

# Social Preferences and the Efficiency of Bilateral Exchange

Daniel J. Benjamin\*

*Cornell University and Institute for Social Research*

November 11, 2008

## Abstract

Without contracting or repetition, purely self-regarding agents will not trade. To what extent can social preferences, such as altruism or a concern for fairness, generate efficient bilateral exchange? I analyze a simple exchange game: A purely self-regarding first mover transfers some amount of a commodity to a second mover. Then the second mover, who has social preferences defined over material payoffs, transfers a commodity back to the first mover. I identify key properties of social preferences that matter for bilateral exchange behavior. I find the equilibrium will be efficient if either (1) the second mover's transfer is monetary (i.e., material payoffs are quasi-linear in the second mover's commodity), or (2) the second mover's social preferences cause him to behave in accordance with a "fairness rule" (such as the 50-50 sharing norm). The results may explain why small-scale transactions with discretionary monetary payment are common, and suggest that social norms that prescribe fair allocations promote efficiency in exchange environments.

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

*Keywords:* social preferences, fairness, altruism, gift exchange, norms, Rotten Kid theorem

---

\*A previous version of this paper circulated as part of an essay, "A Theory of Fairness in Labor Markets." 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, David Laibson, Ted O'Donoghue, 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, 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

The efficiency of bilateral exchange has long been a central issue in economics. Under some institutional arrangements, it is well-understood that exchange will be efficient. When both parties' actions can be bound by an enforceable contract, the Coase theorem implies that the parties will agree to a Pareto efficient transaction (Coase 1960). In the absence of enforceable contracts, if the exchange will be repeated, folk theorems imply that some equilibria are efficient if the parties are sufficiently patient (e.g., Friedman 1971; Fudenberg & Maskin 1986). This paper addresses a third possible source of increased efficiency: a direct concern for the welfare of the other party—often called social (or interpersonal) preferences.

I analyze theoretically a simple bilateral exchange environment where contracting is infeasible and the exchange is one-shot (this environment includes the “gift-exchange game” and “trust game” as special cases). I focus on this setting primarily because it lays bare the role of social preferences by ruling out contracts and repetition. It is also a relevant economic environment in its own right. In a few especially simple situations, a customer's payment for a good or service is not contracted in advance—for example, tipping (Conlin, Lynn, & O'Donoghue 2003) or payment on the honor system (Dawes & Thaler 1988; Dubner & Levitt 2004). More generally, analysis of the no-contracts, no-repetition scenario is useful for understanding factors that may be relevant in environments with contracts or repetition. Hart & Moore (2008) argue that in exchange relationships that *are* contractible, the legal contract can typically only require “perfunctory” performance, so the principal must rely on the good will of the agent for better-than-perfunctory performance. Similarly, Akerlof (1982) and Akerlof & Yellen (1990) argue that workers' social preferences play a major role in determining how much effort they provide in excess of a minimum work standard, even though employment relationships often involve both contracts and repetition (see also Bewley 1999).

In one-shot bilateral exchange environments without contracting, both laboratory experiments and field experiments have found that generous behavior is often reciprocated, enabling trade to occur and sometimes leading to moderate or high levels of efficiency (e.g., Fehr, Kirchsteiger, & Riedl 1993).<sup>1</sup> Several papers have found that particular functional forms for social preferences

---

<sup>1</sup>Hundreds of laboratory experiments since Fehr, Kirchsteiger, & Riedl (1993) and Berg, Dickhaut, & McCabe (1995) have documented reciprocal behavior in gift-exchange games and trust games, respectively. There have also been a number of recent field experiments. For example, in two field experiments, Gneezy & List (2006) hired individuals for data-entry or door-to-door fundraising. They did not tell the workers that the workers were participants in an experiment, and they were careful to rule out non-fairness explanations for why effort would be increasing in the wage. Workers who were paid more entered more data and raised more money, respectively. Although Gneezy & List (2006) found that effort increased only for the first few hours, other field experiments with larger sample sizes have found persistent effort effects (e.g., Al-Ubaydli, Andersen, Gneezy, & List 2006). Additional field experiments include Kube, Maréchal, & Puppe (2007a); Kube, Maréchal, & Puppe (2007b); Cohn, Fehr, & Goette (2007); Bellemare &

can account reasonably well for observed behavior (e.g., Fehr, Klein, & Schmidt 2007). However, little is known theoretically about the conditions for efficiency when exchange is motivated by social preferences.<sup>2</sup> Understanding the conditions for efficiency is important for interpreting the existing evidence, as well as for predicting to which real-world settings the laboratory findings will generalize. This paper asks: What social preferences (if any) lead to efficient exchange? When the equilibrium is not fully efficient, what factors determine how inefficient it will be?

Economic analysis involving fairness preferences generally proceeds by studying a particular functional form (e.g., Fehr, Klein, & Schmidt 2007). Instead, my approach is to study properties of a generalized model of social preferences.<sup>3</sup> This approach has three significant advantages. First, because researchers disagree about important features of social preferences—such as the prevalence and intensity of a desire to harm the player who is ahead (e.g., Engelmann & Strobel 2004; Fehr, Naef, & Schmidt 2006)—it is important to understand whether and how results are sensitive to these features. Second, focusing on properties of preferences makes it possible to discover that assumptions made implicitly by existing fairness models are actually quite central. Finally, studying fairness and altruistic preferences within the same framework allows me to discuss the relationship between conclusions from the literatures on altruism (economics of the family) and fairness (experimental/behavioral economics), literatures which have generally proceeded separately from each other.

I study a bilateral exchange game that has two stages. The first mover takes an action that transfers a commodity from herself to the second mover. Then the second mover takes an action that transfers a commodity from himself to the first mover. Since the second mover has no extrinsic incentive to help the first mover at his own expense, no trade would occur if the second mover were purely self-regarding. Instead, I assume that the second mover's has "social preferences": his utility depends on both his own and the other player's "material payoff," the standard (purely

---

Shearer (2007); Greenberg (1990); and Pritchard, Dunnette, & Jorgenson (1972).

<sup>2</sup>Rabin (1997) addresses the efficiency of 2-player, 2-action, sequential-move games where a purely self-regarding first mover can make a "fair" or "unfair" offer, and the fair-minded second mover can "accept" or "punish." He finds that the equilibrium outcome is relatively efficient (in terms of money payoffs) in games where the first mover's available "unfair" offer is not particularly unfair; where the second mover's available "punish" action does not harm the first mover much; or where the "punish" action harms the first mover a great deal. Rabin's results qualitatively differ from the results in this paper largely because I assume continuous action spaces, while Rabin's conclusions rely on the players not being able to choose the extremeness of their actions. In my set-up with continuous actions, the second mover would never choose to "punish"; see the discussion of part 4 of Lemma 1 below. Also relatedly, Dufwenberg, Heidhues, Kirchsteiger, Riedel, & Sobel (2008) address the efficiency of general equilibrium when agents have social preferences.

<sup>3</sup>Relatedly, Cox, Friedman, & Sadiraj (2008) propose a unifying framework for existing models of social preferences. In order to fit data from experimental games, they formalize assumptions that more generous actions induce more altruistic preferences, and that this effect is stronger for acts of commission than for acts of omission. I do not address these aspects of behavior.

self-regarding) payoff from consumption of goods, leisure, etc. Because it is the second mover's social preferences that drive the main results and key intuitions, and because the analysis becomes much more complicated when the first mover also has social preferences, I assume for most of the analysis that the first mover just maximizes her material payoff. At the end of the paper, I discuss how the results generalize when both players have social preferences.

Because of social preferences, two distinct notions of efficiency are potentially relevant. "Utility Pareto efficiency" takes into account the fact that the second mover's utility depends on both players' material payoffs. By contrast, I call a transaction "materially Pareto efficient" if there is no alternative transaction that could have increased one player's *material* payoff without reducing the other player's. Theorem 1 shows that for a broad class of social preferences, these two efficiency concepts are tightly linked in the bilateral exchange game: every utility Pareto efficient transaction is also a materially Pareto efficient transaction. I also find that if the equilibrium of the game is materially Pareto efficient, then it is also utility Pareto efficient, and it occurs at the transaction that is most preferred by the second mover.

Two properties that the second mover's social preferences could have turned out to be central in determining whether the equilibrium occurs at this efficient transaction. The first property is that 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. Normality is implicitly assumed in all functional forms for fairness preferences that have been proposed, but previous work has not appreciated its role in driving the results of those models. I show that normality is the key property of social preferences that generates behavior that looks like reciprocity (the second mover transferring more in response to a greater transfer by the first mover), often considered a hallmark of fair-minded behavior. Normality also plays a role as a sufficient condition ensuring an efficient equilibrium in the important cases outlined below.

A second important property is that the utility function could be "fairness-kinked." Although in consumer theory a kinked utility function over goods is merely an illustrative extreme case, kinked *social* preferences over material payoffs are realistic. Several leading functional forms for fairness preferences assume that the utility function is kinked wherever both players receive equal material payoffs (Fehr & Schmidt 1999; Charness & Rabin 2002; some versions of Bolton & Ockenfels 2000). That is why these models are consistent with the preponderance of 50-50 sharing in laboratory experiments (Camerer 2003). More generally, a fairness-kinked utility function describes behavior in which people adhere to a social norm (such as the 50-50 sharing norm) that requires that both

players share in the gains from a larger pie.

The central result of the paper (Theorem 2) proves that at least one of two conditions is necessary for the equilibrium exchange to be efficient: (1) both players' material payoffs are linear in the second-mover's action, or (2) the second mover's interpersonal indifference curve at the equilibrium is kinked. When neither is satisfied—that is, when the the second mover's interpersonal indifference curves are smooth *and* the second mover faces a convex cost of increasing his transfer to the first mover—the equilibrium *cannot* be efficient. In other words, the first mover can profitably deviate from the action that (in combination with the second mover's reaction) would generate an efficient transaction. Deviating is profitable despite a second-order loss in the total gains from trade because the first mover gets a first-order increase in her share of the gains from trade.

Theorem 2 suggests that there are two cases in which the equilibrium could be efficient. The first is where both players' material payoffs are linear in the second mover's action—"quasi-linear," a.k.a. "transferable," material payoffs. This is a standard assumption for situations where the second mover's action is a monetary transfer. In the bilateral exchange game, the monetary transfer is not specified in advance—for example, it is the discretionary component of payment to a contractor who provides housework. Theorem 3 shows that linearity in the second mover's action, along with normality of the second mover's social preferences, are together sufficient conditions for efficiency. This result may explain why small-scale commercial transactions with discretionary monetary payment are so common.

Theorem 3 generalizes the well-known Rotten Kid theorem (Becker 1974). That result states that if a parent (the second mover) is altruistic, then a selfish child (the first mover) will act so as to maximize family income. Traditionally, the Rotten Kid theorem has been interpreted as pertaining to family environments but *not* market environments because altruism is thought to be relevant primarily within the family. However, Theorem 3 shows that the efficiency conclusion actually holds for a more general class of social preferences, including fairness concerns that appear to be common in market environments. This is important because it implies that the surprising efficiency conclusion actually applies to contracting situations outside the family.

The second efficient case is where the second-mover's utility is fairness-kinked (even if the material payoffs are not linear in the second-mover's action). That is, the second mover follows a social norm that requires that both players share in the gains from a larger pie (such as the 50-50 sharing norm). In that case, Theorem 4 states that if the second mover's social preferences are normal and sufficiently kinked, the equilibrium will be efficient. This result may explain why social norms have developed that prescribe fair allocations of gains from trade.

Consistent with a long-standing conjecture among social scientists (e.g., Arrow 1971, p.22), the results in this paper suggest that social norms can enable efficient economic exchange, even in the absence of enforceable contracts. Therefore, in many settings, the main consequence of contracting may be redistribution, rather than efficiency enhancement. The efficient equilibrium that relies on social preferences gives the second mover his most-preferred transaction, whereas with an enforceable contract, the division of surplus is determined by bargaining power at the time of contracting, and the timing of actions that carry out the contract does not matter for the division of surplus.

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 results on equilibrium efficiency. Section 6 discusses how the results generalize when both players have social preferences. Section 7 discusses how these results apply to the hold-up problem, as well as three possible other extensions of the analysis: uncertainty about the other players's preferences (as in Fehr, Klein, & Schmidt 2007), social preferences that depend on the perceived intentions of the other player (as in Rabin 1993 and Levine 1998), and concern about how the terms of exchange compare with the terms received by others (often relevant in multi-worker firms). The appendix contains proofs.

## 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 can take an action  $a_1$  that transfers a commodity from herself to the second mover. The second mover can then choose an action  $a_2$  that transfers a commodity from himself to the first mover. 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 no transfers occur, and both players receive an outside option payoff.<sup>4</sup>

The players' **material payoffs** represent the purely self-regarding component of payoffs. The players' respective material payoff functions are

$$\pi_1(a_1, a_2) = a_2 - a_1, \tag{1}$$

$$\pi_2(a_1, a_2) = v(a_1) - c(a_2). \tag{2}$$

---

<sup>4</sup>As long as trade occurs in equilibrium, the results are not sensitive to the exact assumption about when an outside option is available and to whom.

The function  $v(\cdot)$  reflects concave benefits to the second mover of the first mover's transfer:  $v' > 0$ ,  $v'' < 0$ ,  $\lim_{a_1 \rightarrow -\infty} v'(a_1) = \infty$ ; and  $c(\cdot)$  reflects convex costs to the second mover of the second mover's transfer:  $c' > 0$ ,  $c'' > 0$ ,  $\lim_{a_2 \rightarrow \infty} c'(a_2) = \infty$ .<sup>5</sup> I have written the first mover's material payoff function as linear in  $a_1$  and  $a_2$  to keep notation brief; as long as the material payoff functions are additively separable, they can be represented as (1)-(2) via an appropriate change in variables.<sup>6</sup> Additive-separability of the material payoffs simplifies the analysis but is not crucial for the main results in this paper.<sup>7</sup>

The second mover maximizes utility,  $U(\pi_1, \pi_2)$ , that depends on both his own and on the first mover's material payoff and may have some or all of the properties described later, in Section 3. For brevity, I slightly abuse notation by sometimes writing the second mover's utility function as  $U(a_1, a_2)$  instead of  $U(\pi_1(a_1, a_2), \pi_2(a_1, a_2))$ . To keep the analysis simpler, I assume that the first mover just maximizes her own material payoff,  $\pi_1(a_1, a_2)$ , but in Section 6, I discuss how the main conclusions generalize when the first mover also has social preferences.

I normalize  $v(0) = 0$  and  $c(0) = 0$ , each player's outside option material payoff is 0, and the second mover's outside option utility is  $U(0, 0) = 0$ . I assume that  $v'(0) > c'(0)$  so that there are potential (material) gains from trade.

## 2.1 Equilibrium

The solution concept is subgame-perfect Nash equilibrium. Unlike in a typical principal-agent problem, the first mover cannot make her transfer a function of any variable that depends on the second-mover's action. Therefore the second mover has no extrinsic incentive to make a transfer to the first mover. Clearly, if the second mover were purely selfish, with utility function  $U(\pi_1, \pi_2) =$

---

<sup>5</sup>The assumptions that  $\lim_{a_1 \rightarrow -\infty} v'(a_1) = \infty$  and  $\lim_{a_2 \rightarrow \infty} c'(a_2) = \infty$  play a purely technical role in ensuring that the space of individually-rational material payoffs is compact and hence that an equilibrium exists.

<sup>6</sup>For example, suppose the first-mover's material payoff is increasing in her consumption of both Good 1,  $g_1^1$ , and her consumption of Good 2,  $g_1^2$ :

$$\pi_1(g_1^1, g_1^2) = u_1^1(g_1^1) + u_1^2(g_1^2),$$

with  $u_1^1 > 0$ ,  $u_1^{1''} \leq 0$ ,  $u_1^2 > 0$ , and  $u_1^{2''} \leq 0$ . The second-mover's material payoff is also increasing in his consumption of both goods:

$$\pi_2(g_2^1, g_2^2) = u_2^1(g_2^1) + u_2^2(g_2^2),$$

with  $u_2^1 > 0$ ,  $u_2^{1''} \leq 0$ ,  $u_2^2 > 0$ , and  $u_2^{2''} \leq 0$ . Letting  $g^1 = g_1^1 + g_2^1$  and  $g^2 = g_1^2 + g_2^2$  denote the aggregate amounts of the two goods, the following transformation gives (1)-(2):  $a_2 = u_1^2(g_1^2)$ ,  $a_1 = -u_1^1(g^1 - g_2^1)$ ,  $v(a_1) = u_2^1(g^1 - (u_1^1)^{-1}(-a_1))$ , and  $c(a_2) = -u_2^2(g^2 - (u_1^2)^{-1}(a_2))$ . Note that  $c(\cdot)$  is convex if either  $u_1^2(\cdot)$  or  $u_2^2(\cdot)$  is concave.

