1, \widehat{a}_2) > 0$, and so $\frac{\partial U_2^A}{\partial \pi_1} \geq 0$ there (else joint-monotonicity is violated). And when $\frac{\partial U_2^A}{\partial \pi_2} - p(\overline{\overline{a}}_1, \widehat{a}_2)\frac{\partial U_2^A}{\partial \pi_1} < 0$, we must again have $\frac{\partial U_2^A}{\partial \pi_1} \geq 0$ (else $\frac{\partial U_2^A}{\partial \pi_2} < 0$, violating joint-monotonicity). So by choosing a value k slightly larger than $p(\overline{\overline{a}}_1, \widehat{a}_2)$, we must have

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

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

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

at $\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$.

If we draw 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, SM's most-preferred point on l is below $\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ and his most-preferred point on l' is above $\overrightarrow{\pi}(a_1, a_2)$. 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 the normal good assumption is violated; a contradiction.

We now prove that $\overline{\overline{a}}_1$ is the global optimum for FM in two cases, but before proceeding, we make three notes.

First, at any individual 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. By A2, we have $(a_1, a_2) \ll (\widehat{a}_1, \widehat{a}_2)$. Assuming for now that $\frac{-\partial \pi_1(\widehat{a}_1, \widehat{a}_2)/\partial a_1}{\partial \pi_1(\widehat{a}_1, \widehat{a}_2)/\partial a_2} < \frac{\partial \pi_2(\widehat{a}_1, \widehat{a}_2)/\partial a_1}{-\partial \pi_2(\widehat{a}_1, \widehat{a}_2)/\partial a_2}$, then by A2 and strict quasi-concavity of π_2, $\pi_2(a_1, a_2) \leq \pi_2(\widehat{a}_1, \widehat{a}_2)$; a contradiction. We now show that $\frac{-\partial \pi_1(\widehat{a}_1, \widehat{a}_2)/\partial a_1}{\partial \pi_1(\widehat{a}_1, \widehat{a}_2)/\partial a_2} < \frac{\partial \pi_2(\widehat{a}_1, \widehat{a}_2)/\partial a_1}{-\partial \pi_2(\widehat{a}_1, \widehat{a}_2)/\partial a_2}$. Because $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ is MPE, $\frac{-\partial \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)/\partial a_1}{\partial \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)/\partial a_2} = \frac{\partial \pi_2\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)/\partial a_1}{-\partial \pi_2\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)/\partial a_2}$. Since $(\widehat{a}_1, \widehat{a}_2) \ll \left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$, $\pi_1(a_1, a_2) = \pi_1(\widehat{a}_1, \widehat{a}_2)$, and $\pi_2(a_1, a_2) > \pi_2(\widehat{a}_1, \widehat{a}_2)$, the strict quasi-concavity of π_1 and π_2 combined with A2 imply that $\frac{-\partial \pi_1(\widehat{a}_1, \widehat{a}_2)/\partial a_1}{\partial \pi_1(\widehat{a}_1, \widehat{a}_2)/\partial a_2} < \frac{-\partial \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)/\partial a_1}{\partial \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)/\partial a_2}$ and $\frac{\partial \pi_2\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)/\partial a_1}{-\partial \pi_2\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)/\partial a_2} < \frac{\partial \pi_2(\widehat{a}_1, \widehat{a}_2)/\partial a_1}{-\partial \pi_2(\widehat{a}_1, \widehat{a}_2)/\partial a_2}$.

Second, because the material payoff functions are additively separable, the slope of any budget curve, $p(a_1, a_2)$, does not depend on a_1, and so we will write it as $p\left(a_2\right)$.

Third, since $U_2^A\left(\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\right) = U_2^B\left(\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)\right)$, and since SM's fairness rule is strictly increasing, the relevant part of SM's utility function for any (π_1, π_2) with $\pi_1 \geq \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ is U_2^A.

