Is Gift-Exchange Efficient? The Rotten Firm Theorem

Daniel J. Benjamin*
Dartmouth College and Institute for Social Research

May 18, 2007
PRELIMINARY AND INCOMPLETE **

Abstract

To what extent can a preference for fair transactions "substitute for" the availability of binding contracts in generating efficient exchange? I analyze a gift-exchange game, where a profit-maximizing firm offers a wage to a fair-minded worker, who then chooses how much effort to exert. I characterize the set of Pareto efficient transactions, given that the worker's utility function incorporates a concern for both his own self-centered payoff and the firm's profit. If the worker's interpersonal preferences are smooth, then the equilibrium is bounded away from efficiency. On the other hand, if the worker's interpersonal preferences are sufficiently-kinked, then the worker chooses effort according to a "fairness rule," meaning that the worker's self-centered payoff and the firm's profit move in tandem. In that case, the equilibrium is Pareto efficient – a result I call the Rotten Firm Theorem. I relate the Rotten Firm Theorem, which relies on kink interpersonal preferences, to the Rotten Kid Theorem, which relies on "transferable utility."

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

Keywords: fairness, social preferences, gift-exchange, efficiency wages, altruism

*A previous version of this paper circulated under the title, "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, Edward Glaeser, David Laibson, Jesse Shapiro, Andrei Shleifer, and Jón Steinsson. 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, Jeffrey Yip, and especially Hongyi Li for outstanding research assistance. All mistakes are my fault. E-mail: daniel.benjamin@gmail.com.

1 Introduction

When is exchange 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, the folk theorem states that some equilibria are (nearly) efficient if the parties are sufficiently patient (e.g., Bull 1987, MacLeod & Malcomson 1989). In the absence of repetition, interpersonal preferences may promote efficiency.

This paper asks: Can a preference for fair transactions generate efficient exchange? This is an important question because a concern for fairness influences behavior in a wide range of market settings (e.g., Kahneman, Knetsch, & Thaler 1986). For example, Dawes & Thaler (1988) observe that it is common in rural areas for farmers to leave fresh produce on a table by the road. Most customers leave money in a cash box, even though they could take the produce without paying. Similarly, Dubner & Levitt (2004) report on a "Bagel Guy" who earns a living by leaving bagels in workplace kitchens, relying on the honor system to be paid. Tipping is a common social norm that relies on customers' sense of fairness to leave a larger tip for better service (Conlin, Lynn, & O'Donoghue 2003).

In labor markets, there is empirical evidence that workers' concern for fair transactions plays a role in motivating effort. Akerlof (1982) and Akerlof & Yellen (1990) review sociological evidence that effort responds positively to the wage for fairness reasons (see also Bewley 1999). Mas (2005) provides field evidence that police officers' effort responds to plausibly exogenous wage changes: When final-offer arbitration of a compensation dispute rules in favor of the police officers, exogenously raising pay, police initiate more and higher-quality arrests.[1] In a suggestive field experiment, Greenberg (1990) found that employee theft (interpretable as negative effort) increased substantially in two of three non-union manufacturing plants that experienced a temporary wage cut during the duration of the pay change, but not in the plant where pay was held constant. 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 only for the first few hours) (see also Pritchard, Dunnette, & Jorgenson 1972).

Much existing work shows that a concern for fairness enables exchange in settings without

[1] Since an arbitrator's ruling is independent of the intentions of the city employers, this evidence suggests that it is the salary itself (rather than the employer's intentions toward the workers) that induces effort.

enforceable contracts, where purely self-regarding agents would not be willing to trade (e.g., Fehr & Schmidt 1999). In that sense, it is well-known that fairness can increase realized gains from trade. However, existing work has not addressed whether a concern for fairness leads to efficient exchange.

To focus on the implications of the preferences, I study the simplest possible setting: a deterministic gift-exchange game where everything is common knowledge. A profit-maximizing firm pays a wage to a fair-minded worker, who provides effort. The "material payoff" of the firm is profit, and the material payoff of the worker corresponds to what his utility would be if he were entirely self-centered, his benefit from a higher wage and convex cost of effort. The firm maximizes its material payoff, while the worker maximizes utility, which depends on both the worker's own material payoff and the firm's material payoff.

The key result is that the equilibrium of the gift-exchange game is Pareto efficient if the worker's interpersonal preferences are sufficiently-kinked. I call this result the Rotten Firm theorem: even though the firm seeks only to maximize profit, it will offer a wage that leads the worker to reciprocate an efficient level of effort. The reason is that the worker chooses effort according to a "fairness rule," which ensures the worker's material payoff increases (or decreases) in tandem with the firm's profit. As a result, the worker will reciprocate a higher wage with greater effort because a higher wage transfers resources from the firm to the worker. To keep material payoffs increasing in tandem, the worker must make a transfer back to the firm by increasing effort. When determining which wage offer will maximize profit, the firm knows that the worker's effort choice will ensure that both players earn a positive share of the marginal gains from trade. It is in the firm's best interest to offer the wage that induces the worker to exert efficient effort, which maximizes the total gains from trade and therefore also the material payoffs to each player.