<sup>7</sup>An exception is part 2 of Lemma 1, for which additive-separability *is* crucial. Theorem 4 does not require additively-separable material payoffs and therefore also applies (with appropriate boundary conditions imposed) when the first mover has the non-additively-separable material payoff function often used in laboratory gift-exchange experiments:  $\pi_1(a_1, a_2) = (k_1 - a_1)a_2$  and  $\pi_2(a_1, a_2) = a_1 - c(a_2) - k_2$ , where  $a_1 \leq k_1$ ,  $a_2 \geq 0$ , and where  $k_1 > 0$  and  $k_2$  are constants (e.g., Fehr, Kirchsteiger, & Riedl 1993).

$\pi_2$ , then regardless of the first mover's action, the second mover would take minimal action (here actually, negative infinity, since the actions are unbounded). There would be no exchange because the first mover would prefer her outside option. Hence, any exchange that occurs in equilibrium is a consequence of the second mover'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.

At the solution to a typical principal-agent problem, the second mover's "participation constraint,"  $U(a_1, a_2) \geq 0$ , is binding. That is because the first mover can reduce the level of her transfer without affecting the second-mover's best response. Here, because the level of the first mover's action will in general affect the second-mover's reaction, that constraint may not bind. I call an equilibrium  $(a_1, a_2)$  **interior** if  $U(a_1, a_2) > 0$ .

## 2.2 Efficiency

Recall that an exchange is defined to be Pareto efficient if there is no alternative exchange that could have made one party better off without making the other worse off. Here, there are two possible interpretations of Pareto efficiency, depending on whether the second mover's welfare is measured by his material payoff or by his utility.

**Definition 1** *A transaction  $(a_1, a_2)$  is **utility Pareto efficient** if there is no other transaction  $(\hat{a}_1, \hat{a}_2)$  such that  $\pi_1(\hat{a}_1, \hat{a}_2) \geq \pi_1(a_1, a_2)$  and  $U(\hat{a}_1, \hat{a}_2) \geq U(a_1, a_2)$ , at least one inequality strict.*

**Definition 2** *A transaction  $(a_1, a_2)$  is **materially Pareto efficient** if there is no other transaction  $(\hat{a}_1, \hat{a}_2)$  such that  $\pi_1(\hat{a}_1, \hat{a}_2) \geq \pi_1(a_1, a_2)$  and  $\pi_2(\hat{a}_1, \hat{a}_2) \geq \pi_2(a_1, a_2)$ , at least one inequality strict.*

If both players were purely self-regarding, then Pareto efficiency would be characterized by  $v'(a_1) = c'(a_2)$ . It follows that here, a transaction  $(a_1, a_2)$  is materially Pareto efficient if and only if  $v'(a_1) = c'(a_2)$ .

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).<sup>8</sup> To the extent that individuals' social preferences reflect adherence to social norms, a social

---

<sup>8</sup>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." Arrow (1971) also makes this distinction, albeit less explicitly.

planner might be interested in promoting material Pareto efficiency rather than utility Pareto efficiency. Consistent with this view, much of the existing work can be interpreted as asking whether behavior in accordance with social norms (such as tipping) leads to material Pareto efficiency (e.g., Conlin, Lynn, & O’Donoghue 2003). 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. I discuss the relationship between utility Pareto efficiency and material Pareto efficiency in the bilateral exchange game in Section 4.

When the equilibrium is not efficient, it may be of interest to describe the factors that determine how far from efficient it is. Unfortunately, there is not a unique way to measure the degree to which a transaction is inefficient. I define a notion that will turn out to be convenient.

**Definition 3** *The marginal (material) inefficiency of a transaction  $(a_1, a_2)$  is*

$$\left. \frac{d\pi_2(a_1, a_2)}{da_1} \right|_{\pi_1(a_1, a_2) = \bar{\pi}_1} = v'(a_1) - c'(a_2),$$

*the amount by which the second-mover’s material payoff could be increased on the margin, holding the first-mover’s material payoff constant.*

An alternative notion of the degree of inefficiency would be  $\left. \frac{d\pi_1(a_1, a_2)}{da_1} \right|_{\pi_2(a_1, a_2) = \bar{\pi}_2} = \frac{v'(a_1)}{c'(a_2)} - 1$ , the amount by which the first mover’s material payoff could be increased on the margin, holding the second-mover’s material payoff constant. These coincide when  $c(a_2) = a_2$ , and both equal zero at a materially Pareto efficient transaction.

This paper asks: What social preferences for the second mover lead to a materially Pareto efficient equilibrium? Utility Pareto efficient? When the equilibrium is not fully efficient, what factors determine how inefficient it will be?

## 3 Social Preferences

### 3.1 Existing Models

A player with social preferences, say the second mover, is assumed to maximize a utility function,  $U(\pi_1, \pi_2)$ , that depends on both players’ material payoffs.<sup>9</sup> These four prominent existing models help motivate the more general analysis that follows:

---

<sup>9</sup>Models designed to explain laboratory behavior write interpersonal preferences  $U(x^W, x^F)$  as a function of the monetary amounts  $x^W$  and  $x^F$  paid to participants in a laboratory experiment. To allow for more than one commodity (in the gift-exchange game, money and effort), I instead make utility  $U(\pi^W, \pi^F)$  depend on the material payoffs  $\pi^W$  and  $\pi^F$  from the transaction. If the material payoff functions are quasi-linear in money, then the  $U(\pi^W, \pi^F)$  specification specializes to the  $U(x^W, x^F)$  model in the laboratory, where money is the only relevant commodity.

- Fehr & Schmidt’s (1999) “inequity-averse preferences” have the form

$$U(\pi_1, \pi_2) = \pi_2 - \alpha \max\{\pi_1 - \pi_2, 0\} - \beta \max\{\pi_2 - \pi_1, 0\}, \quad (3)$$

where  $\alpha \geq 0$  is the second mover’s aversion to “disadvantageous unfairness” (the first mover earning more than the second mover), and  $\beta \geq 0$  is his aversion to “advantageous unfairness” (the second mover earning more than the first mover).

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

$$U(\pi_1, \pi_2) = \pi_2 - \omega \left( \frac{\pi_2}{\pi_2 + \pi_1} - \frac{1}{2} \right)^2, \quad (4)$$

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 simple exchange game, Bolton & Ockenfels (2000, pp.183-187) instead use

$$U(\pi_1, \pi_2) = -|\pi_1 - \pi_2|. \quad (5)$$

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

$$U(\pi_1, \pi_2) = \pi_2 + \gamma\pi_1 + \delta \min\{\pi_2, \pi_1\}, \quad (6)$$

where  $\gamma \geq 0$  is the second-mover’s positive regard for the other player, and  $\delta \geq 0$  is his additional concern for the person who gains least.

- Becker’s (1974) model of altruism within the family assumes that  $U(\pi_1, \pi_2)$  is continuous, monotonically increasing in both arguments, quasi-concave, and normal (i.e.,  $\pi_1$  and  $\pi_2$  enter  $U$  as normal goods).

Figure 1 illustrates interpersonal indifference curves from (3) and Becker’s (1974) altruism model. Applied economic analysis involving fairness preferences generally proceeds by studying the implications of one of the above functional forms. Instead, I will study the behavioral implications of properties of social preferences in order to understand which implications follow from which classes of models.

---

Whether it is appropriate to write social preferences as a function of the material payoffs depends on whether individuals judge the fairness of an allocation taking into account both of the relevant commodities (rather than, say, only judging the fairness of the monetary allocation). This seems especially reasonable for bilateral transactions, where the role of two commodities is salient.

### 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. By contrast, fairness preferences such as (3), (4), and (5) allow for a type of non-monotonicity: an individual may prefer to reduce the payoff of a person who is ahead. Upward-sloping regions of interpersonal indifference curves in Figure 1 reflect non-monotonicity.

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; Fisman, Kariv, & Markovits 2005). Relatedly, models of positional (or status) preferences also predict a willingness to sacrifice one’s own material payoff to reduce others’s (e.g., Heffetz & Frank 2008).

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 important to rule out too much spitefulness or self-hating, in which case trade would not occur or an equilibrium would not exist. I propose a new condition, joint-monotonicity, that appropriately weakens monotonicity.<sup>10</sup>

**Definition 4**  $U$  is *joint-monotonic* if for any  $(\pi_1, \pi_2)$ :

1. For any  $\varepsilon > 0$ , there is some  $(\hat{\pi}_1, \hat{\pi}_2)$  with  $0 < \hat{\pi}_1 - \pi_1 < \varepsilon$ ,  $0 < \hat{\pi}_2 - \pi_2 < \varepsilon$ , and  $U(\hat{\pi}_1, \hat{\pi}_2) > U(\pi_1, \pi_2)$ .
2. There exist  $\delta_1, \delta_2 > 0$  such that  $U(\pi_1 - \delta_1, \pi_2) < U(\pi_1, \pi_2)$  and  $U(\pi_1, \pi_2 - \delta_2) < U(\pi_1, \pi_2)$ .

---

<sup>10</sup>In studying social preferences in a general equilibrium environment, Dufwenberg, Heidhues, Kirchsteiger, Riedel, & Sobel (2008) independently propose a “social monotonicity” condition, which is the same as condition 1 of my joint-monotonicity property.

Condition 1 states that for any material payoff pair, there is an arbitrarily close alternative material payoff pair giving more to *both* players that the second mover 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.

However, condition 2 states that if the first mover's (second mover's) material payoff is sufficiently small relative to the second mover's (first mover's), the second mover would prefer that the first mover (second mover) have a higher material payoff. This rules out globally spiteful preferences such as  $U(\pi_1, \pi_2) = \pi_2 - \pi_1$ , and because the inequalities are strict, it rules out the purely self-regarding case  $U(\pi_1, \pi_2) = \pi_2$  and the purely altruistic case  $U(\pi_1, \pi_2) = \pi_1$ . Although descriptively reasonable, condition 2 actually plays only a technical role in the analysis that follows, ensuring that optimal strategies exist for both players. All of the above models of social preferences satisfy joint-monotonicity.

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

**Definition 5**  $U$  is **quasi-concave** if for any  $(\pi_1, \pi_2), (\hat{\pi}_1, \hat{\pi}_2)$  such that  $U(\pi_1, \pi_2) \leq U(\hat{\pi}_1, \hat{\pi}_2)$ ,  $U(\pi_1, \pi_2) \leq U(\lambda\pi_1 + (1-\lambda)\hat{\pi}_1, \lambda\pi_2 + (1-\lambda)\hat{\pi}_2)$  for any  $\lambda \in [0, 1]$ .

For social preferences, quasi-concavity implies that along an interpersonal indifference curve, the higher the first mover's material payoff, the less material payoff the second mover is willing to give up to increase the first mover's material payoff. It also ensures 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, the second mover prefers 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 6** Suppose  $\tilde{\pi}_1(I; p)$  and  $\tilde{\pi}_2(I; p)$ , defined by

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

are finite, real-valued functions.  $U$  is (**weakly**) **locally normal at**  $(I; p)$  if  $\tilde{\pi}_1(I; p)$  and  $\tilde{\pi}_2(I; p)$

are (strictly) increasing in  $I$  at  $(I; p)$ .  $U$  is (**weakly**) **normal** if  $U$  is (weakly) locally normal at  $(I; p)$  for all  $I \in \mathbb{R}$  and  $p > 0$ .

(The functions  $\tilde{\pi}_1(I; p)$  and  $\tilde{\pi}_2(I; p)$  will in fact be well defined given other assumptions on  $U$ .) Becker’s (1974) altruism model explicitly assumes normality, and *all* of the above fairness functional forms—(3), (4), (5), and (6)—also satisfy normality or weak normality.<sup>11</sup> Nonetheless, existing work has not recognized that normality is a strong and central assumption in generating fair-minded behavior. For future reference, if  $U$  is continuously twice-differentiable, call  $N(I; p) \equiv \frac{\partial \tilde{\pi}_1(I; p)}{\partial I}$  the **income effect (on  $\pi_1$ )**, and note that if  $U$  is locally normal at  $(I; p)$ , then  $N(I; p) > 0$ .

A fourth property that will also turn out to be important is conformance to rules of fair behavior. The “equal-split fairness rule,” also called the “50-50 sharing norm,” has been documented in a variety of contexts, such as negotiations, asymmetric joint ventures among corporations, share tenancy in agriculture, and bequests to children (Andreoni & Bernheim 2007). Similarly, in “dictator game” experiments, where one player allocates a given amount of money between himself and another player, 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.<sup>12</sup> Indeed, a kink around equal material payoffs is the feature of fairness models (3), (5), and (6) that allows them to explain the preponderance of equal splits.

Although 50-50 splits are the predominant norm in the laboratory and in a variety of field settings, *unequal* fairness norms come into play in other economic environments (see Cappelen, Hole, Sørensen, & Tungodden 2007). For example, customers who leave money for produce in a cash box on the honor system typically pay the requested amount per unit of produce (rather than matching the amount of payment to their own perceived gain). Financial contracts often apportion profit according to unequal percentages that are standard in the industry. Moreover, even if an individual intends to split surplus evenly, self-serving biases may cause the individual to overestimate his own share, causing him to adhere to a fairness rule that is not equal-split (Babcock & Loewenstein 1997). To allow for these possibilities, I do not require that the kinks occur at equal material payoffs.

---

<sup>11</sup>When actions are bounded, piecewise-linear functional forms like (3) and (6) 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 1) and that the equilibrium is unique (Theorems 3 and 4).

<sup>12</sup>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 (2007) for an alternative model based on signaling, which could be viewed as a microfoundation for kinked indifference curves.

**Definition 7**  $U$  is **fairness-kinked** if it can be expressed as  $U = \min \{U^A, U^B\}$ , where  $U^A, U^B$  are utility functions satisfying:

- There exists some  $(\pi_1, \pi_2)$  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  $U^A(\hat{\pi}_1, \pi_2) < U^B(\hat{\pi}_1, \pi_2)$  for all  $\hat{\pi}_1 > \pi_1$ .
- If  $U^A(\pi_1, \pi_2) \geq U^B(\pi_1, \pi_2)$ , then  $U^A(\pi_1, \hat{\pi}_2) > U^B(\pi_1, \hat{\pi}_2)$  for all  $\hat{\pi}_2 > \pi_2$ .

(Note that if  $U^A$  and  $U^B$  are continuous, joint-monotonic, and quasi-concave, then so is  $U$ .)

**Definition 8** For a fairness-kinked  $U$ , the **fairness rule** is the function  $f(\pi_2)$  that, given a material payoff for the second-mover  $\pi_2$ , 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 utility can be interpreted as social preferences that penalize deviations from the fairness rule. To see this, note that  $U = \min \{U^A, U^B\}$  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. As noted above, functional forms (3), (5), and (6) are fairness-kinked, with the equal-split fairness rule.

The single-crossing properties in the definition of “fairness-kinked” have three useful implications. First, the indifference curves are in fact kinked at fair transactions. Second,  $U = U^A$  in the region of **disadvantageously unfair** transactions for the second mover, where the first mover’s material payoff is higher and the second-mover’s material payoff is lower than dictated by the fairness rule; and  $U = U^B$  in the region of **advantageously unfair** transactions for the second mover. Lastly, 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 utility is locally normal when the agent behaves according to the fairness rule.<sup>13</sup>

Finally, the weakening of monotonicity makes necessary a technical assumption. What matters for behavior is whether the second-mover’s *indifference curves* are kinked or smooth. When  $U$  is monotonic, the indifference curves are kinked if and only if  $U$  is kinked. However, when  $U$  is joint-monotonic, there may be saddle points,  $(\pi_1, \pi_2)$  with  $\frac{\partial U}{\partial \pi_1} = \frac{\partial U}{\partial \pi_2} = 0$ , where the indifference

<sup>13</sup>To be precise, suppose  $U$  is fairness-kinked, and the material payoff pair  $(\tilde{\pi}_1, \tilde{\pi}_2)$  that maximizes  $U$  on budget line  $p\pi_1 + \pi_2 = I$  satisfies:  $U^A = U^B$ ,  $\frac{\partial U^A}{\partial \pi_1} - p \frac{\partial U^A}{\partial \pi_2} < 0$ , and  $\frac{\partial U^B}{\partial \pi_1} - p \frac{\partial U^B}{\partial \pi_2} > 0$  evaluated at  $(\tilde{\pi}_1, \tilde{\pi}_2)$  (so that  $(\tilde{\pi}_1, \tilde{\pi}_2)$  occurs at a kink point). Then  $U$  is locally normal at  $(I; p)$ .

curves can be kinked even though  $U$  is smooth.<sup>14</sup> The following technical assumption rules out such points, ensuring that the indifference curves are kinked if and only if  $U$  is kinked.