Case 1: FM is purely self-regarding. To show 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 $(a_1, a_2\left(a_1\right)) \neq \left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ that satisfies $\pi_1(a_1, a_2\left(a_1\right)) \geq \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. Suppose to the contrary that there exists an individually-rational transaction $(a_1', a_2\left(a_1'\right)) \neq \left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ such that $\pi_1(a_1', a_2\left(a_1'\right)) \geq$

$\pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. Then $\frac{\partial U_2^A}{\partial \pi_2} - p(a_2(a_1')) \frac{\partial U_2^A}{\partial \pi_1} = 0$ at $\overrightarrow{\pi}(a_1', a_2(a_1'))$. If $\pi_1(a_1', a_2(a_1')) = \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$, then since $(a_1', a_2(a_1')) \gg (\widehat{a}_1, \widehat{a}_2)$ and $\frac{\partial p}{\partial a_2} < 0$, we have $p(a_2(a_1')) < p(\widehat{a}_2)$. So if $\frac{\partial U_2^A}{\partial \pi_1} > 0$ then $\frac{\partial U_2^A}{\partial \pi_2} - p(a_2(a_1')) \frac{\partial U_2^A}{\partial \pi_1} > \frac{\partial U_2^A}{\partial \pi_2} - p(\widehat{a}_2) \frac{\partial U_2^A}{\partial \pi_1} \geq 0$ at $(a_1', a_2(a_1'))$ (using Preparatory Claim); and if $\frac{\partial U_2^A}{\partial \pi_1} = 0$, then $\frac{\partial U_2^A}{\partial \pi_2} - p(a_2(a_1')) \frac{\partial U_2^A}{\partial \pi_1} > 0$ at $(a_1', a_2(a_1'))$ by TA1; a contradiction.

So we must have $\pi_1(a_1', a_2(a_1')) > \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. By A2, 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 $a_2' < a_2(a_1')$. If we draw budget lines m, m' with slopes $-p(a_2')$ and $-p(a_2(a_1'))$ passing through the two points $\overrightarrow{\pi}(a_1', a_2')$ and $\overrightarrow{\pi}(a_1', a_2(a_1'))$, SM's most-preferred point on line m is below $\overrightarrow{\pi}(a_1', a_2')$ (since $-\frac{\partial U_2^A(\overline{\overline{a}}_1, \overline{\overline{a}}_2)/\partial \pi_2}{\partial U_2^A(\overline{\overline{a}}_1, \overline{\overline{a}}_2)/\partial \pi_1} < -p\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ i.e., at $\overrightarrow{\pi}\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$ SM's indifference curve is steeper than the budget curve $B\left(\overline{\overline{a}}_1\right)$ and using Technical Lemma) and his most-preferred point on m' is (by definition) $\overrightarrow{\pi}(a_1', a_2(a_1'))$. Now if we draw a third line m'' with slope $-p(a_2(a_1')) > -p(a_2')$ going through $\overrightarrow{\pi}(a_1', a_2(a_1'))$ (moving from m' to m'' can be thought of as a Slutsky compensated price change), then SM's most-preferred point on line m'' must be above $\overrightarrow{\pi}(a_1', a_2(a_1'))$. But this is a violation of the normal good assumption; a contradiction.

Case 2: FM's distributional preferences are strictly monotonic and $\pi_1(\overline{a}_1, \overline{a}_2) > \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$. 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 and $\pi_1(\overline{a}_1, \overline{a}_2) > \pi_1\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$, 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 5 (Rotten Kid Theorem). *Suppose U_1 is quasi-concave, and either purely self-regarding or monotonic. Suppose U_2 is monotonic, quasi-concave, and normal. In the rotten kid game (Example 2), the unique equilibrium transaction is SM's favorite transaction $\left(\overline{\overline{a}}_1, \overline{\overline{a}}_2\right)$, which maximizes family income.*

Proof: Follows directly from Theorem 3. 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.