Rather than being kinked, if the worker's interpersonal preferences are smoothly convex (as in, e.g., Bolton & Ockenfels 2000, Cox, Friedman, & Sadiraj 2005) – exhibiting what I call "second-order unfairness aversion" – then the worker will not follow a fairness rule for all wages, and the equilibrium will be bounded away from efficiency. Although several leading fairness specifications have a kink (Fehr & Schmidt 1999, Charness & Rabin 2002), the papers that introduce these models do not emphasize the kink. Therefore, the central role of the kink in the analysis is surprising. One contribution of this paper is to point out the behavioral and equilibrium consequences of kinked interpersonal preferences, which I call "first-order unfairness aversion."

Do individuals actually have kinked interpersonal preferences? If so, are they kinked enough for gift-exchange to be efficient? In laboratory settings where it is clear how to split the surplus equally,

many individuals choose to do so (Fehr & Schmidt 1999, Charness & Rabin 2002), consistent with interpersonal preferences that are kinked around an "equal-split fairness rule." Moreover, calibrations of the model using preference parameter estimates from those settings suggest that about 40% of the population has "sufficiently-kinked" preferences. Hence one interpretation of the Rotten Firm theorem is that a realistic degree of fair-mindedness can fully substitute for the availability of enforceable contracts. However, in some real-world settings, it may not be clear what effort level satisfies a fairness rule. In laboratory settings where it is unclear how to split the surplus equally, most individuals appear to have smooth interpersonal preferences (Fisman, Kariv, & Markovits 2005). In those cases, the results in this paper imply that a concern for fairness cannot generate fully efficient exchange.

Despite some parallels, the Rotten Firm theorem and the Rotten Kid theorem are distinct results. In the gift-exchange game, the Rotten Kid theorem applies as long as the worker's and firm's material payoff functions are both quasi-linear in effort, even if the worker's interpersonal preferences are smooth. By contrast, the Rotten Firm theorem applies as long as the worker's interpersonal preferences are sufficiently-kinked, even if the material payoff functions are not quasi-linear. Nonetheless, the two theorems are cousins and together constitute the *only* two cases in which interpersonal preferences cause the equilibrium to be efficient.

This paper relates to much previous research that has analyzed how fairness preferences affect behavior in variants of the gift-exchange game (e.g., Fehr & Schmidt 1999, Fehr, Klein, & Schmidt 2006). The preferences discussed in this paper are general enough to embed many of the leading models of interpersonal preferences as special cases (such as Becker 1974, Fehr & Schmidt 1999, Bolton & Ockenfels 2000, Charness & Rabin 2002). However, I do not address signaling (Levine 1998) or intentions-based models (Rabin 1993; Dufwenberg & Battiglini 2005), which are often substantially more complex to analyze.[2] The paper also relates to the social choice literature since interpersonal preferences that are "sufficiently-kinked" satisfy the axiom of resource-monotonicity (Kalai 1977; Roemer 1996).

The paper proceeds as follows. Section 2 describes the gift-exchange game. Section 3 introduces a general class of interpersonal preferences. Given those preferences, Section 4 characterizes the set of Pareto efficient outcomes. Section 5 describes the logic of the Rotten Firm theorem by showing that if the worker chooses effort according to a fairness rule, then the equilibrium is efficient. Section 6 proves that if the worker's interpersonal preferences are smooth, then the worker does not obey a

[2] Virtually all of these models imply that effort is increasing in the wage, but the Rotten Firm theorem will apply only if the worker behaves according to a fairness rule (as described in Section 5).

fairness rule, and the equilibrium cannot be efficient. Section 7 presents the Rotten Firm theorem, illustrates it with piece-wise linear interpersonal preferences, and calibrates what proportion of laboratory participants exhibit "sufficiently-kinked" preferences. Section 8 explains the relationship between the Rotten Firm theorem and the Rotten Kid theorem. Section 9 concludes by discussing possible extensions of the analysis. Because the proofs help give intuition for the results, I include them in the text rather than in an appendix.

2 The Gift-Exchange Game

There are two players, a firm and a worker. The firm can either offer a wage w to the worker or instead choose not to employ the worker. If the worker accepts the wage offer, then the worker chooses effort e. The firm's profit, or **material payoff function**, is

$$\pi^F(w,e) = e - w. \tag{1}$$

The firm maximizes its material payoff. The worker's material payoff function is

$$\pi^W(w,e) = v(w) - c(e). \tag{2}$$

The function $v(\cdot)$ reflects concave benefits of having more money: $v(0) = 0$, $v' > 0$, $v'' < 0$; and $c(\cdot)$ reflects convex costs of exerting greater effort: $c(0) = 0$, $c' > 0$, $c'' > 0$, $\lim_{e\to\infty} c'(e) = \infty$. I assume that $v'(0) > c'(0)$ so that there are potential gains from trade.

Following a common modeling strategy in the literature on altruism (e.g., Barro 1974, Bergstrom 1989), these material payoff functions represent the purely self-centered aspect of payoffs, but the worker maximizes utility $U\left(\pi^W, \pi^F\right)$ (described in the next section), which may include the firm's as well as his own material payoff. If the firm does not offer employment to the worker, or if the worker refuses employment, then both players receive outside option material payoff 0, and the worker earns utility $U(0,0) = \overline{U}$. If instead trade occurs, call the wage-effort pair (w,e) a **transaction**. The solution concept is subgame-perfect Nash equilibrium. If the firm is indifferent about employing the worker, I assume as a tie-breaker that the firm does not offer employment.