**Technical Assumption (TA)** *At any point where  $U$  is differentiable,  $U$  has non-vanishing first derivative: There is no  $(\pi_1, \pi_2)$  such that  $\frac{\partial U}{\partial \pi_1} = \frac{\partial U}{\partial \pi_2} = 0$  at  $(\pi_1, \pi_2)$ .*

## 4 Some Preliminaries

This section presents preliminary observations about efficiency, about the second mover's behavior, and about the first-mover's behavior that will be useful in the subsequent analysis.

### 4.1 Characterizing Utility Pareto Efficient Transactions

Which transactions are utility Pareto efficient? Describing the set of utility Pareto efficient transactions may seem challenging because the second-mover's utility function could be a complicated function of the material payoffs, possibly kinked. Perhaps surprisingly, it turns out that for a very general class of social preferences, there is a straightforward characterization that highlights a tight link between utility Pareto efficiency and material Pareto efficiency.

As is standard, call a transaction  $(a_1, a_2)$  **individually-rational** if both players earn at least their outside option:  $\pi_1(a_1, a_2) \geq 0$  and  $U(\pi_1(a_1, a_2), \pi_2(a_1, a_2)) \geq 0$ . Let

$$(a_1^*, a_2^*) \equiv \arg \max_{\{(a_1, a_2) | \pi_1(a_1, a_2) \geq 0\}} U(\pi_1(a_1, a_2), \pi_2(a_1, a_2))$$

be called the second-mover's **favorite transaction**, his most-preferred transaction among the individually-rational transactions. It is necessarily materially Pareto efficient.

**Theorem 1** *Suppose  $U$  is continuous, joint-monotonic, and quasi-concave. The second-mover's favorite transaction  $(a_1^*, a_2^*)$  exists and is unique. Moreover, a transaction is utility Pareto efficient if and only if it is materially Pareto efficient and satisfies  $\pi_2(a_1, a_2) \leq \pi_2(a_1^*, a_2^*)$ .*

Figure 2 illustrates the relationship between the material Pareto efficiency frontier and the utility Pareto efficiency frontier. A transaction that gives the second mover higher material payoff than

---

<sup>14</sup>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  $\min\{x, y\} = 0$ .

his favorite transaction cannot be utility Pareto efficient because both players would prefer the second-mover's favorite transaction. Material Pareto efficiency is a necessary condition for utility Pareto efficiency because the second-mover's preferences are joint-monotonic. Starting from a materially Pareto-inefficient transaction, there is an alternative transaction that increases both players' material payoffs in a direction that increases the second-mover's utility. In a general equilibrium setting, Dufwenberg, Heidhues, Kirchsteiger, Riedel, & Sobel (2008) independently prove that material Pareto efficiency is a necessary condition for utility Pareto efficiency.

## 4.2 The Second-Mover's Behavior

Given the first-mover's action  $a_1$ , the second mover can be thought of as selecting a pair of material payoffs on the **(material payoff) budget curve**  $B(a_1) = \{(\pi_1(a_1, a_2), \pi_2(a_1, a_2))\}_{a_2 \in \mathbb{R}}$  by his choice of action  $a_2$ . At a point  $(a_1, a_2)$  on the budget curve, the (material payoff) budget *line* that first-order approximates the budget curve is given by  $p\pi_1 + \pi_2 = I$ , where  $p = p(a_2) \equiv -\frac{d\pi_2}{d\pi_1}\Big|_{B(a_1)} = c'(a_2)$  and  $I = I(a_1, a_2) \equiv p(a_2)\pi_1(a_1, a_2) + \pi_2(a_1, a_2)$ . This notation,  $p(a_2)$  and  $I(a_1, a_2)$ , will be useful because local normality is defined with respect to a linear budget set.

Lemma 1 establishes results about the second-mover's behavior that are helpful for backward-inducting the equilibrium, as well as informative about what various properties of social preferences imply for behavior.

**Lemma 1** *Suppose  $U$  is continuous, joint-monotonic, and quasi-concave. Then:*

1. *For any transfer by the first mover,  $a_1$ , the second mover has a unique optimal action,  $a_2(a_1)$ , that is a continuous function of  $a_1$ .*
2. *If  $U$  is (weakly) locally normal at  $(I(\hat{a}_1, a_2(\hat{a}_1)); p(a_2(\hat{a}_1)))$ , then  $a_2(a_1)$  is (weakly) increasing in  $a_1$  at  $\hat{a}_1$ . Hence if  $U$  is (weakly) normal, then  $a_2(a_1)$  is (weakly) increasing in  $a_1$  at all  $\hat{a}_1$ .*
3. *If  $(\hat{a}_1, a_2(\hat{a}_1))$  is an equilibrium, then  $a_2(a_1)$  is strictly increasing in  $a_1$  at  $\hat{a}_1$ .*
4. *If  $U$  is continuously differentiable at some  $(\hat{a}_1, a_2(\hat{a}_1))$  and satisfies (TA), then  $\frac{\partial U}{\partial \pi_1} > 0$  and  $\frac{\partial U}{\partial \pi_2} > 0$  at  $(\hat{a}_1, a_2(\hat{a}_1))$ .*

The second-mover's optimal action is unique because his indifference curves are convex and his budget curve is strictly concave. Existence follows from part 2 of joint-monotonicity and the limit

conditions on  $v(\cdot)$  and  $c(\cdot)$ , which together rule out the possibility that the second-mover might make an unboundedly positive or negative transfer.

A reciprocity motive—roughly speaking, a preference to be more benevolent toward individuals who are more benevolent—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. An influential defense of using models without a built-in reciprocity motive is that they are much simpler to analyze, while nonetheless generating similar behavior (e.g., Fehr & Schmidt 2003). Lemma 1 shows that, in the kind of game considered here, (local) normality of  $U$  is the critical assumption that causes the second mover to behave in a way that looks like reciprocity in models where social preferences are defined only over material payoffs. Holding constant the second-mover’s action, an increase in the first-mover’s transfer locally shifts the budget curve without changing its slope. Normality states that the second mover prefers both players’ material payoffs to increase or decrease together in exactly such a circumstance, which requires the second mover to strictly increase his transfer. In other games (or with non-additively-separable material payoff functions), normality would *not* necessarily lead to reciprocal behavior. Note that if  $a_2(a_1)$  is increasing in  $a_1$ , then the marginal inefficiency,  $v'(a_1) - c'(a_2(a_1))$ , is strictly decreasing in  $a_1$ .

The third part of the lemma observes that even without normality, a higher transfer by the first mover leads to a higher transfer by the second mover at any equilibrium. If it did not, the first mover could profitably deviate by reducing her transfer. (Hence local normality is a sufficient condition for  $a_2(a_1)$  to be increasing, but not a necessary condition.)

While part 2 of the lemma highlights a central role for normality in bilateral exchange, part 4 points out the irrelevance of generalizing monotonicity to joint-monotonicity: if the second-mover’s optimum occurs at a smooth region of his indifference curves, then his social preferences are monotonic on the margin, even if they are not monotonic in general. If the second mover instead preferred to reduce either his own or the first mover’s material payoff on the margin, then his action could not be optimal because he could get higher utility by either increasing or reducing the size of his transfer, respectively. Graphically, since the budget curve is always downward-sloping in the space of material payoffs, a smooth indifference curve must also be downward sloping at a tangency point. Even if the second-mover’s optimum occurs at a kink, the weakening of monotonicity to joint-monotonicity does not really matter because non-monotonicities away from the kink are not relevant for behavior.

Why do non-monotonicities feature so prominently in the evidence that motivates models of

fairness if they are essentially irrelevant in the bilateral exchange game? The key difference between the bilateral exchange game and settings where non-monotocities matter—like the payoff allocation problems described in Section 3—is that in the latter, the “budget set” (the set of material payoff pairs available to the fair-minded player) is upward-sloping in the space of material payoffs. For example, in a prototypical two-option problem, the “unfair” option may give higher material payoffs to both players than the “fair” option does. In that setting, an individual with joint-monotonic utility might well prefer the materially-dominated “fair” option. By contrast, in the bilateral exchange game, the budget curve is downward-sloping. Because the second-mover’s utility is joint-monotonic, rather than reducing both players’ material payoffs away from an “unfair” option, he always prefers to increase the material payoff of the player whose material payoff is low at the same time that he decreases the material payoff of the player whose material payoff is high. Since the budget curve is continuous, there is always a point on the budget curve available to him that gives him higher utility than a materially-dominated outcome would.

### 4.3 The First-Mover’s Behavior

The first part of Lemma 2 states that the second-mover’s favorite transaction is the *only* materially Pareto efficient transaction that is possible for the first mover to induce. If there were two materially Pareto efficient transactions that the first mover could induce, then  $U$  would have two local maxima on the material Pareto efficiency frontier, which is ruled out by quasi-concavity. Because of Lemma 2, I will sometimes refer to the second mover’s favorite transaction as “the” efficient transaction, even though technically there are many other materially/utility Pareto efficient transactions.

**Lemma 2** *Suppose  $U$  is continuous, joint-monotonic, and quasi-concave. Then:*

1. *There exists a unique  $\hat{a}_1$  such that the resulting transaction  $(\hat{a}_1, a_2(\hat{a}_1))$  is materially Pareto efficient. This transaction is the second-mover’s favorite transaction (and so is utility Pareto efficient).*
2. *An equilibrium exists. Moreover, if there is some individually rational, materially efficient transaction,  $(\hat{a}_1, \hat{a}_2)$ , and some material payoff pair on the same interpersonal indifference curve,  $(\tilde{\pi}_1, \tilde{\pi}_2)$ , such that  $-\frac{\tilde{\pi}_2 - \pi_2(\hat{a}_1, \hat{a}_2)}{\tilde{\pi}_1 - \pi_1(\hat{a}_1, \hat{a}_2)} > c'(\hat{a}_2)$ , then there exists an equilibrium in which the players exchange rather than taking their outside options.*

This first part has an immediate corollary: The second-mover’s favorite transaction is the *only* candidate for a materially Pareto efficient equilibrium. In turn, that observation has two important

implications. First, because it is his favorite transaction, the second mover never prefers to deviate. Therefore, a necessary condition for an equilibrium to be materially/utility Pareto efficient is that the *first mover* does not prefer to deviate from the second-mover's favorite transaction. For this reason, subsequent sections will focus on whether the first mover has an incentive to make the transfer that induces the efficient transaction.

Second, even when the equilibrium of the bilateral exchange game is efficient, the first mover would *always* prefer to contract for the second mover's action, as long as writing and enforcing a contract is not too costly. Figure 2 illustrates that if the first mover can make a take-it-or-leave-it offer  $(a_1, a_2)$  to the second mover, the first mover will choose the transaction that maximizes her material payoff subject to ensuring that the second mover earns at least his outside option utility. This transaction occurs at the intersection of the material Pareto-efficiency frontier with the second mover's outside option indifference curve. More generally, if the second mover has bargaining power, the contract will lie somewhere on the material Pareto-efficiency frontier between the second-mover's outside option indifference curve and the second-mover's favorite transaction. At *any* of these transactions, the first mover earns a higher material payoff than she does at the second-mover's favorite transaction.

The second part of Lemma 2 addresses existence of equilibrium. Since the set of individually rational transactions is compact, an equilibrium always exists. However, this equilibrium may involve the first mover choosing her outside option if the second mover is so selfish that his transfer will be very small (or negative), regardless of the first mover's action. Lemma 2 gives a sufficient condition for trade to occur in equilibrium. Roughly, the condition is that the second mover is altruistic enough (his interpersonal indifference curve is flat enough) at some individually rational, materially efficient transaction.

## 5 Efficiency of Equilibrium

This section addresses the central question of the paper: How efficient is the equilibrium of the bilateral exchange game?

### 5.1 The General Case

If  $U$  is smooth and  $c(a_2)$  is convex, the equilibrium cannot be materially Pareto efficient.

**Theorem 2** *Suppose  $U$  is joint-monotonic, quasi-concave, continuously twice-differentiable, and satisfies (TA), and suppose  $c'' > 0$ . Then no equilibrium is materially Pareto efficient. Furthermore,*

at any interior equilibrium  $(a_1, a_2(a_1))$ , the marginal inefficiency is:

$$v'(a_1) - c'(a_2) = \frac{1}{N(I(a_1, a_2); p(a_2))} \frac{c''(a_2)}{\frac{d^2\pi_2}{d(\pi_1)^2} \Big|_{U=U(a_1, a_2)}} \neq 0. \quad (7)$$

At an interior equilibrium, the magnitude of marginal inefficiency is decreasing in the magnitude of the income effect (hereafter abbreviated as  $N(a_1, a_2)$ ), increasing in the convexity of the cost of the second mover's transfer, and decreasing in the convexity of the interpersonal indifference curve. Since  $c'' > 0$  and  $\frac{d^2\pi_2}{d(\pi_1)^2} \Big|_{U=U(a_1, a_2)} > 0$ , the sign of the income effect determines the sign of the marginal inefficiency.

To understand why no equilibrium is efficient, first consider an equilibrium where the second mover gets exactly his outside option utility  $U = 0$ . Since  $v'(0) > c'(0)$  and  $U$  is joint-monotonic, the second mover's favorite transaction gives him strictly higher utility. Since the equilibrium transaction is not the second mover's favorite, it cannot be materially Pareto efficient.

Now consider an interior equilibrium. The first mover's incentive on the margin can be decomposed into two effects:

$$\begin{aligned} \frac{d\pi_1(a_1, a_2(a_1))}{da_1} &= \underbrace{(v'(a_1) - c'(a_2)) N(a_1, a_2) \left( \frac{\frac{d^2\pi_2}{d(\pi_1)^2} \Big|_{U=U(a_1, a_2)}}{\frac{d^2\pi_2}{d(\pi_1)^2} \Big|_{U=U(a_1, a_2)} + c''(a_2)} \right)}_{\text{size-of-the-pie effect}} \quad (8) \\ &+ \underbrace{\left( \frac{-c''(a_2)}{\frac{d^2\pi_2}{d(\pi_1)^2} \Big|_{U=U(a_1, a_2)} + c''(a_2)} \right)}_{\text{share-of-the-pie effect}} \end{aligned}$$

The “size-of-the-pie effect” is the effect on the first mover's material payoff from a change in the gains from trade. If the change in the players' actions increases efficiency, and if the second mover's utility is locally normal, then the size-of-the-pie effect is positive. The “share-of-the-pie effect” captures how the second mover optimally reallocates the change in gains from trade (rather than giving all of it to the first mover) as a result of the change in the rate of tradeoff between the players' material payoffs. While the size-of-the-pie effect could in general have either sign, the share-of-the-pie effect is always negative. An increase in  $a_2$  raises the “price” of a unit of the first mover's material payoff relative to a unit of the second mover's since  $c(a_2)$  is convex. Because  $U$  is quasi-concave and smooth, this makes the second mover want to reduce the first mover's material payoff on the margin and increase the second mover's.

The materially Pareto-efficient transaction  $(a_1^*, a_2^*)$  is *not* be an equilibrium. At  $(a_1^*, a_2^*)$ , the marginal inefficiency,  $v'(a_1^*) - c'(a_2^*)$ , equals zero, so the size-of-the-pie effect is second order. The

share-of-the-pie effect dominates, and the first mover can profitably deviate by making a smaller transfer.

At an equilibrium, the two effects must exactly offset each other. Equation (7) follows from setting (8) equal to zero. The equilibrium has inefficiently low transfers if and only if the utility function is locally normal; and inefficiently high transfers if and only if it is locally inferior. ( $N(a_1, a_2)$  cannot equal zero at a local equilibrium because the first mover's first-order condition would be violated.) Figure 3 illustrates how material payoffs vary with the first mover's action and how the equilibrium occurs at a materially Pareto-inefficient transaction.

In the discrete payoff allocation problems that are often studied, a fair-minded second-mover is more likely to choose an “efficient but unfair” outcome over an “inefficient but fair” outcome the less materially costly it is to help the first mover and the more weight he puts on the first-mover's material payoff. It is notable that here, it is the *second derivatives* of  $c(a_2)$  and of the interpersonal indifference curve that determines the magnitude of marginal inefficiency, rather than the slopes *per se*. The reason is that, with a continuous budget curve, the slopes are taken into account in how the second mover chooses to *distribute* a given amount of surplus; loosely speaking, when  $c'$  is small or  $d\pi_1/d\pi_2|_U$  is large in magnitude, the second mover is willing to transfer more for any given level of transfer by the first mover. However, the degree of efficiency (as measured by the gains from marginally increasing the transfers) depends on how these slopes change, rather than on their levels.

A paradoxical implication is that if the second mover cares exclusively about social efficiency—that is, if utility is a weighted sum of the material payoffs,  $U = (1 - \lambda)\pi_1 + \lambda\pi_2$  for some  $0 < \lambda < 1$ —then the equilibrium will be maximally *inefficient* (as measured by marginal inefficiency). In that case, the second mover's transfer is independent of the first mover's, so the first mover will choose the lowest transfer that makes the second mover indifferent between trading and taking his outside option. The equilibrium will be more efficient to the extent that the second-mover's indifference curves are convex, even if there are many transactions where the second mover has a preference for reducing one of the players' material payoffs.