□

Theorem 6. *Suppose U_1 and U_2 are joint-monotonic, quasi-concave, and potentially fairness-kinked. If the equilibrium $(a_1^*, a_2(a_1^*))$ is UPE, then at least one of the following must be true:*

1. *$(a_1^*, a_2(a_1^*))$ is also MPE.*

2. *U_2 is fairness-kinked at $\overrightarrow{\pi}(a_1^*, a_2(a_1^*))$, and SM's indifference curve for disadvantageously unfair transactions is tangent to SM's fairness rule at $\overrightarrow{\pi}(a_1^*, a_2(a_1^*))$.*

3. *U_1 and U_2 are fairness-kinked at $\overrightarrow{\pi}(a_1^*, a_2(a_1^*))$, and the fairness rules f_1 and f_2 have different slopes at $\overrightarrow{\pi}(a_1^*, a_2(a_1^*))$.*

Proof: First we show that if U_2 is continuously differentiable at $\overrightarrow{\pi}(a_1^*, a_2(a_1^*))$, then any UPE equilibrium is also MPE. From the proof of Theorem 1, we know that a material payoff pair can be UPE but not MPE only if it is strictly within the materially-feasible set and occurs at a tangency between FM's and SM's interpersonal indifference curves. Further, the tangency line through $\overrightarrow{\pi}(a_1^*, a_2(a_1^*))$ must be weakly positively sloped. However, according to Lemma 1, if U_2 is continuously differentiable at $\overrightarrow{\pi}(a_1^*, a_2(a_1^*))$, then SM's indifference curve is strictly negatively-sloped there. This proves the claim.

Next we show that if a UPE equilibrium occurs at a non-MPE transaction $(a_1^*, a_2(a_1^*))$, U_2 is fairness-kinked at $\overrightarrow{\pi}(a_1^*, a_2(a_1^*))$, and 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 $\overrightarrow{\pi}(a_1^*, a_2(a_1^*))$. As above, we know that $\overrightarrow{\pi}(a_1^*, a_2(a_1^*))$ must lie strictly within the materially-feasible set and occur at a weakly upward-sloping tangency between FM's and SM's interpersonal indifference curves. Since U_2 is fairness-kinked at $\overrightarrow{\pi}(a_1^*, a_2(a_1^*))$, the curve traced out by $\{\overrightarrow{\pi}(a_1, a_2(a_1))\}_{a_1 \in \mathbb{R}}$ must be tangent to SM's fairness rule at $\overrightarrow{\pi}(a_1^*, a_2(a_1^*))$. Since $\overrightarrow{\pi}(a_1^*, a_2(a_1^*))$ is an equilibrium and U_1 is continuously differentiable at $\overrightarrow{\pi}(a_1^*, a_2(a_1^*))$, FM's indifference curve is tangent to the curve traced out by $\{\overrightarrow{\pi}(a_1, a_2(a_1))\}_{a_1 \in \mathbb{R}}$ at $\overrightarrow{\pi}(a_1^*, a_2(a_1^*))$ and hence also tangent to SM's fairness rule at that point. Since U_2 is fairness-kinked at $\overrightarrow{\pi}(a_1^*, a_2(a_1^*))$, SM's indifference curve in the region of disadvantageously unfair transactions must be weakly steeper at $\overrightarrow{\pi}(a_1^*, a_2(a_1^*))$ than SM's fairness rule. But if it were strictly steeper, then $\overrightarrow{\pi}(a_1^*, a_2(a_1^*))$ would be utility-Pareto dominated by an arbitrarily close material payoff pair that is also on SM's disadvantageously-unfair indifference curve. We conclude that SM's indifference curve must be tangent.