Unlike in a typical principal-agent problem, the firm cannot make the wage depend on output in this gift-exchange game. Hence the worker has no extrinsic incentive to exert effort. If the worker were purely selfish, with utility function $U \equiv \pi^W(w,e)$, there would be no exchange. The firm would prefer its outside option because it could not guarantee that the worker would exert any effort. This stark case makes clear the implications of the worker's preference for fair transactions because any employment that occurs in equlibrium is the result of the worker's interpersonal preferences.

3 Interpersonal Preferences

Models designed to explain laboratory behavior write interpersonal preferences $U\left(x^W, x^F\right)$ 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, wage and effort), I instead make utility $U\left(\pi^W, \pi^F\right)$ depend on the material payoffs π^W and π^F from the transaction. If the material payoff functions are quasi-linear in money, then the $U\left(\pi^W, \pi^F\right)$ specification specializes to the $U\left(x^W, x^F\right)$ model in the laboratory, where money is the only relevant commodity. I assume throughout that $U\left(\pi^W, \pi^F\right)$ is continuous.

In consumer theory, it is usual to assume that utility is monotonically increasing in all its arguments. In the context of a concern for fairness, it is important to allow interpersonal preferences to be non-monotonic. For example, if the firm's material payoff is much larger than the worker's, the worker might prefer that the firm have a lower material payoff. I define a new condition, joint-monotonicity, that appropriately weakens monotonicity.

Definition 1 *The preferences U exhibit **joint-monotonicity** if for any $\left(\pi^W, \pi^F\right)$ and $\varepsilon > 0$, there is some $\left(\pi^{W\prime}, \pi^{F\prime}\right)$ with $\pi^{W\prime} > \pi^W$ and $\pi^{F\prime} > \pi^F$ such that $\left\|\left(\pi^{W\prime}, \pi^{F\prime}\right) - \left(\pi^W, \pi^F\right)\right\| < \varepsilon$ and $U\left(\pi^{W\prime}, \pi^{F\prime}\right) > U\left(\pi^W, \pi^F\right)$.*

Joint-monotonicity states that for any material payoff pair, there is always some very close alternative material payoff pair giving more to *both* players that the worker strictly prefers. It differs from local non-satiation in requiring that it is always possible to find a more-preferred allocation in a particular direction, a direction which jointly increases both players' material payoffs. Joint-monotonicity is equivalent to requiring that, at any material payoff pair, the utility function is not decreasing in both arguments.

I assume that the interpersonal preferences are quasi-concave, a standard condition that implies convex indifference curves.

Definition 2 *The preferences U exhibit **weak quasi-concavity** if for any $\left(\pi^W, \pi^F\right)$ and $\left(\pi^{W\prime}, \pi^{F\prime}\right)$ such that $U\left(\pi^W, \pi^F\right) \leq U\left(\pi^{W\prime}, \pi^{F\prime}\right)$, $U\left(\pi^W, \pi^F\right) \leq U\left(\lambda\pi^W + (1-\lambda)\,\pi^{W\prime}, \lambda\pi^F + (1-\lambda)\,\pi^{F\prime}\right)$ for any $\lambda \in [0,1]$. **Strict quasi-concavity** replaces the final weak inequality with a strict inequality.*

Figure 1 illustrates indifference curves from utility functions satisfying joint-monotonicity and quasi-concavity.

3.1 Second-order unfairness aversion

For interpersonal preferences, quasi-concavity captures a kind of concern for fairness. Quasi-concavity means that the worker's and the firm's material payoffs enter the worker's utility function as complements. The marginal rate of substitution varies with the level of the material payoffs. Holding fixed the worker's material payoff, the higher the firm's material payoff, the less material payoff the worker is willing to give up to increase the firm's profit. In fact, if the worker's interpersonal preferences are non-monotonic and the firm's profit is sufficiently high, the worker may be willing to reduce his own payoff to hurt the firm.

Definition 3 *The preferences U exhibit* ***second-order unfairness aversion*** *if U is jointly-monotonic, strictly quasi-concave, and smooth (infinitely-differentiable).*

This kind of concern for fairness is "second-order" in the sense that it arises from the smooth curvature of the utility function. Examples of preferences exhibiting second-order unfairness aversion include Bolton & Ockenfels (2000) and the constant-elasticity-of-substitution functional form (used in, *inter alia*, Cox, Friedman, & Sadiraj 2005), as well as Becker's (1974) altruistic preferences. By contrast, "simple altruism" $U = \pi^W + \gamma\pi^F$, in which the other player's material payoff enters utility linearly, does not exhibit any kind of unfairness aversion. Altruism can be understood as the slope of the interpersonal indifference curves (a first-order property), while the change in the slope as material payoffs vary corresponds to concern for fairness.

3.2 First-order unfairness aversion

In laboratory games, people often choose to split money equally between themselves and other players. Consider dictator games (Forsythe et al 1994), in which one player unilaterally decides how to split a pie between himself and another player. A typical finding is that a disproportionate fraction of participants choose to divide the pie exactly evenly between the two players (Camerer 2003). This equal-split behavior is difficult to explain with smooth interpersonal preferences unless the participant's utility function treats the two players symmetrically, putting equal weight on the players' material payoffs. Yet many of these same people 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). A smooth, symmetric interpersonal utility function that explains equal-split behavior in symmetric dictator games would not predict equal-split behavior in this case. Adherence to equal-split norms is often attributed to a concern for fairness, but second-order unfairness aversion cannot generate this behavior.