Equation (8) implies that there are two cases in which the share-of-the-pie effect vanishes at an efficient transaction: (1) both players' material payoff functions are linear in the second commodity, in which case the numerator of the share-of-the-pie term is zero, and (2) the second-mover's utility function is kinked at the equilibrium, in which case the denominator is infinite. Hence at least one of these two conditions is necessary for the equilibrium to be fully efficient. The next two subsections explore these cases in more detail.

## 5.2 Efficient Case I: Material Payoffs Are Linear in the Second Action

This subsection, and this subsection only, replaces the assumption that  $c'' > 0$  with the assumption that  $c'' = 0$ . In this case, where the players' material payoff functions are linear in the second-mover's action, the second mover faces the same rate of tradeoff between the players' material payoffs, regardless of either player's action. The share-of-the-pie effect in (8) is zero because the first mover cannot influence the relative price of the players' material payoffs with his action. As a result, the first mover does not face a tradeoff between maximizing the size of the pie and creating a more favorable relative price. That makes it possible for the equilibrium to be fully efficient.

In fact, linear material payoff functions in the second-mover's action and the normality of  $U$ , taken together, are sufficient to ensure that the equilibrium is efficient—and also unique.

**Theorem 3** *Suppose  $U$  is continuous, joint-monotonic, quasi-concave, and normal. If there is an equilibrium in which the players trade, and if  $c(a_2) = \kappa_1 + \kappa_2 a_2$  for  $\kappa_2 > 0$ , then the unique equilibrium transaction is the second-mover's favorite transaction (and so is utility Pareto efficient).*

Because  $U$  is normal, and because the rate of tradeoff between material payoffs is independent of the size of the pie, the second mover will choose his action such that both players' material payoffs are increasing in the total surplus to be divided. Hence, the first mover maximizes her own material payoff by choosing the unique action that leads to a materially Pareto-efficient outcome. Figure 4 illustrates Theorem 3.

In a variety of economic applications, it is common to assume that  $c(a_2)$  is linear in  $a_2$ . That is considered a reasonable model when the first mover is providing a good or service that has diminishing marginal utility or increasing marginal cost of provision, and the second mover is paying money in exchange (because the marginal utility of money declines much more slowly than the marginal utility for any particular good). Hence the theorem applies to commercial transactions where the exact payment is not specified in advance, such as the discretionary component of payment to a contractor (e.g., for housework). Theorem 3 says that when monetary transactions have a discretionary component, and the second mover is an agent with social preferences, the transactions will be efficient. As long as the specified assumptions on  $U$  are satisfied, the conclusion does not depend on how selfish or altruistic the second-mover is, or whether  $U$  is kinked or smooth; the first mover will choose the same action in any case, since there is a unique efficient action (satisfying  $v'(a_1^*) = \kappa_2$ ) when  $c(a_2) = \kappa_1 + \kappa_2 a_2$ . For that reason, even though I have assumed that  $U$  is known to the first mover, the first mover would choose the efficient action even if she were

uncertain about the second-mover’s social preferences (and hence uncertain about the action the second mover will choose). Theorem 3 may explain the ubiquity of commercial transactions where the monetary payment is (at least partly) discretionary.

Theorem 3 generalizes the Rotten Kid theorem (Becker 1974), a well-known result in the economics of the family. In the classic Rotten Kid setup, a child can take an action that increases family income but reduces his own personal income. Then the parent transfers some amount of family income to the child. The Rotten Kid theorem says that if the parent has altruistic (i.e., monotonic) social preferences, and if both players’ material payoffs enter the parent’s utility as normal goods, then even a purely self-regarding child (“rotten kid”) will act so as to maximize family income.

Relative to the Rotten Kid theorem, Theorem 3 relaxes the assumption that utility is monotonic, allowing instead for joint-monotonicity. It follows fairly directly from Lemma 1 that this relaxation does not matter for equilibrium behavior. Although only a slight mathematical generalization, the relaxation of monotonicity is an important economic generalization because it means that the surprising efficiency conclusion applies for a large and empirically-relevant class of (non-monotonic) fairness preferences. Traditionally, the Rotten Kid theorem has been interpreted as applying to family environments but *not* market environments because altruism is assumed to be relevant primarily within the family (Becker 1974). By contrast, a concern for fairness appears to be widespread in market settings (e.g., Kahneman, Knetsch, & Thaler 1986; Bewley 1999). Theorem 3 therefore suggests that the Rotten Kid logic may be applicable in a far wider range of settings, including market environments.

On the other hand, the analysis in this paper also clarifies the limitations of Rotten-Kid-type results. Bergstrom (1989) proved that linearity of the material payoff functions in the second-mover’s action is a sufficient condition for the Rotten Kid theorem, and he conjectured that it is also a necessary condition. Theorem 2 proves that it is in fact a necessary condition.<sup>15</sup> Moreover, although existing literature interprets altruism as central and does not emphasize normality, the analysis here makes clear that normality is crucial while altruism can be weakened.

### 5.3 Efficient Case II: Second-Mover’s Social Preferences Are Fairness-Kinked

Returning to the assumption that  $c(a_2)$  is strictly convex, this subsection considers the case where the second-mover’s indifference curve is kinked at his favorite transaction. In this case, the second-

---

<sup>15</sup>The game I analyze is slightly less general than Bergstrom’s (1989) in that it has only two players with material payoff functions that are additively separable, but is similar in essential respects.

mover's optimal action is insensitive to a marginal change in the rate of tradeoff between the players' material payoffs. The share-of-the-pie effect in (8) is zero here because—even though the first mover *can* influence the relative price of the players' material payoffs with his action, unlike in the case of linear material payoff functions—changing the relative price of the players' material payoffs will not affect the second-mover's action. As in the case of linear material payoff functions, the fact that the share-of-the-pie effect vanishes makes it possible for the equilibrium to be fully efficient.

In fact, if the second-mover's utility is very slightly kinked at his favorite transaction, then the action that induces the second-mover's favorite transaction will be a *local* optimum for the first mover (though possibly not a global optimum).

**Lemma 3** *Suppose  $U$  is fairness-kinked, with  $U^A$  and  $U^B$  being joint-monotonic, quasi-concave, continuously twice-differentiable, and satisfying (TA). Let  $(a_1^*, a_2^*)$  denote the second-mover's favorite transaction. If the utility function satisfies*

$$U^A = U^B \tag{9}$$

$$\frac{\partial U^A}{\partial \pi_1} - c'(a_2^*) \frac{\partial U^A}{\partial \pi_2} < 0 \tag{10}$$

$$\frac{\partial U^B}{\partial \pi_1} - c'(a_2^*) \frac{\partial U^B}{\partial \pi_2} > 0 \tag{11}$$

at  $(a_1^*, a_2^*)$ , then the second-mover's optimal strategy  $a_2(a_1)$  satisfies the fairness rule for all  $a_1$  in a neighborhood of  $a_1^*$ , and  $a_1^*$  is a locally optimal action for the first mover.

Equality (9) simply says that the second-mover's favorite transaction occurs at a kink (and therefore, on the fairness rule). At the favorite transaction, the weak inequalities

$$\frac{\partial U^A(\pi_1(a_1^*, a_2^*), \pi_2(a_1^*, a_2^*))}{\partial a_2} = \frac{\partial U^A}{\partial \pi_1} - c'(a_2^*) \frac{\partial U^A}{\partial \pi_2} \leq 0 \tag{12}$$

$$\frac{\partial U^B(\pi_1(a_1^*, a_2^*), \pi_2(a_1^*, a_2^*))}{\partial a_2} = \frac{\partial U^B}{\partial \pi_1} - c'(a_2^*) \frac{\partial U^B}{\partial \pi_2} \geq 0 \tag{13}$$

necessarily hold, ensuring that the second-mover's action is optimal. If these inequalities hold strictly—that is, if (10) and (11) are satisfied—then the second-mover's optima continue to occur at kinks for small changes in the first-mover's action. As a result, the second mover obeys the fairness rule locally, and so inducing the second-mover's favorite transaction is locally optimal for the first mover.

Intuitively, the fairness rule is characterized by the players' material payoffs increasing or decreasing in tandem. As long as the second-mover's favorite transaction occurs at a kink, the second

mover responds to small changes in the first-mover's action by adjusting his own transfer to ensure that the fairness rule remains satisfied. This behavior means that the first-mover's material payoff is increasing in the size of the pie, which gives the first mover an incentive to maximize the size of the pie.

However, fairness-kinked social preferences alone do not ensure that the equilibrium occurs at the efficient transaction. Theorem 4 provides sufficient conditions: the second-mover's social preferences are normal and *sufficiently* fairness-kinked at the efficient transaction. These conditions also imply that the equilibrium is unique.

**Theorem 4** *Suppose  $U$  is fairness-kinked, with  $U^A$  and  $U^B$  being joint-monotonic, quasi-concave, normal, continuously twice-differentiable, and satisfying (TA). Let  $(a_1^*, a_2^*)$  denote the second-mover's favorite transaction, and let  $(\hat{a}_1, \hat{a}_2)$  denote the (necessarily unique) transaction with  $\hat{a}_1 < a_1^*$  such that  $\pi_1(\hat{a}_1, \hat{a}_2) = \pi_1(a_1^*, a_2^*)$  and  $U(\hat{a}_1, \hat{a}_2) = 0$ . If there is an equilibrium in which the players trade, and if the utility function satisfies*

$$U^A = U^B \tag{14}$$

$$\frac{\partial U^A}{\partial \pi_1} - c'(\hat{a}_2) \frac{\partial U^A}{\partial \pi_2} < 0 \tag{15}$$

$$\frac{\partial U^B}{\partial \pi_1} - c'(a_2^*) \frac{\partial U^B}{\partial \pi_2} > 0 \tag{16}$$

at  $(a_1^*, a_2^*)$ , then the second-mover's optimal strategy  $a_2(a_1)$  satisfies the fairness rule for all  $a_1$  in a neighborhood of  $a_1^*$ , and  $(a_1^*, a_2^*)$  is the unique equilibrium transaction (and so is utility Pareto efficient).

Inequality (16) is the same as (11),<sup>16</sup> but inequality (15) is more stringent than (11), requiring that the second mover put *sufficiently* low relative weight on the first-mover's material payoff at disadvantageously unfair transactions that are near the efficient transaction. Combined with normality, this condition prevents the second mover from being willing (due to altruism) to make a large transfer when the first mover makes a small transfer. Loosely put, normality allows the local aversion to disadvantageous unfairness from (15) to imply a sufficient aversion to disadvantageous unfairness on the outside option indifference curve. Without (15) and normality, even if inducing the efficient transaction were locally optimal for the first mover, the first mover might earn a still greater material payoff by making a much smaller transfer. Figure 5 illustrates the efficient

---

<sup>16</sup>The conclusion of Theorem 4 that  $(a_1^*, a_2^*)$  is the unique equilibrium transaction actually requires only the weak inequality (13), rather than the strict inequality (16). Moreover, with the weak inequality, the second-mover's optimal strategy  $a_2(a_1)$  still satisfies the fairness rule for all  $a_1 < a_1^*$  sufficiently close to  $a_1^*$ . However, with weak inequality (13), the second mover may not obey the fairness rule when  $a_1$  is marginally larger than  $a_1^*$ .

equilibrium with fairness-kinked social preferences, as well as how the equilibrium could fail to be efficient if the assumptions of Theorem 4 are not satisfied.

Roughly speaking, these results state that if the second mover is behaving according to a fairness rule, then trade will be efficient. Unlike Theorem 3, Lemma 3 and Theorem 4 put no restrictions on the material payoff functions, and so apply to more kinds of exchange, including in-kind trades, such as barter or exchange of favors. Moreover, as long as the second mover 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.

However, the theorem requires that the first mover know what fairness rule the second mover is following. Otherwise, the first mover would not know which action would induce an efficient transaction. The fact that efficiency depends on both players knowing the fairness rule suggests a role for customs or laws that make the “fair” rate of exchange between commodities common knowledge. It also might explain why social norms like 50-50 sharing have developed, and are considered to apply when there is no norm that is more specific to the situation. Indeed, even if the first mover is uncertain about what fairness rule the second mover is following, as long as a sufficiently large proportion of second movers follow the same rule, it will be optimal for the first mover to make the transfer that would lead to an efficient outcome given that rule.

## 6 Robustness: Both Players Have Social Preferences

To keep things as simple as possible, I have assumed throughout that only the second-mover’s utility,  $U_2(\pi_1, \pi_2)$  (now subscripted for clarity), depends on both players’ material payoffs. Here I discuss the extent to which the earlier results generalize when both players have social preferences.

If the first-mover’s utility function  $U_1(\pi_1, \pi_2)$  is monotonic, then all of the main points from earlier sections generalize. In this case, a transaction is utility Pareto efficient if and only if it lies on the material Pareto-efficiency frontier between the second-mover’s favorite transaction and the first-mover’s favorite transaction. If both players’ material payoff functions are linear in the second-mover’s action, or if the second-mover’s social preferences are sufficiently-kinked, then the equilibrium occurs at the second-mover’s favorite transaction (and so is utility Pareto efficient) as long as  $U_2$  is normal. These full efficiency results only require that the first-mover’s social preferences be continuous and monotonic because the first mover maximizes *both* players’ material payoffs by inducing the efficient transaction. If the first mover has fairness-kinked utility with a different fairness rule than the second mover’s, it is the second-mover’s fairness rule that determines

the equilibrium.

New complications arise if the first-mover's utility is merely joint-monotonic rather than monotonic. In that case, there may be non-materially-Pareto-efficient transactions that both players consider "unfair" to themselves in the sense that both prefer to reduce the other player's material payoff. These transactions are utility Pareto efficient, despite being materially Pareto inefficient. However, such transactions cannot occur in equilibrium because at least one player can profitably deviate. In the extreme case where at least one player wants to reduce the other player's material payoff at *every* transaction, then no trade will occur in equilibrium—yet this no-trade outcome will actually be utility Pareto efficient.

At an interior local equilibrium  $(a_1, a_2(a_1))$ , if  $U_1$  is continuously differentiable and  $U_2$  is continuously twice-differentiable, then the formula for marginal (material) inefficiency is

$$v'(a_1) - c'(a_2) = \frac{1}{M(I(a_1, a_2); p(a_2))} \frac{c''(a_2)}{\frac{d^2\pi_2}{d(\pi_1)^2} \Big|_{U_2=U_2(\pi_1, \pi_2)}} \left( -\frac{\partial U_1(a_1, a_2)}{\partial a_1} \right), \quad (17)$$

where  $M(I; p) \equiv \frac{\partial U_1(\tilde{\pi}_1(I; p), \tilde{\pi}_2(I; p))}{\partial I}$  is the income effect on  $U_1$ , and  $-\frac{\partial U_1(a_1, a_2)}{\partial a_1}$  is the effect on  $U_1$  from marginally reducing  $a_1$ , *holding*  $a_2$  constant.

Naturally, equation (17) specializes to (7) when the first mover is purely self-regarding ( $U_1 \equiv \pi_1$ , so  $-\frac{\partial U_1(a_1, a_2)}{\partial a_1} = -\frac{\partial \pi_1(a_1, a_2)}{\partial a_1} = 1$ ). However, the first-mover's social preferences introduce a third possible way to get full efficiency: both players share the same favorite transaction ( $-\frac{\partial U_1(a_1, a_2)}{\partial a_1} = 0$ ), in which case that transaction is the equilibrium and is utility Pareto efficient. Relative to when the first mover is purely self-regarding, the equilibrium could be *less* efficient if either the first mover considers her material-payoff-maximizing transaction to be "unfair" to herself ( $-\frac{\partial U_1(a_1, a_2)}{\partial a_1} > 1$ ), or if the first mover is so altruistic that her favorite transaction gives the second mover a higher material payoff than the second mover's own favorite transaction does ( $-\frac{\partial U_1(a_1, a_2)}{\partial a_1} < 0$ ). However, in the most realistic case where the first mover is altruistic on the margin, but prefers to give herself a higher material payoff than the second mover does ( $0 < -\frac{\partial U_1(a_1, a_2)}{\partial a_1} < 1$ ), equation (17) implies that the equilibrium will be *more* efficient than it would be in the case where the first mover is purely self-regarding.

## 7 Discussion and Concluding Remarks

When contracting is costly, social preferences may nonetheless enable relatively efficient exchange. The results in this paper predict that exchange will be efficient in two cases: (1) when the party

who is trusted to reciprocate is paying money to the other party, or (2) when he behaves according to a rule of fair behavior.