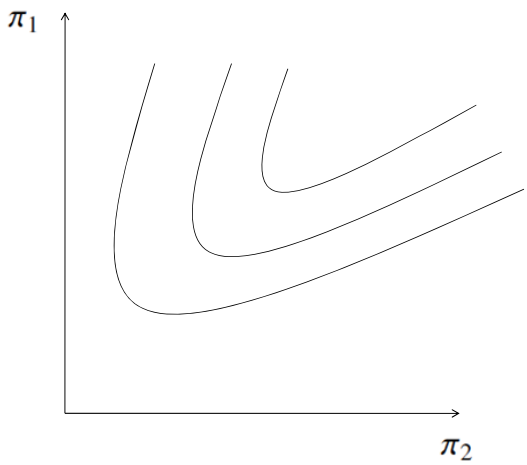
Finally, note that if a UPE equilibrium occurs at a non-MPE transaction $(a_1, a_2(a_1))$, and if U_1 and U_2 are each fairness-kinked at $(\pi_1(a_1, a_2(a_1)), \pi_2(a_1, a_2(a_1)))$, then the players must be following different fairness rules at $\overrightarrow{\pi}(a_1, a_2(a_1))$. If they were following the same fairness rule locally, then FM could get higher utility by deviating to an action that induces SM to implement a material payoff pair further northeast along the shared fairness rule.

□

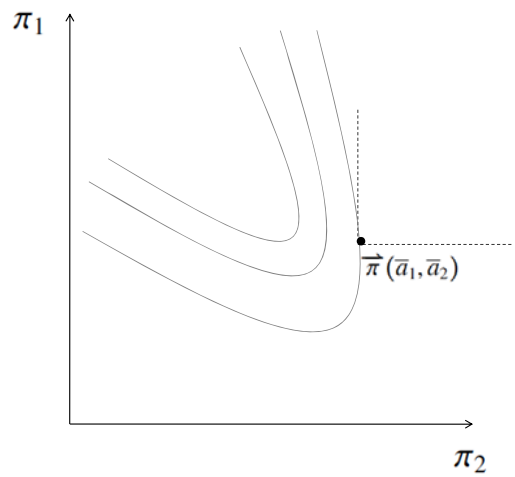
References

[1] Rangarajan K. Sundaram. *A First Course in Optimization Theory.* Cambridge University Press, Cambridge, UK, 1996.

[1a]



[1b]



[1c]

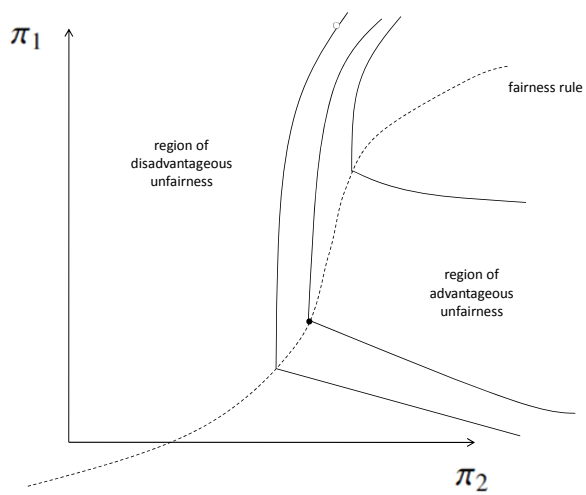
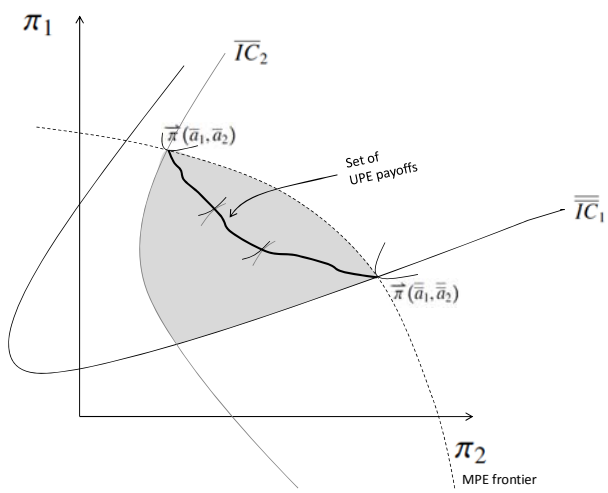
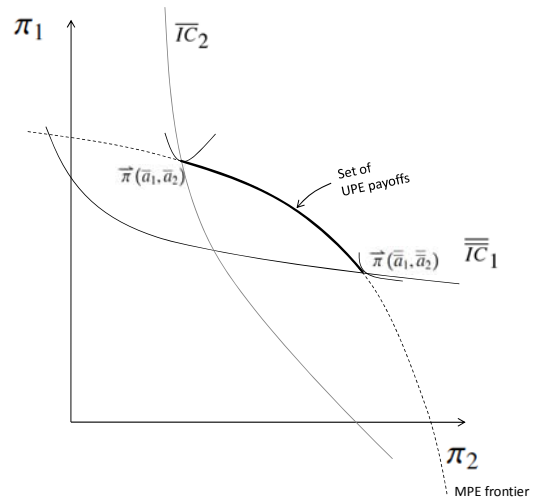