Instead, a substantial minority of people appear to behave according to a "fairness rule."

Definition 4 *A **fairness rule** g is a continuous, strictly increasing function that maps each possible material payoff for the worker into a material payoff for the firm.*

The **equal-split fairness rule** is the identity function $g(x) = x$, which assigns equal material payoffs to both players. The definition of a fairness rule is more general than equal splits only, allowing for the possibility that the fairness rule may be non-linear and biased in a self-serving way.

A fairness rule partitions the space of material payoffs into two regions. The northwest region corresponds to **disadvantageously unfair** transactions for the worker, where the firm's material payoff is higher and the worker's material payoff is lower than dictated by the fairness rule. The southeast region corresponds to **advantageously unfair** transactions for the worker. Psychologically, adherence to a salient fairness rule may be the result of learned social norms. Economically, fairness-rule behavior can be modeled as a path of kinks in the utility function along the fairness rule. If the preferences put appreciably more weight on the worker's material payoff (relative to the firm's) at a transaction that is disadvantageously unfair than at a transaction that is advantageously unfair, even when those two transactions are very close, then I call the preferences (locally) "first-order unfairness-averse."

Definition 5 *Suppose U is jointly-monotonic and weakly quasi-concave, and for some fairness rule g, $\left(\pi^W, \pi^F\right)$ satisfies $\pi^F = g\left(\pi^W\right)$. The preferences U exhibit **first-order unfairness aversion about** $\left(\pi^W, \pi^F\right)$ if for all $\left(\widehat{\pi}^W, \widehat{\pi}^F\right)$ in a neighborhood of $\left(\pi^W, \pi^F\right)$,*

$$\frac{\lim_{\widehat{\pi}^W \uparrow \pi^W} \partial U\left(\pi^W, \widehat{\pi}^F\right)/\partial \pi^W}{\lim_{\widehat{\pi}^F \downarrow \pi^F} \partial U\left(\widehat{\pi}^W, \pi^F\right)/\partial \pi^F} > \frac{\lim_{\widehat{\pi}^W \downarrow \pi^W} \partial U\left(\pi^W, \widehat{\pi}^F\right)/\partial \pi^W}{\lim_{\widehat{\pi}^F \uparrow \pi^F} \partial U\left(\widehat{\pi}^W, \pi^F\right)/\partial \pi^F}.$$

*The preferences U exhibit **first-order unfairness aversion about fairness rule** g if they exhibit first-order unfairness aversion about all $\left(\pi^W, \pi^F\right)$ satisfying $\pi^F = g\left(\pi^W\right)$.*

Preferences that are first-order unfairness-averse will cause the player to make choices that adhere to the fairness rule in a variety of situations. First-order unfairness aversion can be understood as a limit case of second-order unfairness aversion as the degree of convexity becomes large. The distinction between second-order and first-order unfairness aversion is analogous to the distinction between second-order and first-order risk aversion (Segal & Spivak 1990). Figure 1c illustrates indifference curves for preferences that are first-order unfairness-averse about a fairness rule.

The most widely-applied fairness model, Fehr & Schmidt's (1999) inequity-aversion, exhibits first-order unfairness aversion about the equal-split fairness rule:

$$U = \pi^W - \alpha \max\left\{\pi^F - \pi^W, 0\right\} - \beta \max\left\{\pi^W - \pi^F, 0\right\}$$

$(\alpha, \beta \geq 0)$, as do the Rawlsian "social welfare preferences" advocated by Charness & Rabin (2002):

$$U = \pi^W + \gamma\pi^F + \delta \min\left\{\pi^W, \pi^F\right\}$$

$(\gamma, \delta \geq 0)$. The papers that introduced these models did not emphasize the kink. However, the kink will turn out to be central for generating rule-based behavior and therefore efficient equilibrium.

4 Efficiency With Interpersonal Preferences

The analysis that follows addresses whether the equilibrium of the gift-exchange game is Pareto efficient. This section clarifies what Pareto efficiency means when the worker has interpersonal preferences.

Recall that an exchange is Pareto efficient if it makes both parties at least as well off as any alternative exchange could have.

Definition 6 *A transaction* (w, e) *is* ***Pareto efficient*** *if there is no other transaction* $(\widehat{w}, \widehat{e})$ *such that* $\pi^F(\widehat{w}, \widehat{e}) \geq \pi^F(w, e)$ *and* $U(\widehat{w}, \widehat{e}) \geq U(w, e)$*, at least one inequality strict.*

Let P denote the set of Pareto efficient transactions.

Economists usually assume that both parties to an exchange are purely selfish, seeking to maximize their material payoffs. That scenario corresponds to the special case where $U(w, e) \equiv \pi^W(w, e)$. A transaction that *would be* Pareto efficient if the worker were selfish (rather than also concerned about the firm's material payoff) is called materially-efficient.

Definition 7 *A transaction* (w, e) *is* ***materially-efficient*** *if there is no other transaction* $(\widehat{w}, \widehat{e})$ *such that* $\pi^F(\widehat{w}, \widehat{e}) \geq \pi^F(w, e)$ *and* $\pi^W(\widehat{w}, \widehat{e}) \geq \pi^W(w, e)$*, at least one inequality strict.*

Let M denote the set of materially-efficient transactions.