While the analysis in this paper directly applies to settings where the second-mover's action is entirely discretionary, it is relevant for understanding a range of other contracting and organizational issues, such as the hold-up problem. In the classic hold-up problem, a first mover makes an irreversible, relationship-specific investment and then Nash-bargains with the relationship partner over how to split the surplus. Because the first-mover's investment is sunk, she is forced to share the *gross* returns (rather than the returns net of the investment) with the partner. Anticipating this, it may not be profitable for the first mover to make the investment at all, even if making the investment is socially efficient. The bilateral exchange game analyzed in this paper does not directly apply to the hold-up problem because the bilateral exchange game is the special case where the second-mover has all of the bargaining power in the Nash bargaining. That extreme case is precisely when the hold-up problem is most severe. Yet that is the case where Theorem 3 delivers a generic efficiency result for a very broad class of social preferences when the players' material payoffs are assumed to be linear in the second mover's transfer (the standard assumption in hold-up settings). Theorem 4 applies even without quasi-linear material payoffs when the second-mover behaves according to a rule for allocating surplus fairly. Hence the results here suggest that social preferences provide a full solution to the hold-up problem exactly when other potential solutions do not apply.<sup>17</sup>

A key assumption about social preferences that enables them to resolve the hold-up problem is that the second-mover cares about the players' material payoffs (i.e., returns) *net* of the first-mover's action (i.e., investment). Viewed in this light, a crucial feature of social preferences is that they do not ignore sunk costs. However, the extent to which judgments about what is fair ignore sunk costs is likely sensitive to how the situation is framed. This suggests that in settings where social preferences play an important role in exchange, strategic framing of the situation may also become important.

Besides incorporating strategic framing, there are three ways to extend the analysis that are particularly urgent for improving the applicability of the theory. Carefully implementing any of these extensions will require further elaborating the theory of social preferences. First, I have assumed throughout that the players' preferences are common knowledge, but uncertainty about

---

<sup>17</sup>Interestingly, I believe that the equilibrium will *not* be efficient in the less severe hold-up problem where the first-mover has some bargaining power because the second-mover will anticipate Nash bargaining when choosing his action. A careful analysis of this case requires an expected utility structure on the social preferences in order to operationalize the Nash bargaining.

both the material payoff functions and the social preferences is more realistic (see, e.g., Fehr, Klein, & Schmidt 2007). In some cases, uncertainty about social preferences will make little difference—such as when the first mover 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 the first mover to take the action that induces a 50-50-rule-following second mover to respond efficiently. However, in many cases, it seems likely that uncertainty will reduce efficiency. For example, even if all second-mover types obey a fairness rule, if each obeys a *different* fairness rule, then the first-mover’s uncertainty will make it optimal for him to play against a “representative second-mover” who has smooth preferences, ensuring that the equilibrium is generically inefficient. A careful analysis of the uncertainty case will require a theory of social preferences under uncertainty, which raises challenges I have not confronted in this paper.

Second, I have assumed that social preferences depend only on material payoffs, but there is substantial evidence that social preferences also include a reciprocity motive (see, e.g., Rabin 1993 and Charness & Rabin 2002). In terms of generating reciprocal behavior, a direct reciprocity motive will qualitatively serve much the same role as the assumption of normality in the present paper. However, in the psychological games apparatus, it is not clear how to formulate the question of whether the equilibrium exchange is efficient because preferences themselves depend on beliefs about the other player. Nonetheless, since models of reciprocity require an underlying material-payoff-based theory of what is considered a fair allocation, the analysis in the present paper may help provide a foundation for analyses that incorporate intentions.

Finally, although this paper has focused on bilateral exchange, there are important real-world settings with many players, such as a multiple-worker firm. Although it is important to analyze how a concern for fairness affects transactions in such environments, the conclusions are sensitive to how the bilateral model is generalized. The simplest extension of the model to a situation with multiple workers hired by a firm is to assume that each worker’s social preferences apply to that worker’s bilateral transaction with the firm. In that case, 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, the fact that real-world workers compare their own wage and effort to those of other workers in the firm (e.g., Bewley 1999) suggests that the bilateral model misses important aspects of reality. It is not difficult to write down a game with an employer and two workers, along with social preferences for the workers, such that the workers sabotage each other in order to come out ahead, leading to an equilibrium that is materially and

Pareto inefficient. Unfortunately, at present there is very little evidence to suggest which of many possible generalizations of social preferences is most appropriate.

## 8 Proofs

We begin with a technical lemma before proving the results in the text. Throughout, for notational compactness, let  $\pi \equiv (\pi_1, \pi_2)$  denote the vector of material payoffs.

**Technical Lemma** *Suppose  $U$  is continuous, joint-monotonic, and quasi-concave. Then:*

1. *The set of individually-rational transactions*

$$T \equiv \{(a_1, a_2) \mid U(\pi(a_1, a_2)) \geq 0, \pi_1(a_1, a_2) \geq 0\}$$

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

2. *Along any budget curve  $B(a_1) = \{(\pi_1(a_1, a_2), \pi_2(a_1, a_2))\}_{a_2 \in \mathbb{R}}$ , the second-mover's utility has a unique maximum and strictly decreases as  $a_2$  moves away from this maximum.*
3. *Along the material-efficiency frontier, the second-mover's utility has a unique maximum and strictly decreases as  $a_2$  moves away from this maximum.*

**Proof of part 1:** The transaction  $(a_1, a_2) = (0, 0)$  gives material payoffs  $\pi(0, 0) = (0, 0)$  and utility  $U(\pi(0, 0)) = 0$ , so both sets are non-empty.

By joint-monotonicity (part 2), there is some  $\pi'_2 < 0$  with  $U(0, \pi'_2) < 0$ . Quasi-concavity implies that  $U(0, \pi_2) \leq U(0, \pi'_2) < 0$  for all  $\pi_2 \leq \pi'_2$ . By joint-monotonicity (part 2) and continuity, there is some  $\pi'_1 < 0$  with  $U(0, \pi'_2) < U(\pi'_1, 0) < 0$ . Quasi-concavity then implies that any point  $(\pi_1, \pi_2)$  with  $\pi_1 \geq 0$  that lies to the left of the line connecting  $(0, \pi'_2)$  to  $(\pi'_1, 0)$  will have  $U(\pi_1, \pi_2) < 0$ . Therefore, every point of  $T$  will lie to the right of this line.

The line is downward sloping, so it may be written in the form  $b\pi_1 + d\pi_2 = k$  for some constants  $b, d, k$  with  $b, d > 0$ . So every point  $(a_1, a_2) \in T$  satisfies  $b\pi_1(a_1, a_2) + d\pi_2(a_1, a_2) > k$ . Rewrite this left-hand side as  $(dv(a_1) - ba_1) + (ba_2 - dc(a_2))$ . Now,  $ba_2 - dc(a_2)$  is concave and goes to  $-\infty$  as  $a_2 \rightarrow \infty$ , so it is either decreasing everywhere or has some maximum value  $M$  at a point  $a_2^*$ .

In the first case,  $a_2 \geq a_1$  (which is true for all points in  $T$  because  $\pi_1 \geq 0$ ) implies  $ba_2 - dc(a_2) \leq ba_1 - dc(a_1)$ , hence  $k < b\pi_1(a_1, a_2) + d\pi_2(a_1, a_2) \leq d(v(a_1) - c(a_1))$  for all  $(a_1, a_2) \in T$ . The right-hand side of this goes to  $-\infty$  as  $a_1 \rightarrow \pm\infty$ , since  $\lim_{a_1 \rightarrow -\infty} v'(a_1) = \infty$  and  $\lim_{a_2 \rightarrow \infty} c'(a_2) = \infty$ . Hence,  $a_1$  must be bounded for all  $(a_1, a_2) \in T$ . Now this implies that  $av(a_1) - ba_1$  is also bounded above in  $T$ , which in turn implies that  $ba_2 - ac(a_2)$  must be bounded below in  $T$ , so  $a_2$  is bounded above in  $T$ . And  $a_2 \geq a_1$  shows that  $a_2$  is bounded below in  $T$ .

In the second case, where  $ba_2 - dc(a_2)$  has a maximum  $a_2^*$ , we can apply the same argument to show that  $a_1$  is bounded above in  $T$  because for  $a_1 > a_2^*$ ,  $ba_2 - dc(a_2)$  is decreasing (since  $a_2 \geq a_1$ ). Moreover,  $a_1$  must be bounded below in  $T$  because  $dv(a_1) - ba_1 > k - (ba_2 - dc(a_2)) \geq k - M$ . Then we proceed as in the previous paragraph to show that  $a_2$  is bounded in  $T$ .

So in either case, we have shown that both  $a_1$  and  $a_2$  are bounded for all  $(a_1, a_2) \in T$ , as required. Hence  $T$  is bounded. Since  $T$  is clearly closed and  $U$  is continuous, the set  $T$  is in fact compact, as is  $T_\pi$ .

**Proof of part 2:** We first show that for each fixed  $a_1$  and real number  $k$ , the set of second-mover actions  $\{a_2 \mid U(\pi(a_1, a_2)) \geq k\}$  is an interval (possibly unbounded). Let  $a'_2 < a''_2$  be two values in this set. By construction,  $U \geq k$  at  $(x_1, y_1) = \pi(a_1, a'_2)$  and  $(x_2, y_2) = \pi(a_1, a''_2)$ . It follows that  $U \geq k$  at  $(x_1, y_2)$ . (To see this, let  $\bar{y} = \max\{y \in [y_1, y_2] \mid U(x_1, y) \geq k\}$  (the maximum exists by continuity). If  $\bar{y} = y_2$  then we are done, so assume  $\bar{y} < y_2$ . By joint-monotonicity, we can choose  $\hat{x}, \hat{y}$  with  $x_1 < \hat{x}$  and  $\bar{y} < \hat{y} < y_2$  so that  $U(\hat{x}, \hat{y}) > U(x_1, \bar{y}) \geq k$ . The line segment connecting  $(x_2, y_2)$  and  $(\hat{x}, \hat{y})$  meets the line  $x = x_1$  at a point with some  $y$ -coordinate strictly between  $\bar{y}$  and  $y_2$ . By quasi-concavity, the value of  $U$  at this point is  $\geq k$ . This contradicts the maximality of  $\bar{y}$ .) Now, for any  $a'_2 < a_2 < a''_2$ , the point  $\pi(a_1, a_2)$  lies strictly inside the triangle defined by these three points (by the convexity of  $c(a_2)$ ). Since  $U$  is quasi-concave,  $U(\pi(a_1, a_2)) \geq k$  also.

This shows that there cannot be three values  $a'_2 < a_2 < a''_2$  with  $U(\pi(a_1, a'_2)) > U(\pi(a_1, a_2)) < U(\pi(a_1, a''_2))$ . It follows that  $U$  is either weakly monotonic everywhere, or weakly increasing on  $(a_2^*, -\infty)$  and weakly decreasing on  $(a_2^*, \infty)$  for some  $a_2^*$ .

We now show that  $U$  cannot be constant on any interval along the budget curve. Suppose  $U$  assumes the constant value  $k$  on the interval  $[a'_2, a''_2]$ . Quasi-concavity implies that  $U$  is  $\geq k$  at the point  $(x_0, y_0) = (\pi(a_1, a'_2) + \pi(a_1, a''_2))/2$ . For sufficiently small  $\epsilon > 0$ , the box  $[x_0, x_0 + \epsilon] \times [y_0, y_0 + \epsilon]$  lies entirely below and to the left of the curve  $C = \{\pi(a_1, a_2) \mid a'_2 < a_2 < a''_2\}$ . Joint-monotonicity ensures that  $U$  assumes a value  $k' > k$  at some point  $(x', y')$  inside this box. Now, let  $S = \{(x, y) \mid x \geq x', y \geq y', U(x, y) \geq U(x', y')\}$ . We know that  $S$  does not intersect  $C$  because  $U \geq k'$  on  $S$ , whereas  $U$  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  $x'' > x', y'' > y'$  with  $U(x'', y'') > U(x', y') \geq k'$ , contradicting maximality. It follows that  $U$  cannot be constant on  $[a'_2, a''_2]$  after all.

We complete the proof by ruling out that  $U$  is monotonic along the budget curve; in particular, we show that for any  $(a_1, a_2)$ , there are  $a'_2 < a_2 < a''_2$  such that  $U(\pi(a_1, a'_2)) < U(\pi(a_1, a_2)) >$

$U(\pi(a_1, a_2''))$ . Let  $(\hat{x}_0, \hat{y}_0) = \pi(a_1, a_2)$  and  $U_0 = U(\pi(a_1, a_2))$ . By joint-monotonicity (part 2), there exist  $\delta_x, \delta_y > 0$  such that for  $\hat{x}_1 = \hat{x}_0 - \delta_x$  and  $\hat{y}_1 = \hat{y}_0 - \delta_y$ ,  $U(\hat{x}_0, \hat{y}_1) < U_0$  and  $U(\hat{x}_1, \hat{y}_0) < U_0$ . Since  $U$  is continuous and quasi-concave, we can assume that  $U(\hat{x}_0, \hat{y}_1) = U(\hat{x}_1, \hat{y}_0)$ ,  $U(\hat{x}_0, \hat{y}_1 - \varepsilon) < U(\hat{x}_0, \hat{y}_1)$ , and  $U(\hat{x}_1 - \varepsilon, \hat{y}_0) < U(\hat{x}_1, \hat{y}_0)$  for  $\varepsilon > 0$  sufficiently small. Because the slope of the budget curve is asymptotes to  $-\infty$  and  $0$  as  $a_2 \rightarrow -\infty$  and  $a_2 \rightarrow +\infty$ , respectively, the line through  $(\hat{x}_0, \hat{y}_1)$  and  $(\hat{x}_1 - \varepsilon, \hat{y}_0)$  intersects the budget curve at some point  $(\hat{x}_2, \hat{y}_2)$  with  $\hat{x}_2 < \hat{x}_0$  and  $\hat{y}_2 > \hat{y}_0$ . Since  $U$  is quasi-concave,  $U(\hat{x}_2, \hat{y}_2) < U_0$ . Similarly, the line through  $(\hat{x}_1, \hat{y}_0)$  and  $(\hat{x}_0, \hat{y}_1 - \varepsilon)$  intersects the budget curve at some point  $(\hat{x}_3, \hat{y}_3)$  with  $\hat{x}_3 > \hat{x}_0, \hat{y}_3 < \hat{y}_0$ , and  $U(\hat{x}_3, \hat{y}_3) < U_0$ .

**Proof of part 3:** The argument in part 2 applies not just to the budget curve, but to any graph of the form  $(\pi_1, g(\pi_1))$  where  $g$  is a decreasing, concave function. In particular, it can be applied to the material Pareto-efficiency frontier. □

**Theorem 1** *Suppose  $U$  is continuous, joint-monotonic, and quasi-concave. The second-mover's favorite transaction  $(a_1^*, a_2^*)$  exists and is unique. Moreover, a transaction is utility Pareto efficient if and only if it is materially Pareto efficient and satisfies  $\pi_2(a_1, a_2) \leq \pi_2(a_1^*, a_2^*)$ .*

**Proof:** Note that any point not on the material Pareto-efficiency frontier cannot be either utility Pareto efficient or a maximum for the second-mover's utility  $U$ : If  $(\pi_1, \pi_2)$  is a materially Pareto-inefficient payoff pair, then for sufficiently small  $\epsilon > 0$ , the box  $[\pi_1, \pi_1 + \epsilon] \times [\pi_2, \pi_2 + \epsilon]$  still lies below the material Pareto-efficiency frontier (and so consists entirely of feasible payoff pairs), and joint-monotonicity implies that some point  $(\pi_1', \pi_2')$  inside this box gives the second-mover higher utility than  $(\pi_1, \pi_2)$ . Clearly  $(\pi_1', \pi_2')$  also gives the first-mover higher utility than  $(\pi_1, \pi_2)$ , so we have an individually-rational point whose payoffs for both parties are higher than at  $(\pi_1, \pi_2)$ .

Now, since the space of individually-rational payoffs is compact (by Technical Lemma), there must be some point that maximizes  $U$ ; as we have just seen, this point must be materially Pareto-efficient. Technical Lemma further implies that there must be a *unique* payoff pair that achieves the maximum utility. Since this payoff pair is on the material Pareto-efficiency frontier, there is in turn only one transaction  $(a_1^*, a_2^*)$  that achieves these payoffs. (Any other transaction that gives the same material payoff for the first-mover must take the form  $(a_1^* + \Delta, a_2^* + \Delta)$ . But there is exactly one value of  $\Delta$  that maximizes  $\pi_2(\Delta) = v(a_1^* + \Delta) - c(a_2^* + \Delta)$ .)

Let  $(\pi_1^*, \pi_2^*) = \pi(a_1^*, a_2^*)$  be the materially Pareto-efficient payoff pair that maximizes  $U$ . If  $(\pi_1', \pi_2')$  is any other materially Pareto-efficient pair with  $\pi_2' > \pi_2^*$ , then  $\pi_1' < \pi_1^*$  (by material Pareto-efficiency of  $(\pi_1^*, \pi_2^*)$ ), so both the first-mover's payoff and the second-mover's utility are

lower at  $(\pi'_1, \pi'_2)$  than at  $(\pi_1^*, \pi_2^*)$ . So  $(\pi'_1, \pi'_2)$  is not utility Pareto efficient.