Figure 1. Interpersonal indifference curves. Panel (a): Preferences that are joint-monotonic but not monotonic. Panel (b): Preferences that violate joint-monotonicity: At material payoff pair $\vec{\pi}(\bar{a}_1, \bar{a}_2)$, there is no alternative material payoff pair to the northeast that gives higher utility. Panel (c): Fairness-kinked preferences. Due to the non-monotonicity, the black point is preferred to the white point that gives higher material payoffs to both players.

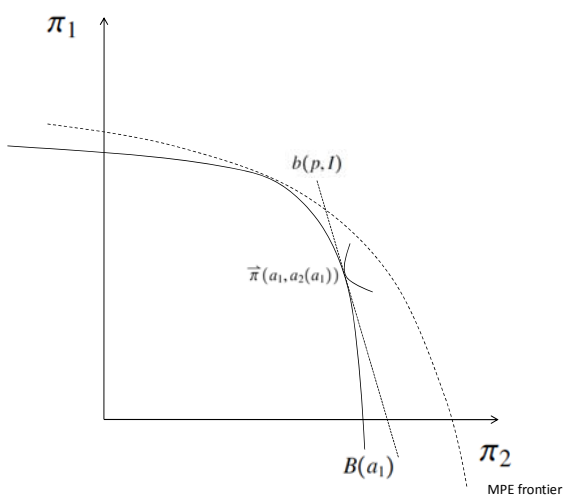
[2a]



[2b]



[3]



[4]

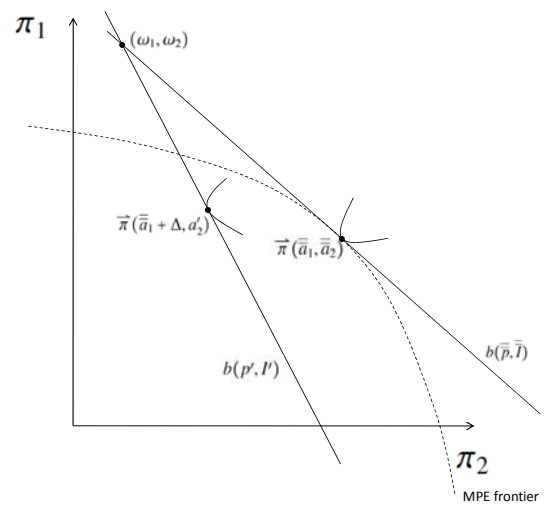
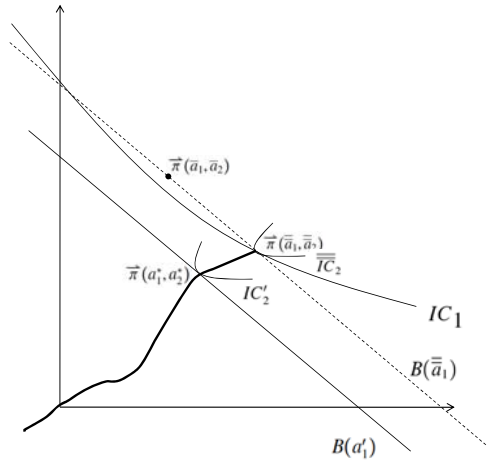


Figure 2. Relationship between utility Pareto efficiency and material Pareto efficiency. Panel (a): Both players have joint-monotonic preferences. Panel (b): One of the players (here, SM) has monotonic preferences.