In general, a transaction is materially-efficient if and only if it equates the marginal rates of substitution of the players' material payoffs,

$$\frac{\partial \pi^F(w, e)/\partial w}{\partial \pi^F(w, e)/\partial e} = \frac{\partial \pi^W(w, e)/\partial w}{\partial \pi^W(w, e)/\partial e}. \tag{3}$$

It would be useful to find a simple characterization of Pareto efficient transactions. This might seem unlikely at first blush since the worker's utility function could be a complicated function of the material payoffs, possibly kinked. Nonetheless, it turns out that (3) remains a necessary condition for Pareto efficiency.

Let $\widetilde{(w,e)} = \arg\max_{(w,e)\in M} U\left(\pi^W(w,e), \pi^F(w,e)\right)$ be called the worker's **favorite materially-efficient transaction**, his most-preferred transaction among the transactions that are materially-efficient.

Proposition 1 *Suppose U exhibits second-order unfairness aversion or first-order unfairness aversion about some fairness rule. A transaction is Pareto efficient if and only if it is materially-efficient and gives the worker a lower material payoff than his favorite materially-efficient transaction gives him:*

$$P = \left\{(w,e) \in M | \pi^W(w,e) \leq \pi^W\widetilde{(w,e)}\right\}.$$

Proof. To see why $P \subset \left\{(w,e) \in M | \pi^W(w,e) \leq \pi^W\widetilde{(w,e)}\right\}$, consider a transaction that is not materially-efficient (and hence lies in the interior of the material payoff possibility set). Joint-monotonicity implies that there is some direction of slightly greater material payoffs for both players that increases the worker's utility. Since both players are better off, the original transaction cannot be Pareto efficient. Now consider a transaction that is materially-efficient but gives the worker a higher material payoff than his favorite materially-efficient transaction gives him. That transaction cannot be Pareto efficient because both players prefer the worker's favorite materially-efficient transaction.

Now we show that $P \supset \left\{(w,e) \in M | \pi^W(w,e) \leq \pi^W\widetilde{(w,e)}\right\}$. First note that the worker's utility is monotonically decreasing along the material-efficiency frontier as transactions move farther away from his favorite materially-efficient transaction. Now suppose there were a transaction that is materially-efficient and that gives the worker a lower material payoff than his favorite materially-efficient transaction gives him, but that were not Pareto efficient. Then there would be some alternative transaction that gives the firm a higher material payoff and gives the worker higher utility. Since the original transaction is materially-efficient, this alternative transaction must give the worker a lower material payoff. If this alternative transaction lies in the interior of the material payoff possibility set, then there is some materially-efficient transaction that gives even higher material payoff to the firm and even higher utility to the worker. But since the worker's utility along the material-efficiency frontier is lower for transactions farther away from his favorite materially-efficient transaction, this materially-efficient transaction must give the worker lower (not higher)

utility than the original transaction, which is a contradiction. ■

Figure 2 illustrates the proposition.

5 The Logic of the Rotten Firm Theorem

Unfairness-averse preferences do *not* imply a preference for materially-efficient allocations.[3] In fact, much of the laboratory evidence for unfairness-averse preferences comes from showing that people choose to allocate resources across individuals in a way that is *materially-inefficient* but more equal. For example, 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. Experimental participants also make materially-inefficient choices when real money is at stake, though less commonly (e.g., Fisman, Kariv, & Markovits 2005; Charness & Rabin 2002). Figure 1c illustrates how non-monotonicity makes it possible to prefer a materially-inefficient allocation.

In a gift-exchange game, however, the worker's adherence to a fairness rule leads to an equilibrium that is materially-efficient. A worker is said to obey a fairness rule if the worker's effort choice as a function of the wage, $e(w)$, causes the material payoffs of the players to co-move in accordance with that fairness rule.

Definition 8 *The worker* ***obeys a fairness rule at*** w *if for some fairness rule* g *and for all* $\widehat{w}$ *in a neighborhood of* w*,*

$$e(\widehat{w}) \;\; \textit{satisfies} \; \pi^F(\widehat{w}, e(\widehat{w})) = g\left(\pi^W(\widehat{w}, e(\widehat{w}))\right).$$

If the worker obeys a fairness rule at w, then $e'(w) = \frac{v'+g'}{c'+g'} > 0$. Intuitively, an increase in the wage increases the worker's material payoff and reduces the firm's. In order to keep the material payoffs moving in tandem as required by the fairness rule, the worker must increase effort, reducing his own material payoff and increasing the firm's.

Because the worker obeys a fairness rule, the firm and worker both gain on the margin from an increase in the gains from trade. This makes it optimal for the firm to maximize the gains from trade by offering the wage that induces an efficient level of effort.

[3] Indeed, Charness & Rabin (2002) and Engelmann & Strobel (2004) criticize Fehr & Schmidt's (1999) inequity-aversion model on the grounds that it does not build in any desire for material efficiency.

Proposition 2 *Suppose the worker obeys a fairness rule at all w. Then the equilibrium of the gift-exchange game is materially-efficient.*

Proof. Because the worker obeys a fairness rule at all w, $\frac{d\pi^F(w,e(w))}{dw} = e'(w) - 1$ and $\frac{d\pi^W(w,e(w))}{dw} = v'(w) - c'(e(w))e'(w)$ have the same sign at all w. Since the set of feasible material payoffs is convex, there is some wage offer w^* that maximizes the firm's profit, and hence also the worker's material payoff. But $\frac{d\pi^F(w^*,e(w^*))}{dw} = 0$ and $\frac{d\pi^W(w^*,e(w^*))}{dw} = 0$ jointly imply that $v'(w^*) = c'(e(w^*))$, the condition for material-efficiency. ■

Note that at any other wage offer, the resulting wage-effort pair would be inefficient. Figure 3 illustrates that the profit-maximizing wage offer occurs at the unique point on the material-efficiency frontier that satisfies the fairness rule.