We have now shown that any utility Pareto-efficient transaction  $(a_1, a_2)$  is materially Pareto efficient and satisfies  $\pi_2(a_1, a_2) \leq \pi_2(a_1^*, a_2^*)$ . Conversely, consider any transaction  $(a_1, a_2)$  meeting these conditions. If it is not utility Pareto efficient, then there exists another transaction  $(a'_1, a'_2)$  (which we may assume to be materially Pareto-efficient) giving at least equally high material payoff to the first-mover and utility to the second-mover. Now  $\pi(a'_1, a'_2), \pi(a_1, a_2), \pi(a_1^*, a_2^*)$  lie in that order along the material Pareto-efficiency frontier, and  $U(\pi(a'_1, a'_2)) \geq U(\pi(a_1, a_2))$ . But this contradicts the fact (from Technical Lemma) that  $U$  is strictly decreasing as we move away from  $(a_1^*, a_2^*)$  along the material Pareto-efficiency frontier.

□

**Lemma 1** *Suppose  $U$  is continuous, joint-monotonic, and quasi-concave. Then:*

1. *For any transfer by the first-mover  $a_1$ , the second-mover has a unique optimal action  $a_2(a_1)$  that is a continuous function of  $a_1$ .*
2. *If  $U$  is (strictly) locally normal at  $(I(\hat{a}_1, a_2(\hat{a}_1)); p(a_2(\hat{a}_1)))$ , then  $a_2(a_1)$  is (strictly) increasing in  $a_1$  at  $\hat{a}_1$ . Hence if  $U$  is (strictly) normal, then  $a_2(a_1)$  is (strictly) increasing in  $a_1$  at all  $\hat{a}_1$ .*
3. *If  $(\hat{a}_1, a_2(\hat{a}_1))$  is an equilibrium, then  $a_2(a_1)$  is strictly increasing in  $a_1$  at  $\hat{a}_1$ .*
4. *If  $U$  is continuously differentiable at some  $(\hat{a}_1, a_2(\hat{a}_1))$  and satisfies (TA), then  $\frac{\partial U}{\partial \pi_1} > 0$  and  $\frac{\partial U}{\partial \pi_2} > 0$  at  $(\hat{a}_1, a_2(\hat{a}_1))$ .*

**Proof of part 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-semi-continuous correspondence. Since  $a_2(a_1)$  is single-valued, it is a continuous function.

**Proof of part 2:** Consider an increase in the first-mover's action  $\hat{a}'_1 > \hat{a}_1$ . If the second-mover did not change his action, then the second-mover's material payoff would rise while the first-mover's would fall:  $\pi_2(\hat{a}'_1, a_2(\hat{a}_1)) > \pi_2(\hat{a}_1, a_2(\hat{a}_1))$  and  $\pi_1(\hat{a}'_1, a_2(\hat{a}_1)) < \pi_1(\hat{a}_1, a_2(\hat{a}_1))$ . But the slope of the second-mover's budget curve  $p(a_2(\hat{a}_1))$  would remain the same. If  $U$  is weakly locally normal at  $(I(\hat{a}_1, a_2(\hat{a}_1)); p(a_2(\hat{a}_1)))$ , then for a small enough  $\varepsilon > 0$ , the second mover

weakly prefers the material payoff pair on the budget line that gives slightly more material payoff to the first mover and slightly less to himself,  $(\pi_2(\hat{a}'_1, a_2(\hat{a}_1)) - p(a_2(\hat{a}_1))\varepsilon, \pi_1(\hat{a}'_1, a_2(\hat{a}_1)) + \varepsilon)$ , to  $(\pi_2(\hat{a}'_1, a_2(\hat{a}_1)), \pi_1(\hat{a}'_1, a_2(\hat{a}_1)))$ . It follows from continuity of  $U$  that  $a_2(\hat{a}'_1) \geq a_2(\hat{a}_1)$ . Clearly, if  $U$  is weakly normal, then this argument applies at all  $\hat{a}_1$ . If  $U$  is locally normal, then the same argument implies that  $a_2(\hat{a}'_1) > a_2(\hat{a}_1)$ , and this argument applies at all  $\hat{a}_1$  if  $U$  is normal.

**Proof of part 3:** If  $a_2(a_1)$  were not strictly increasing in  $a_1$  at equilibrium action  $\hat{a}_1$ , then the first mover could increase  $\pi_1$  by slightly reducing  $a_1$ , contradicting equilibrium.

**Proof of part 4:** Since  $U$  is continuously differentiable at  $(\hat{a}_1, a_2(\hat{a}_1))$ , the second-mover's unique optimum is characterized by the first-order condition,  $U_{a_2}(\hat{a}_1, a_2) = \frac{\partial U}{\partial \pi_1} - c'(a_2) \frac{\partial U}{\partial \pi_2} = 0$ . Joint-monotonicity rules out that both partial derivatives  $\frac{\partial U}{\partial \pi_1}$  and  $\frac{\partial U}{\partial \pi_2}$  are negative, and (TA) rules out that they both equal 0. Therefore, the first-order condition implies that both are positive.  $\square$

**Lemma 2** *Suppose  $U$  is continuous, joint-monotonic, and quasi-concave. Then:*

1. *There exists a unique  $\hat{a}_1$  such that the resulting transaction  $(\hat{a}_1, a_2(\hat{a}_1))$  is materially Pareto efficient. This transaction is the second-mover's favorite transaction (and so is utility Pareto efficient).*
2. *An equilibrium exists. Moreover, if there is some individually rational, materially efficient transaction,  $(\hat{a}_1, \hat{a}_2)$ , and some material payoff pair on the same interpersonal indifference curve,  $(\tilde{\pi}_1, \tilde{\pi}_2)$ , such that  $-\frac{\tilde{\pi}_2 - \pi_2(\hat{a}_1, \hat{a}_2)}{\tilde{\pi}_1 - \pi_1(\hat{a}_1, \hat{a}_2)} > c'(\hat{a}_2)$ , then there exists an equilibrium in which the players exchange rather than taking their outside options.*

**Proof of part 1:** We will prove that given any action  $\hat{a}_1$ , the transaction  $(\hat{a}_1, a_2(\hat{a}_1))$  resulting from the unique best-response  $a_2(\hat{a}_1)$  is materially Pareto-efficient if and only if  $(\hat{a}_1, a_2(\hat{a}_1))$  is the second-mover's favorite transaction. The “if” direction follows immediately from the fact that the second-mover's favorite transaction is materially Pareto-efficient, so we focus on the “only if” direction. Suppose  $(\hat{a}_1, a_2(\hat{a}_1))$  is materially Pareto-efficient but is not the second-mover's favorite transaction  $(a_1^*, a_2^*)$ . Every budget curve  $B(a_1)$  touches the material Pareto-efficiency frontier at exactly one point (the unique  $a_2$  solving  $v'(a_1) = c'(a_2)$ ) and is tangent to the material Pareto-efficiency frontier at that point. Hence the second-mover's indifference curve passing through  $(\hat{a}_1, a_2(\hat{a}_1))$  touches the material Pareto-efficiency frontier only at  $(\hat{a}_1, a_2(\hat{a}_1))$ . So there is some  $(a'_1, a'_2)$  on the material Pareto-efficiency frontier between  $(\hat{a}_1, a_2(\hat{a}_1))$  and  $(a_1^*, a_2^*)$ , sufficiently close

to  $(\hat{a}_1, a_2(\hat{a}_1))$ , such that  $U(\pi(a'_1, a'_2)) < U(\pi(\hat{a}_1, a_2(\hat{a}_1)))$ . But this contradicts the fact that  $U$  is strictly decreasing as we move away from  $(a_1^*, a_2^*)$  along the material Pareto-efficiency frontier (as proved in Technical Lemma).

**Proof of part 2:** Let  $\hat{\pi}_1 = \pi_1(\hat{a}_1, \hat{a}_2)$  and  $\hat{\pi}_2 = \pi_2(\hat{a}_1, \hat{a}_2)$ . Recall that, since  $(\hat{a}_1, \hat{a}_2)$  is materially efficient,  $B(\hat{a}_1)$  (the budget curve facing the second mover after action  $\hat{a}_1$ ) is tangent to the material efficiency frontier at  $(\hat{\pi}_1, \hat{\pi}_2)$ , and both have slope  $c'(\hat{a}_2)$  at  $(\hat{\pi}_1, \hat{\pi}_2)$ . The assumption that  $-\frac{\tilde{\pi}_2 - \hat{\pi}_2}{\tilde{\pi}_1 - \hat{\pi}_1} > c'(\hat{a}_2)$  implies that the line  $l$  through the points  $(\tilde{\pi}_1, \tilde{\pi}_2)$  and  $(\hat{\pi}_1, \hat{\pi}_2)$  lies below  $B(\hat{a}_1)$  in a neighborhood to the left of  $(\hat{\pi}_1, \hat{\pi}_2)$ . Since  $(\tilde{\pi}_1, \tilde{\pi}_2)$  and  $(\hat{\pi}_1, \hat{\pi}_2)$  lie on the same interpersonal indifference curve, this interpersonal indifference curve must lie below  $l$  in a neighborhood to the left of  $(\hat{\pi}_1, \hat{\pi}_2)$  (using continuity, joint-monotonicity, and quasi-concavity of  $U$ ). Hence the indifference curve lies strictly below  $B(\hat{a}_1)$  in a neighborhood to the left of  $(\hat{\pi}_1, \hat{\pi}_2)$ . But then joint-monotonicity implies there is some  $(\bar{\pi}_1, \bar{\pi}_2) \in B(\hat{a}_1)$  with  $\bar{\pi}_1 > \hat{\pi}_1$  (and  $\bar{\pi}_2 > \hat{\pi}_2$ ) and  $U(\bar{\pi}_1, \bar{\pi}_2) > U(\hat{\pi}_1, \hat{\pi}_2)$ . Technical Lemma then implies that the second-mover's most-preferred  $(\pi_1, \pi_2)$  in  $B(\hat{a}_1)$  also gives  $\pi_1 > \hat{\pi}_1$  and  $U(\pi_1, \pi_2) > U(\hat{\pi}_1, \hat{\pi}_2)$ .

We have shown that if the first mover chooses action  $\hat{a}_1$ , she will get some material payoff  $\pi_1 > \hat{\pi}_1 \geq 0$ , where the last inequality follows from the fact that  $(\hat{a}_1, \hat{a}_2)$  is individually rational. (The second mover will choose to exchange rather than taking his outside option because  $U(\pi_1, \pi_2) > U(\hat{\pi}_1, \hat{\pi}_2) \geq 0$ , where the last inequality follows from the fact that  $(\hat{a}_1, \hat{a}_2)$  is individually rational.) Since some other action may give an even higher material payoff than  $\hat{a}_1$  does, this is a lower bound on the first mover's equilibrium payoff. From Technical Lemma, the set of individually rational transactions  $T$  is compact. Since  $\pi_1(a_1, a_2(a_1))$  is continuous, there exists an optimal action  $a_1$  in  $T$ . The result follows. □

**Theorem 2** *Suppose  $U$  is joint-monotonic, quasi-concave, continuously twice-differentiable, and satisfies (TA), and suppose  $c'' > 0$ . Then no equilibrium is materially Pareto-efficient. Furthermore, at any interior equilibrium  $(a_1, a_2(a_1))$ , the marginal inefficiency is:*

$$v'(a_1) - c'(a_2) = \frac{1}{N(I(a_1, a_2); p(a_2))} \frac{c''(a_2)}{\frac{d^2 \pi_2}{d(\pi_1)^2} \Big|_{U=U(a_1, a_2)}} > 0.$$

**Proof:** The first-mover's optimal action gives him more than his outside option material payoff 0 (Lemma 2). Joint-monotonicity implies that some point  $(\pi_1, \pi_2)$  inside the box  $[0, \epsilon] \times [0, \epsilon]$  gives the second-mover strictly higher utility than  $U(0, 0) = 0$ . Therefore, the second-mover's favorite

transaction  $(\pi_1^*, \pi_2^*)$  gives him  $U(\pi_1^*, \pi_2^*) > 0$ . It follows that any equilibrium  $(a_1, a_2)$  where the constraint  $U(a_1, a_2) \geq 0$  binds cannot be the second-mover's favorite. Hence Lemma 2 implies that  $(a_1, a_2)$  is not materially Pareto efficient.

So suppose  $(a_1, a_2)$  is an equilibrium where the constraint  $U(a_1, a_2) \geq 0$  does not bind. Since  $U$  is continuously twice-differentiable, both the first-mover's first-order condition and the second-mover's first-order condition hold. Implicitly differentiating the second-mover's first-order condition,  $\frac{\partial U}{\partial \pi_1} - c'(a_2) \frac{\partial U}{\partial \pi_2} = 0$ , gives

$$\begin{aligned} \frac{da_2(a_1)}{da_1} &= -\frac{-\frac{\partial^2 U}{\partial(\pi_1)^2} + v'(a_1) \frac{\partial^2 U}{\partial \pi_1 \partial \pi_2} + c'(a_2) \frac{\partial^2 U}{\partial \pi_1 \partial \pi_2} - c'(a_2) v'(a_1) \frac{\partial^2 U}{\partial(\pi_2)^2}}{\frac{\partial^2 U}{\partial(\pi_1)^2} - 2c'(a_2) \frac{\partial^2 U}{\partial \pi_1 \partial \pi_2} - (c'(a_2))^2 \frac{\partial^2 U}{\partial(\pi_2)^2} - c''(a_2) \frac{\partial U}{\partial \pi_2}} \\ &= (v'(a_1) - c'(a_2)) \frac{-\frac{\partial U}{\partial \pi_1} \frac{\partial^2 U}{\partial(\pi_2)^2} + \frac{\partial U}{\partial \pi_2} \frac{\partial^2 U}{\partial \pi_1 \partial \pi_2}}{\left(\frac{\partial U}{\partial \pi_2}\right)^2 \left(\frac{d^2 \pi_2}{d(\pi_1)^2} + c''(a_2)\right)} + \frac{\frac{d^2 \pi_2}{d(\pi_1)^2}}{\frac{d^2 \pi_2}{d(\pi_1)^2} + c''(a_2)}, \end{aligned}$$

where the second equality uses  $c'(a_2) = \frac{\partial U}{\partial \pi_1} \frac{\partial U}{\partial \pi_2}$  and  $\frac{d^2 \pi_2}{d(\pi_1)^2} \Big|_{U=U(a_1, a_2)} = \frac{-\frac{\partial^2 U}{\partial(\pi_1)^2} \left(\frac{\partial U}{\partial \pi_2}\right)^2 + 2\frac{\partial U}{\partial \pi_1} \frac{\partial U}{\partial \pi_2} \frac{\partial^2 U}{\partial \pi_1 \partial \pi_2} - \frac{\partial^2 U}{\partial(\pi_2)^2} \left(\frac{\partial U}{\partial \pi_1}\right)^2}{\left(\frac{\partial U}{\partial \pi_2}\right)^3}$ .