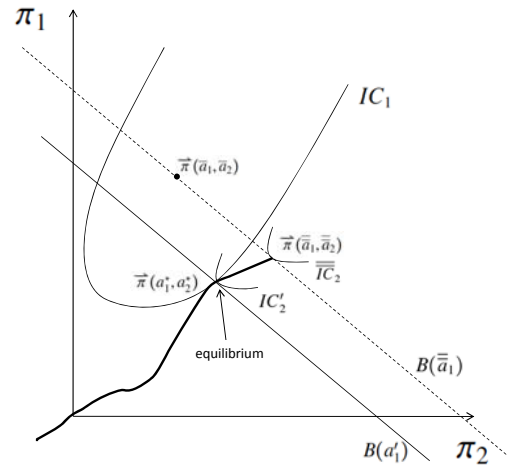
Figure 3. SM's optimal choice on the budget curve $B(a_1)$, $a_2(a_1)$, is the same as SM's optimal choice on the budget line $b(p, I)$ that first-order approximates the budget curve.

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

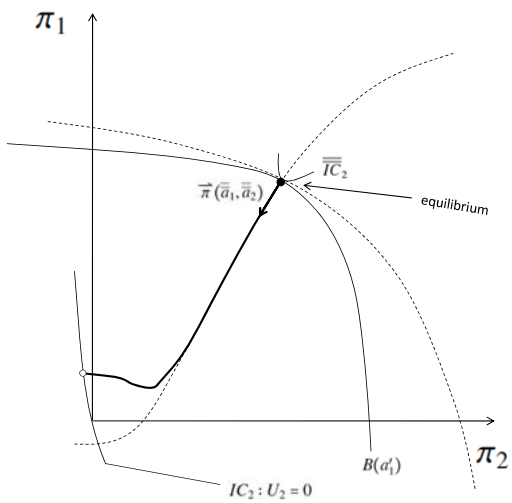
[5a]



[5b]



[6a]



[6b]

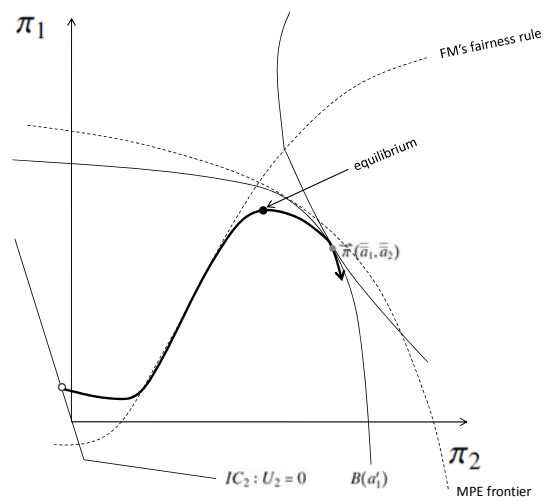


Figure 5. Case I: Budget curves are parallel shifts. In both panels, the dark line with the arrow shows the path of material payoff pairs that could occur, $\vec{\pi}(a_1, a_2(a_1))$, given different possible actions by FM. This path increases up to SM's favorite transaction and then goes back down the same path (if FM takes an "inefficiently high" level of her action). Panel (a): FM's social preferences are monotonic, and the equilibrium occurs at SM's favorite transaction. Panel (b): FM's social preferences are joint-monotonic, and the equilibrium occurs elsewhere.

Figure 6. Case II: SM's has fairness-kinked social preferences. In both panels, the dark line with the arrow shows the path of material payoff pairs that could occur, $\vec{\pi}(a_1, a_2(a_1))$, 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.