Although the equilibrium is materially-efficient, determining whether the equilibrium is *Pareto* efficient requires specifying the worker's preferences. If the worker's fairness-rule behavior is the result of maximizing jointly-monotonic, quasi-concave preferences, then the equilibrium is the worker's favorite materially-efficient transaction and is therefore Pareto efficient.

The worker's adherence to a fairness rule aligns the firm's incentives with the worker's, causing the equilibrium to be efficient. The next two sections show that the worker obeys a fairness rule in the neighborhood of a materially-efficient transaction only if the worker's preferences exhibit first-order unfairness aversion.

6 Second-Order Unfairness Aversion: An Impossibility Result

If the worker's interpersonal preferences exhibit second-order unfairness aversion, then the equilibrium cannot be materially-efficient. The problem is that at a materially-efficient transaction, the worker will be "too altruistic." The firm can profitably deviate by offering a lower wage. The worker is willing to choose effort to allow the firm to earn a greater material payoff at the worker's expense, as long as the firm's gain is large enough relative to the worker's loss. Due to the convexity of the cost of effort, this is possible at the materially-efficient transaction.

To be more precise, consider a materially-efficient transaction $\left(w^{\text{eff}}, e^{\text{eff}}\right)$. If the worker's effort choice is optimal, then the worker is indifferent to a marginal change in effort. That means the worker is indifferent on the margin between increasing own material payoff by c' and increasing the firm's material payoff by 1. Now suppose the firm deviated to a slightly lower wage $\widehat{w} < w^{\text{eff}}$, which induces effort $\widehat{e} < e^{\text{eff}}$. (If a lower wage induced greater effort, then the firm could not have been maximizing.) Because the cost of effort function $c(\cdot)$ is convex, it is now possible – and indeed

optimal – for the worker's effort $\widehat{e}$ to increase the firm's payoff by 1 while reducing own payoff by less than c'. Since the firm's profit increases, the firm will prefer offering $\widehat{w}$ to offering w^{eff}. Figure 4 shows this logic graphically.

Theorem 1 *Suppose U exhibits second-order unfairness aversion. Then there is no equilibrium transaction that is materially-efficient.*

Proof. Given the worker's effort function $e(w)$, the equilibrium is characterized by the wage that solves the firm's first-order condition, $\frac{d\pi^F(w,e(w))}{dw} = 0$. This derivative can be written as the sum of two terms:

$$\frac{d\pi^F(w,e(w))}{dw} = \frac{\partial\pi^F\left(w,\pi^W\right)}{\partial\pi^W}\frac{d\pi^W(w,e(w))}{dw} + \frac{\partial\pi^F\left(w,\pi^W\right)}{\partial w}.$$

The second term is the effect of a change in wage on the firm's material payoff, *holding π^W constant* (that is, effort is adjusted to hold π^W constant). It can be interpreted as the effect of an incremental change in wage on the "size of the pie" that is to be split between the worker and the firm. In this equation, the "size of the pie" is measured in units of the firm's material payoff. The first term is the effect of an incremental change in wage on the worker's material payoff, multiplied by how much the firm's material payoff comoves with the worker's material payoff. That is, the first term can be interpreted as the effect of a marginal change in wage on how the pie is split.

The theorem says that $\frac{d\pi^F(w,e(w))}{dw}$ cannot equal zero at a materially-efficient transaction. At a materially-efficient transaction, the second term, $\frac{\partial\pi^F\left(w,\pi^W\right)}{\partial w}$, equals zero (an envelope condition). Note that $\frac{\partial\pi^F\left(w,\pi^W\right)}{\partial\pi^W} < 0$ because a change in effort causes the worker's material payoff and the firm's material payoff to move in opposite directions. The proof of the theorem amounts to showing that $\frac{d\pi^W(w,e(w))}{dw}$, the marginal effect of an increase in wage on the worker's material payoff, is positive at an efficient transaction.

The worker's effort $e(w)$ solves the first-order condition $-U_1c'(e)+U_2 = 0$, and both U_1 and U_2 are positive at $(w,e(w))$. Let H denote the determinant of the bordered Hessian matrix of the utility function, and $\widetilde{H} \equiv H/U_2^3$ is a transformation of H that is invariant to monotonic transformations of U. A straightforward calculation reveals that at a materially-efficient transaction,

$$\frac{d\pi^W}{dw} = \frac{v'}{1-\frac{\widetilde{H}}{c''}}. \tag{4}$$

Since $\widetilde{H} < 0$, it follows that $0 < \frac{d\pi^W}{dw} < v'$. When the firm increases the wage by a dollar, the worker's material payoff increases by less than the marginal value of a dollar because the worker

increases effort. It follows that $\frac{d\pi^F\left(w^{\text{eff}},e\left(w^{\text{eff}}\right)\right)}{dw} < 0$, which means the firm would be better off offering a slightly lower wage. ■