Now consider maximizing  $U$  with respect to a budget line:  $\tilde{\pi}_1(I; p) = \arg \max_{\{\pi_1: p\pi_1 + \pi_2 = I\}} U(\pi_1, \pi_2)$ , which has first-order condition,  $\frac{\partial U(\tilde{\pi}_1, I - p\tilde{\pi}_1)}{\partial \pi_1} - p \frac{\partial U(\tilde{\pi}_1, I - p\tilde{\pi}_1)}{\partial \pi_2} = 0$ . Implicitly differentiating to find the local income effect:

$$\begin{aligned} \frac{\partial \tilde{\pi}_1(I; p)}{\partial I} &= \frac{-p \frac{\partial^2 U}{\partial(\pi_2)^2} + \frac{\partial^2 U}{\partial \pi_1 \partial \pi_2}}{\frac{\partial^2 U}{\partial(\pi_1)^2} - 2p \frac{\partial^2 U}{\partial \pi_1 \partial \pi_2} + p^2 \frac{\partial^2 U}{\partial(\pi_2)^2}} \\ &= \frac{-\frac{\partial U}{\partial \pi_1} \frac{\partial^2 U}{\partial(\pi_2)^2} + \frac{\partial U}{\partial \pi_2} \frac{\partial^2 U}{\partial \pi_1 \partial \pi_2}}{\left(\frac{\partial U}{\partial \pi_2}\right)^2 \frac{d^2 \pi_2}{d(\pi_1)^2}} \end{aligned}$$

where the second equality uses  $p = \frac{\frac{\partial U}{\partial \pi_1}}{\frac{\partial U}{\partial \pi_2}}$  and the formula for  $\frac{d^2 \pi_2}{d(\pi_1)^2} \Big|_{U=U(a_1, a_2)}$ . Substituting,

$$\frac{da_2(a_1)}{da_1} = (v'(a_1) - c'(a_2)) \frac{\partial \tilde{\pi}_1(I; p)}{\partial I} \frac{\frac{d^2 \pi_2}{d(\pi_1)^2}}{\frac{d^2 \pi_2}{d(\pi_1)^2} + c''(a_2)} + \frac{\frac{d^2 \pi_2}{d(\pi_1)^2}}{\frac{d^2 \pi_2}{d(\pi_1)^2} + c''(a_2)}. \quad (18)$$

Equation (8) in the text follows from substituting (18) into

$$\frac{d\pi_1(a_1, a_2(a_1))}{da_1} = \frac{da_2(a_1)}{da_1} - 1.$$

Setting this equal to zero leads to equation (7). □

**Theorem 3** *Suppose  $U$  is continuous, joint-monotonic, quasi-concave, and normal. If  $c(a_2) = \kappa_1 + \kappa_2 a_2$  for  $\kappa_2 > 0$ , then the unique equilibrium transaction is the second-mover's favorite transaction (and so is utility Pareto efficient).*

**Proof:** Since  $c(a_2) = \kappa_1 + \kappa_2 a_2$ , the budget constraints (and material Pareto-efficiency frontier) are parallel lines with slope  $-\kappa_2$ , so they can be written in the form  $\kappa_2 \pi_1 + \pi_2 = v(a_1) - \kappa_2 a_1 - \kappa_1 \equiv I(a_1)$ . Because  $U$  is normal, the second-mover chooses  $a_2(a_1)$  from each  $B(a_1)$  such that  $\pi_1$  and  $\pi_2$  are both strictly increasing in  $I(a_1)$ . The first-mover maximizes her material payoff by taking the action  $\tilde{a}_1$  that maximizes  $I(a_1)$ , which satisfies  $v'(\tilde{a}_1) = \kappa_2$ , the condition that characterizes material Pareto efficiency. Since the budget constraint  $B(\tilde{a}_1)$  is the material Pareto-efficiency frontier, the resulting transaction is the second-mover's favorite transaction.  $\square$

**Lemma 3** *Suppose  $U$  is fairness-kinked, with  $U^A$  and  $U^B$  being joint-monotonic, quasi-concave, continuously twice-differentiable, and satisfying (TA). Let  $(a_1^*, a_2^*)$  denote the second-mover's favorite transaction. If the utility function satisfies*

$$U^A = U^B \tag{19}$$

$$\frac{\partial U^A}{\partial \pi_1} - c'(a_2^*) \frac{\partial U^A}{\partial \pi_2} < 0 \tag{20}$$

$$\frac{\partial U^B}{\partial \pi_1} - c'(a_2^*) \frac{\partial U^B}{\partial \pi_2} > 0 \tag{21}$$

*at  $(a_1^*, a_2^*)$ , then the second-mover's optimal strategy  $a_2(a_1)$  satisfies the fairness rule for all  $a_1$  in a neighborhood of  $a_1^*$ , and  $a_1^*$  is a locally optimal action for the first mover.*

**Proof:** Since the given inequalities hold strictly at  $(a_1^*, a_2^*)$ , there are some neighborhoods  $X, Y$  of  $a_1^*, a_2^*$ , 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  $a_1^*$ , the second-mover will choose action  $a_2(a_1)$  such that  $U^A = U^B$  at  $\pi(a_1, a_2(a_1))$ . The single-crossing properties in the definition of fairness-kinked utility imply that along the curve  $U^A = U^B$ ,  $\pi_1$  and  $\pi_2$  are increasing in tandem—that is, for points  $(\pi_1', \pi_2')$  and  $(\pi_1'', \pi_2'')$  on the curve,  $\pi_2' < \pi_2''$  if and only if  $\pi_1' < \pi_1''$ . It follows immediately that  $a_1^*$  is a locally optimal action for the first mover.  $\square$

**Theorem 4** *Suppose  $U$  is fairness-kinked, with  $U^A$  and  $U^B$  being joint-monotonic, quasi-concave, normal, continuously twice-differentiable, and satisfying (TA). Let  $(a_1^*, a_2^*)$  denote the second-mover's favorite transaction, and let  $(\hat{a}_1, \hat{a}_2)$  denote the (necessarily unique) transaction with  $\hat{a}_1 < a_1^*$  such that  $\pi_1(\hat{a}_1, \hat{a}_2) = \pi_1(a_1^*, a_2^*)$  and  $U(\hat{a}_1, \hat{a}_2) = 0$ . If the utility function satisfies*

$$U^A = U^B \tag{22}$$

$$\frac{\partial U^A}{\partial \pi_1} - c'(\widehat{a}_2) \frac{\partial U^A}{\partial \pi_2} < 0 \quad (23)$$

$$\frac{\partial U^B}{\partial \pi_1} - c'(a_2^*) \frac{\partial U^B}{\partial \pi_2} > 0 \quad (24)$$

at  $(a_1^*, a_2^*)$ , then the second-mover's optimal strategy  $a_2(a_1)$  satisfies the fairness rule for all  $a_1$  in a neighborhood of  $a_1^*$ , and  $(a_1^*, a_2^*)$  is the unique equilibrium transaction (and so is utility Pareto efficient).

**Proof:** We first show that  $(\widehat{a}_1, \widehat{a}_2)$  exists and is unique. For each possible real number  $a_1$ , consider the unique  $\widehat{a}_2(a_1)$  such that  $\pi_1(a_1, \widehat{a}_2(a_1)) = \pi_1(a_1^*, a_2^*)$  (namely  $\widehat{a}_2 \equiv a_1 + \pi_1(a_1^*, a_2^*)$ ). Then  $\pi_2(a_1, \widehat{a}_2(a_1)) = v(a_1) - c(a_1 + \pi_1(a_1^*, a_2^*))$ ; the derivative of this expression with respect to  $a_1$  is  $v'(a_1) - c'(a_1 + \pi_1(a_1^*, a_2^*))$ , which is decreasing. That is,  $\pi_2(a_1, \widehat{a}_2(a_1))$  is concave in  $a_1$ ; and it must reach a maximum at  $a_1 = a_1^*, \widehat{a}_2(a_1) = a_2^*$ , since  $(a_1^*, a_2^*)$  is materially Pareto efficient. Therefore, any value less than  $\pi_2(a_1^*, a_2^*)$  is achieved exactly once by  $\pi_2(a_1, \widehat{a}_2(a_1))$  for  $a_1 < a_1^*$ . In particular, the (unique) value of  $\pi_2 \leq \pi_2(a_1^*, a_2^*)$  such that  $U(\pi_1(a_1^*, a_2^*), \pi_2) = 0$  is achieved exactly once, which is what was required. Note that  $\widehat{a}_2 < a_2^*$ .

By hypothesis,

$$\frac{\partial U^A}{\partial \pi_1} - c'(\widehat{a}_2) \frac{\partial U^A}{\partial \pi_2} < 0$$

at  $\pi(a_1^*, a_2^*)$ . This ensures  $\frac{\partial U^A}{\partial \pi_2} \geq 0$  at  $\pi(a_1^*, a_2^*)$  (otherwise we would have to have  $\frac{\partial U^A}{\partial \pi_1} < 0$ , violating joint-monotonicity). Since  $\widehat{a}_2 < a_2^*$ , we have  $c'(\widehat{a}_2) < c'(a_2^*)$ , so

$$\frac{\partial U^A}{\partial \pi_1} - c'(a_2^*) \frac{\partial U^A}{\partial \pi_2} < 0$$

at  $\pi(a_1^*, a_2^*)$  also. Hence Lemma 3 implies that the second-mover's optimal strategy  $a_2(a_1)$  satisfies the fairness rule for all  $a_1$  in a neighborhood of  $a_1^*$ .

We next show that

$$\frac{\partial U^A}{\partial \pi_1} - c'(\widehat{a}_2) \frac{\partial U^A}{\partial \pi_2} < 0$$

at all individually-rational transactions  $\pi(a_1, a_2)$  such that  $\pi_1(a_1, a_2) > \pi_1(a_1^*, a_2^*)$ . Suppose to the contrary there were some  $\pi(a_1, a_2)$  at which  $\frac{\partial U^A}{\partial \pi_1} - c'(\widehat{a}_2) \frac{\partial U^A}{\partial \pi_2} \geq 0$ .  $\frac{\partial U^A}{\partial \pi_2} \geq 0$  there (else  $\frac{\partial U^A}{\partial \pi_1} < 0$ , violating joint-monotonicity), so by choosing a value  $k$  slightly smaller than  $c'(\widehat{a}_2)$ , we have

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

strictly at  $\pi(a_1, a_2)$ . Since  $k$  is very close to  $c'(\widehat{a}_2)$ , we know that

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

at  $\pi(a_1^*, a_2^*)$ . So if we draw “budget lines”  $l, l'$  each with slope  $-k$  passing through the two points  $\pi(a_1^*, a_2^*)$  and  $\pi(a_1, a_2)$ , respectively, the second-mover’s most-preferred point on line  $l$  is below  $\pi(a_1^*, a_2^*)$  and his most-preferred point on  $l'$  is above  $\pi(a_1, a_2)$ . (Even though we have only local conditions, we know this holds for the global optimum on each line by Technical Lemma.) By assumption,  $\pi_1(a_1, a_2) > \pi_1(a_1^*, a_2^*)$ . So the normal good assumption is violated unless  $l$  lies to the left of  $l'$ . But the slope of the material Pareto-efficiency frontier at  $\pi(a_1^*, a_2^*)$  is  $-c'(a_2^*) < -k$ , which means that  $(\pi_2 - \pi_2^*)/(\pi_1 - \pi_1^*) < -k$ ; hence the line  $l$  lies to the right of  $l'$ . So, we have a contradiction.

To prove that  $a_1^*$  is the first-mover’s global optimum, we show that at any individually-rational transaction  $(a_1, a_2)$  that gives the first-mover a higher material payoff, the second-mover would prefer to reduce his action. First, we claim that at any such transaction  $(a_1, a_2)$ , we have  $a_2 > \widehat{a}_2$ . Accepting this for now, we get  $c'(a_2) > c'(\widehat{a}_2)$ ; hence, using the result from the previous paragraph,  $\frac{\partial U^A}{\partial \pi_1} - c'(a_2) \frac{\partial U^A}{\partial \pi_2} < 0$  at  $\pi(a_1, a_2)$ . (Again,  $\frac{\partial U^A}{\partial \pi_2} \geq 0$ ; else joint-monotonicity is violated.) This implies that, at this transaction, the second-mover would strictly prefer to lower his action (note that since we assumed  $\pi_1(a_1, a_2) > \pi_1(a_1^*, a_2^*)$ ,  $U^A$  is indeed the relevant part of the utility function here).

It remains only to show that, for any individually-rational transaction  $(a_1, a_2)$  with  $\pi_1(a_1, a_2) > \pi_1(a_1^*, a_2^*)$ , we have  $a_2 > \widehat{a}_2$ . So suppose  $a_2 \leq \widehat{a}_2$ . Then we must have  $a_1 < \widehat{a}_1$ . Using concavity of  $v$  and convexity of  $c$ , we get  $v'(\widehat{a}_1) > v'(a_1^*) = c'(a_2^*) > c'(\widehat{a}_2)$ , and concavities also give

$$\begin{aligned} \frac{\pi_2(a_1, a_2) - \pi_2(\widehat{a}_1, \widehat{a}_2)}{\pi_1(a_1, a_2) - \pi_1(\widehat{a}_1, \widehat{a}_2)} &< \frac{(a_1 - \widehat{a}_1)v'(\widehat{a}_1) - (a_2 - \widehat{a}_2)c'(\widehat{a}_2)}{(a_2 - \widehat{a}_2) - (a_1 - \widehat{a}_1)} \\ &< \frac{-(a_2 - \widehat{a}_2)c'(\widehat{a}_2)}{(a_2 - \widehat{a}_2)} = -c'(\widehat{a}_2). \end{aligned}$$

Thus, the line segment from  $\pi(\widehat{a}_1, \widehat{a}_2)$  upward to  $\pi(a_1, a_2)$  has a slope lying between  $-c'(\widehat{a}_2)$  and 0.

At the point  $(\widehat{a}_1, \widehat{a}_2)$ , we have  $U = U^A = 0$ . At this point, the directional derivative of  $U^A$  in the direction  $(-c'(\widehat{a}_2), 1)$  is  $\leq 0$  (this is given); and the directional derivative in the direction  $(-1, 0)$  is also  $\leq 0$  (because  $U$  must increase in the direction leading to  $\pi(a_1^*, a_2^*)$ , namely  $(1, 0)$ ). At least one of these directional derivatives is strictly negative, by (TA). Hence, the derivative in the direction leading to  $\pi(a_1, a_2)$  is strictly negative. It follows that  $U(\pi(a_1, a_2)) \leq U^A(\pi(a_1, a_2)) < 0$ , which contradicts the assumption that this point was individually-rational. This completes the proof.  $\square$

## References

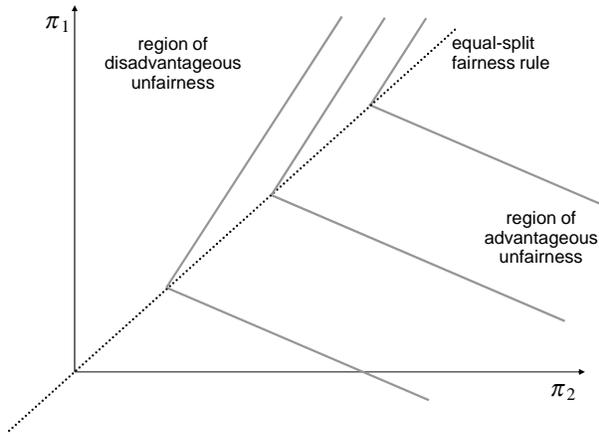
- [1] George A. Akerlof. Labor contracts as partial gift exchange. *Quarterly Journal of Economics*, 97(4):543–569, November 1982.
- [2] George A. Akerlof and Janet L. Yellen. The fair wage-effort hypothesis and unemployment. *Quarterly Journal of Economics*, 105(2):255–283, May 1990.
- [3] Omar Al-Ubaydli, Steffen Andersen, Uri Gneezy, and John List. For love or money? testing non-pecuniary and pecuniary incentive schemes in a field experiment. 2006. Manuscript.
- [4] James Andreoni and B. Douglas Bernheim. Social image and the 50-50 norm. 2007. Manuscript.
- [5] 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.
- [6] Kenneth Arrow. Political and economic evaluation of social effects and externalities. In M.D. Intriligator, editor, *Frontiers of Quantitative Economics*, chapter 1, pages 3–25. North-Holland Publishing Company, 1971.
- [7] Linda Babcock and George Loewenstein. Explaining bargaining impasse: The role of self-serving biases. *Journal of Economic Perspectives*, 11(1):109–126, 1997.
- [8] 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.
- [9] Gary S. Becker. A theory of social interactions. *Journal of Political Economy*, 82(6):1063–93, November/December 1974.
- [10] Charles Bellemare and Bruce Shearer. Gift exchange within a firm: Evidence from a field experiment. 2007. CIRPÉE Working Paper 07-08.
- [11] Joyce Berg, John Dickhaut, and Kevin McCabe. Trust, reciprocity, and social history. *Games and Economic Behavior*, 10:122–142, 1995.
- [12] Theodore C. Bergstrom. A fresh look at the rotten kid theorem – and other household mysteries. *Journal of Political Economy*, 97(5):1138–59, 1989.
- [13] Truman F. Bewley. *Why Wages Don't Fall During a Recession*. Harvard University Press, Cambridge, MA, 1999.

- [14] Gary E. Bolton and Axel Ockenfels. Erc: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1):166–193, March 2000.
- [15] Colin F. Camerer. *Behavioral Game Theory*. Princeton University Press, Princeton, NJ, 2003.
- [16] Alexander W. Cappelen, Astri Drange Hole, Erik Ø. Sørensen, and Bertil Tungodden. The pluralism of fairness ideals: An experimental approach. *American Economic Review*, 97(3):818–827, 2007.
- [17] Gary Charness and Matthew Rabin. Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117(3):817–869, August 2002.
- [18] Ronald H. Coase. The problem of social cost. *Journal of Law and Economics*, 3:1–44, 1960.
- [19] Alain Cohn, Ernst Fehr, and Lorenz Goette. Gift exchange and effort: Evidence from a field experiment. 2007. University of Zürich Manuscript.
- [20] Michael Conlin, Michael Lynn, and Ted O’Donoghue. The norm of restaurant tipping. *Journal of Economic Behavior and Organization*, 52:297–321, 2003.
- [21] James C. Cox, Daniel Friedman, and Vjollca Sadiraj. Revealed altruism. *Econometrica*, 76(1):31–69, 2008.
- [22] Robyn M. Dawes and Richard H. Thaler. Cooperation. *Journal of Economic Perspectives*, 2(3):187–197, 1988.
- [23] Stephen J. Dubner and Steven D. Levitt. What the bagel man saw. *New York Times Magazine*, June 6 2004.
- [24] 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.
- [25] Dirk Engelmann and Martin Strobel. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review*, 94(4):857–869, September 2004.
- [26] Ernst Fehr, Georg Kirchsteiger, and Arno Riedl. Does fairness prevent market clearing? an experimental investigation. *Quarterly Journal of Economics*, 108(2):437–459, 1993.