Note that the transformed determinant of the bordered Hessian $\widetilde{H}$ in equation (4) is a measure of how convex the worker's indifference curves are in a neighborhood of the materially-efficient transaction. Roughly speaking, it measures the complementarity between the firm's and worker's material payoffs in the worker's preferences.[4] The larger is $\widetilde{H}$, the smaller is $\frac{d\pi^W}{dw}$ (because the worker adjusts effort more in response to a wage change), so the smaller is $\frac{d\pi^F(w,e(w))}{dw}$ also. The limit case $\widetilde{H} \to -\infty$ corresponds to first-order unfairness aversion. The fact that $\frac{d\pi^F\left(w^{\text{eff}},e\left(w^{\text{eff}}\right)\right)}{dw} \to 0$ as $\widetilde{H} \to -\infty$ suggests it is possible that the equilibrium transaction may be materially-efficient if the worker's interpersonal preferences are kinked. Indeed, the next section shows that if the worker's interpersonal preferences are sufficiently-kinked, then the equilibrium is efficient.

7 Sufficiently-Kinked Interpersonal Preferences Imply Efficiency: The Rotten Firm Theorem

7.1 The Rotten Firm Theorem

Although second-order unfairness aversion cannot generate efficient exchange, first-order unfairness aversion can if the kink is sufficiently large. The reason is that, if the worker's interpersonal preferences are sufficiently kinked about a materially-efficient transaction, then the worker follows a fairness rule (at least in a neighborhood of that transaction). As in Section 5, when the worker chooses effort according to a fairness rule, it is optimal for the firm to offer a wage that leads to an efficient transaction. Because the worker's and the firm's material payoffs increase in tandem as the firm offers a wage up to the efficient point, in equilibrium the firm offers the wage that maximizes both players' material payoffs.

First-order unfairness aversion about a materially-efficient transaction is a necessary condition for the equilibrium of the gift-exchange game to be Pareto efficient, but it is not sufficient. Even if the worker is choosing effort in accordance with a fairness rule in a neighborhood of an efficient transaction, the worker's optimal effort choice at lower wages may not follow a fairness rule. In that case, the firm may be able to earn higher profit by offering a lower wage. Figure 5a illustrates how a materially-efficient transaction may be locally optimal for the firm without being globally optimal.

[4] Of course, $\widetilde{H}$ is a local measure, not a global measure, of complementarity. The proceeding "comparative statics" on $\widetilde{H}$ should therefore be taken loosely.

A global condition that plays a role in ensuring that the equilibrium is efficient is that the firm's material payoff enters the worker's interpersonal preferences as a "normal good." This is the "normal good" assumption from standard consumer theory (defined for a jointly-monotonic, quasi-concave U): For any fixed "price" $p > 0$ of the firm's material payoff in terms of the worker's material payoff, $\pi^F(I;p)$ is strictly increasing in I, where $\pi^F(I;p)$ is that value $\widetilde{\pi}^F$ that solves $\left(\widetilde{\pi}^W, \widetilde{\pi}^F\right) = \arg\max_{\{(\pi^W,\pi^F):\pi^W+p\pi^F\leq I\}} U\left(\pi^W, \pi^F\right)$. The normal-good assumption is less natural here than in the consumer theory context because, in the gift-exchange game, the worker does not actually face tradeoffs that are linear in material payoffs. The assumption is perhaps best understood as a weak regularity condition on the utility function, a condition implied by stronger assumptions such as homotheticity (which holds with the piecewise-linear preferences considered later in this section).

With the assumption that π^F enters U as a normal good, it is possible to be precise about how kinked the worker's interpersonal preferences must be for the equilibrium to be efficient.

Theorem 2 *(Rotten Firm Theorem) Suppose U is jointly-monotonic and weakly quasi-concave.*

1. *If the equilibrium transaction is materially-efficient, then the worker's interpersonal preferences are first-order unfairness averse about the equilibrium transaction.*

2. *Let $\left(w^{eff}, e^{eff}\right)$ be a materially-efficient transaction with associated material payoffs $\left(\pi^{W*}, \pi^{F*}\right)$, and let $(\widehat{w}, \widehat{e})$ be the transaction such that $\pi^F(\widehat{w}, \widehat{e}) = \pi^{F*}$ and $U\left(\pi^W(\widehat{w}, \widehat{e}), \pi^F(\widehat{w}, \widehat{e})\right) = \overline{U}$. If π^F enters U as a normal good, and if the worker's interpersonal preferences are sufficiently-kinked in the sense that*

$$\frac{\lim_{\pi^W \downarrow \pi^{W*}} \partial U\left(\pi^W, \pi^{F*}\right)/\partial \pi^W}{\lim_{\pi^F \uparrow \pi^{F*}} \partial U\left(\pi^{W*}, \pi^F\right)/\partial \pi^F} \leq \frac{1}{c'(\widehat{e})} < \frac{1}{c'\left(e^{eff}\right)} \leq \frac{\lim_{\pi^W \uparrow \pi^{W*}} \partial U\left(\pi^W, \pi^{F*}\right)/\partial \pi^W}{\lim_{\pi^F \downarrow \pi^{F*}} \partial U\left(\pi^{W*}, \pi^F\right)/\partial \pi^F},$$

then $\left(w^{eff}, e^{eff}\right)$ is the equilibrium transaction, and it is Pareto efficient.