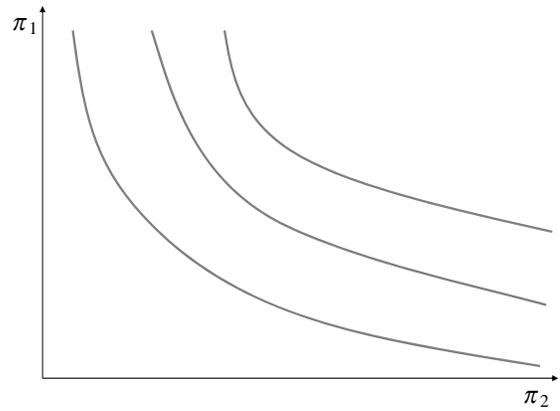
- [27] Ernst Fehr, Alexander Klein, and Klaus M. Schmidt. Fairness and contract design. *Econometrica*, 75(1):121–154, 2007.
- [28] Ernst Fehr, Michael Naef, and Klaus M. Schmidt. Inequality aversion, efficiency, and maximin preferences in simple distribution games: Comment. *American Economic Review*, 96(5):1912–1917, 2006.
- [29] 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.
- [30] Ernst Fehr and Klaus M. Schmidt. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3):817–868, August 1999.
- [31] Ray Fisman, Shachar Kariv, and Daniel Markovits. Pareto-damaging behaviors. *UC Berkeley mimeo*, 2005.
- [32] James W. Friedman. A non-cooperative equilibrium for supergames. *Review of Economic Studies*, 38(1):1–12, 1971.
- [33] Drew Fudenberg and Eric Maskin. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3):533–554, 1986.
- [34] Uri Gneezy and John List. Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, forthcoming, 2006.
- [35] Jerald Greenberg. Employee theft as a reaction to underpayment inequity: The hidden cost of pay cuts. *Journal of Applied Psychology*, 75(5):561–568, 1990.
- [36] Oliver Hart and John Moore. Contracts as reference points. *Quarterly Journal of Economics*, 123(1):1–48, February 2008.
- [37] Ori Heffetz and Robert H. Frank. Preferences for status: Evidence and economic implications. In Jess Benhabib, Alberto Bisin, and Matthew Jackson, editors, *Handbook of Social Economics*. Elsevier, forthcoming.
- [38] 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.

- [39] Sebastian Kube, Clemens Puppe, and Michel Maréchal. The currency of reciprocity. 2006a. Manuscript.
- [40] Sebastian Kube, Clemens Puppe, and Michel Maréchal. Putting reciprocity to work- positive versus negative responses in the field. 2006b. University of St. Gallen Discussion Paper no. 2006-27.
- [41] David K. Levine. Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, 1(3):593–731, 1998.
- [42] Sandra Maximiano, Randolph Sloof, and Joep Sonnemans. Gift exchange in a multi-worker firm. *Economic Journal*, 117:1025–1050, 2007.
- [43] Marvin D. Dunnette Pritchard, Robert D. and Dale O. Jorgenson. Effects of perceptions of equity and inequity on worker performance and satisfaction. *Journal of Applied Psychology*, 56(1):75–94, 1972.
- [44] Matthew Rabin. Incorporating fairness into game theory and economics. *American Economic Review*, 83(5):1281–1302, December 1993.
- [45] Matthew Rabin. Bargaining structure, fairness, and efficiency. *Berkeley Department of Economics Working Paper*, February 1997.
- [46] Amartya Sen. Behaviour and the concept of preference. *Economica*, 40(159):241–259, 1973.
- [47] Rangarajan K. Sundaram. *A First Course in Optimization Theory*. Cambridge University Press, Cambridge, UK, 1996.

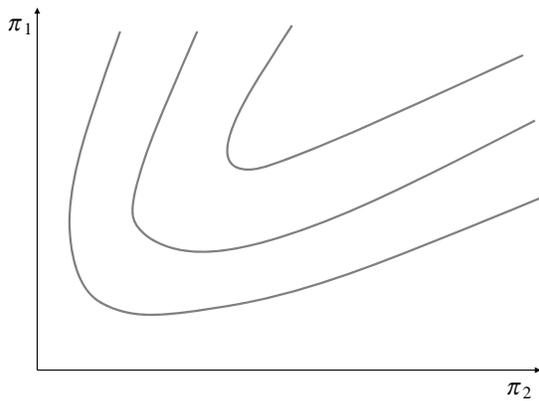
(a)



(b)



(c)



(d)

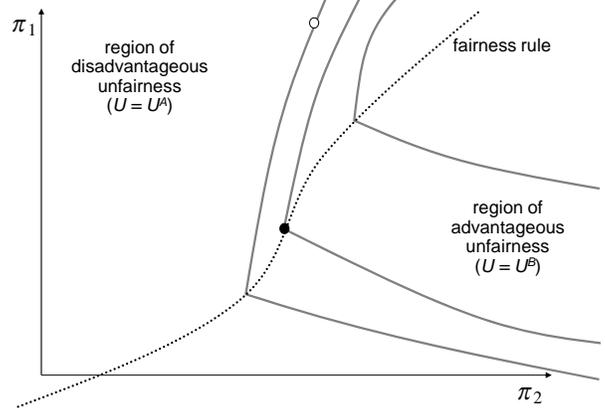


Figure 1. Indifference curves corresponding to social preferences that satisfy joint-monotonicity and quasi-concavity. Panel (a): Inequity-averse preferences. Panel (b): Becker's altruism model, which has monotonic preferences. Panel (c): Smooth preferences that are not monotonic. Panel (d): Fairness-kinked preferences, where the fairness rule is not the equal-sharing rule. Due to the non-monotonicity, the black point is preferred to the white point that gives higher material payoffs to both players.

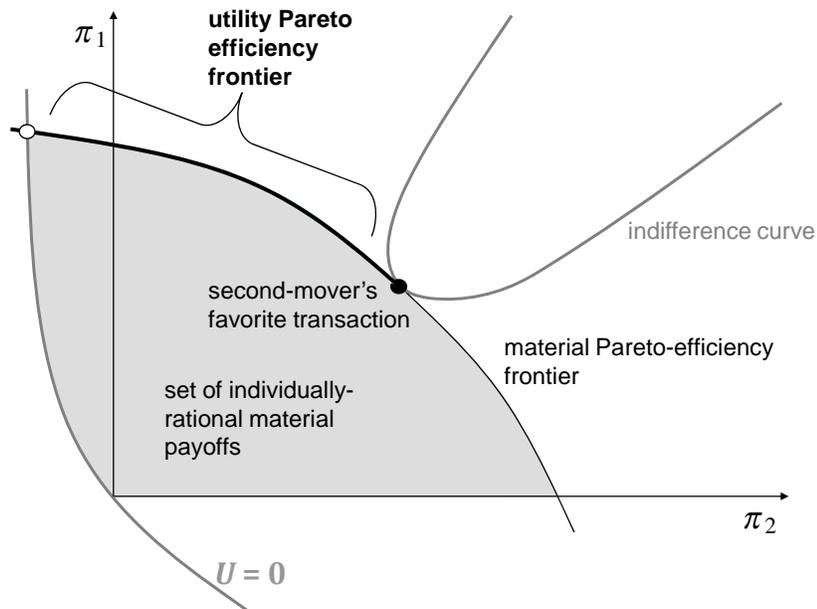


Figure 2. Relationship between utility Pareto efficiency and material Pareto-efficiency. The set of individually-rational material payoffs, shown in gray, is convex and compact. The black point corresponds to the second-mover's favorite transaction. The utility Pareto-efficiency frontier is the subset of the material Pareto-efficiency frontier that gives lower material payoff to the second mover than the second-mover's favorite transaction does. The white point is the outcome if contracts are enforceable and the first mover can make a take-it-or-leave offer to the second mover.

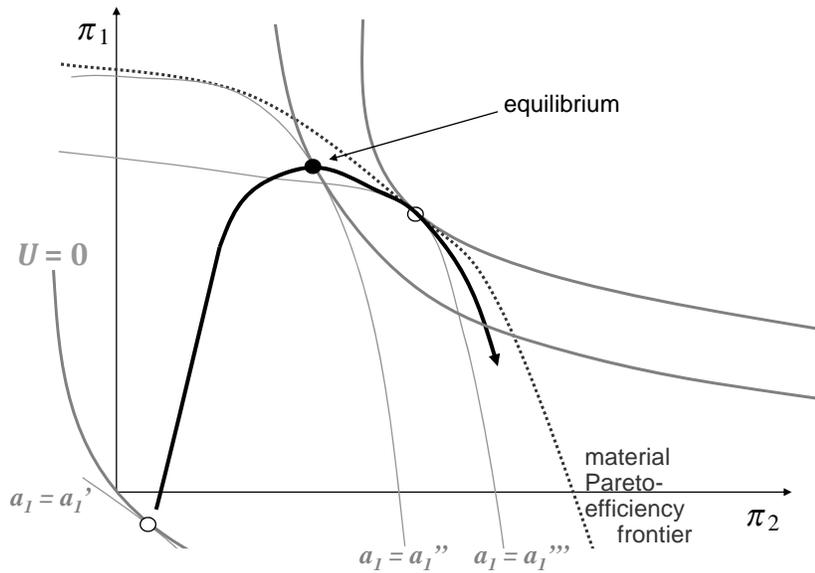


Figure 3. Equilibrium with smooth social preferences and strictly convex  $c(a_2)$  is inefficient. The material payoff budget curves are shown for first-mover actions  $a_1' < a_1'' < a_1'''$ . The white points show the action the second mover would choose at non-equilibrium actions by the first mover. The arrow illustrates how material payoffs vary as the first-mover's action increases. Note that the materially Pareto-efficient transaction along the path of the arrow is not an equilibrium because the first mover can profitably deviate to a lower action. The black point is the equilibrium.

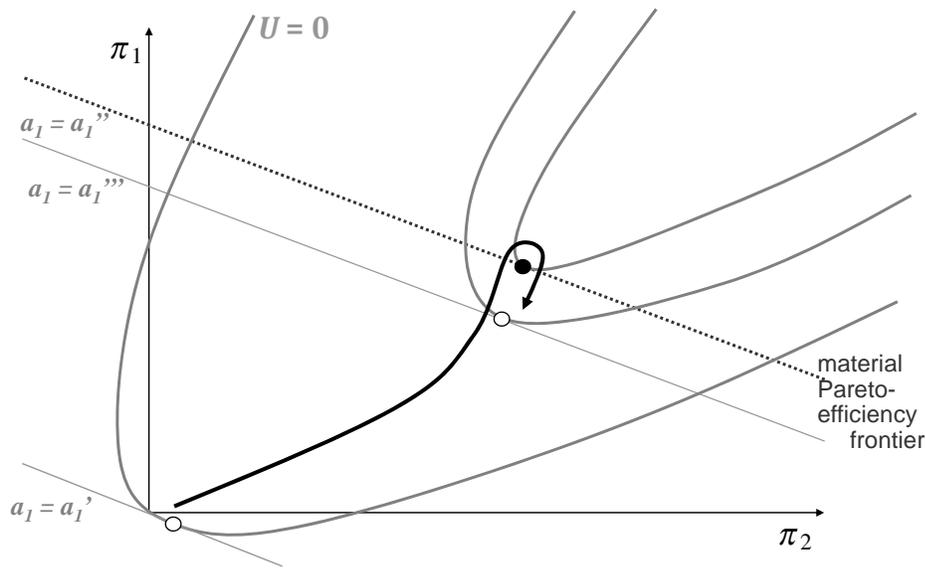


Figure 4. If both players' material payoff functions are linear in the second-mover's action, and if  $U$  is normal, then the equilibrium is efficient. The material payoff budget curves are lines parallel to the material Pareto-efficiency frontier (shown for first-mover actions  $a_1' < a_1'' < a_1'''$ ). The white points show the action the second mover would choose at non-equilibrium actions by the first mover. The arrow illustrates how material payoffs vary as the first-mover's action increases. The black point is the equilibrium.

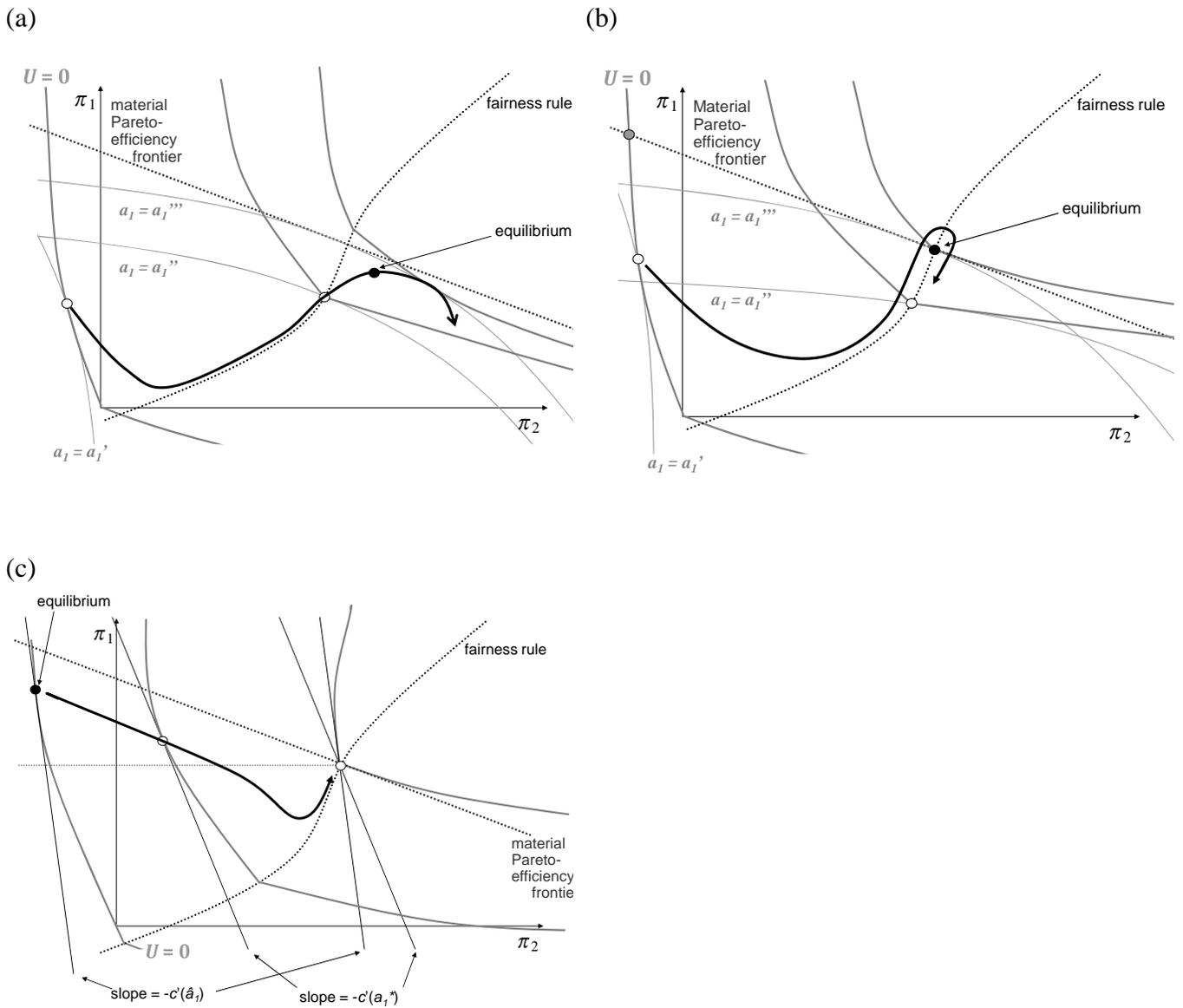


Figure 5. The three possible equilibria with fairness-kinked preferences. The material payoff budget curves are shown for first mover actions  $a_1' < a_1'' < a_1'''$ . The white points show the action the second mover would choose at non-equilibrium actions by the first mover. The arrow illustrates how material payoffs vary as the first-mover's action increases. The black point is the equilibrium. (a) The second-mover's favorite transaction is not on the fairness rule. Equilibrium is inefficient. (b) Equilibrium with sufficiently-kinked and normal social preferences is utility Pareto efficient. (c) The first mover maximizes her material payoff by taking a small action  $\hat{a}_1$  instead of the action  $a_1^*$  that would lead to an efficient outcome. The social preferences shown violate the assumption that  $U$  is normal.