The Rotten Firm theorem shows that a concern for fairness generates efficient exchange only if the interpersonal preferences are sufficiently kinked that the worker follows a fairness rule in a neighborhood of a materially-efficient transaction. Figure 5b illustrates the result.

It is important to recognize that equilibrium efficiency is *not* a consequence of the worker caring directly about material efficiency. Rather, equilibrium efficiency emerges from the firm's optimal reaction to the worker's fairness-rule behavior. At any other wage offer, the resulting wage-effort pair would be inefficient.

Contrast the Rotten Firm theorem equilibrium with the equilibrium that would occur if enforceable contracts were available. If the firm could make a take-it-or-leave-it offer to the worker, the firm would offer the materially-efficient transaction that gives the worker exactly his outside option level of utility $\overline{U}$ (see Figure 5b). If the worker could make a take-it-or-leave-it offer to the firm, the worker would offer his favorite materially-efficient transaction – which is exactly the equilibrium in Theorem 2. More generally, if the worker and the firm each have some bargaining power, the Coase theorem implies that the equilibrium with enforceable contracts will be materially-efficient and give the worker an intermediate level of utility. Hence *in terms of efficiency* Theorem 2 implies that strong enough first-order unfairness aversion can fully substitute for the availability of binding contracts. However, the firm would prefer the equilibrium with enforceable contracts because relying on fairness requires sharing more of the gains from trade with the worker.

7.2 Example: Piece-Wise Linear Preferences

In this subsection, I solve for the equilibrium of the gift-exchange game using a general piece-linear specification of interpersonal preferences. This example is of special interest for several reasons. For one thing, the parameters are psychologically-interpretable, which helps develop intuition for what the kink means. Moreover, the specification embeds important leading models of fairness preferences as special cases. As a result, the preferences can be calibrated based on existing parameter estimates from laboratory data, which allows for a judgment of whether "sufficiently-kinked" is extreme or is instead an empirically-plausible preference.

Specifically, suppose the worker's preferences take the form

$$U\left(\pi^W, \pi^F\right) = \pi^W + f\left(\pi^W, \pi^F\right) + \gamma\pi^F, \tag{5}$$

where the "fairness function" f is given by

$$f\left(\pi^W, \pi^F\right) = -\alpha \max\left\{\pi^F - \pi^W, 0\right\} - \beta \max\left\{\pi^W - \pi^F, 0\right\}. \tag{6}$$

The fairness function represents the purely fair-minded component of interpersonal preferences, adherence to an equal-split fairness rule. The fairness function takes a maximum value of zero when $\pi^W = \pi^F$. A transaction that gives equal material payoffs is called a **fair transaction**.

The three preference parameters have a straightforward interpretation.[5] The first term of (6), parameterized by $\alpha \geq 0$, captures **aversion to disadvantageous unfairness** for the worker: it

[5] Although it would be possible to parameterize utility function (5) more parsimoniously, with two parameters rather than three, the interpretation would be less clear.

is unfair when the firm's material payoff exceeds the worker's. It is also unfair when the worker's material payoff exceeds the firm's. The second term captures this **aversion to advantageous unfairness** for the worker, parameterized by $\beta \geq 0$. Altruism is parameterized by $0 \leq \gamma \leq 1$. As special cases, the utility function (5) embeds inequity-aversion (Fehr & Schmidt 1999) as well as "Rawlsian" fairness preferences, which have been advocated as a good descriptive model of laboratory behavior by several authors (Charness & Rabin 2002, Engelmann & Strobel 2004).[6]

7.2.1 Worker's effort choice

In describing how the worker's actual choice of effort depends on the wage, it is useful first to consider the "fair" choice of effort:

$$e^{\text{fair}}(w) \equiv \arg\max_{e} f\left(\pi^W(w,e), \pi^F(w,e)\right).$$

The fair choice of effort $e^{\text{fair}}(w)$ satisfies the equal-split fairness rule

$$\pi^W\left(w, e^{\text{fair}}\right) = \pi^F\left(w, e^{\text{fair}}\right), \tag{7}$$

equating the worker's material payoff with the firm's material payoff. It follows that $\frac{de^{\text{fair}}(w)}{dw} > 0$. The fair effort level is strictly increasing in the wage because, all else equal, an increase in the wage reduces the firm's material payoff and raises the worker's. Maintaining equal material payoffs requires that the worker make a transfer back to the firm by increasing effort.

In choosing how much effort to exert, the worker takes into account his own material payoff, his concern for a fair transaction, and his altruism:

$$e(w) \equiv \arg\max_{e} \pi^W(w,e) + f\left(\pi^W(w,e), \pi^F(w,e)\right) + \gamma\pi^F(w,e). \tag{8}$$

The worker's most-preferred choice of effort turns out to be equal to the most fair level of effort, above some **reciprocity lower bound** $\underline{e}$ and up to some **reciprocity upper bound** $\overline{e}$. That is, the worker fairly reciprocates a higher wage with higher effort, except that the worker never works harder than $\overline{e}$. Moreover, even if the firm offers a very low wage, the worker will never exert less effort than $\underline{e}$.

[6] If $\gamma = 0$, then the preferences are inequity-averse. If $\alpha = 0$, then utility function (5) can be written as

$$U\left(\pi^W, \pi^F\right) = \pi^W + \beta \min\left\{\pi^W, \pi^F\right\} + (\gamma - \beta)\pi^F.$$

Lemma 1 *Suppose the worker has piece-wise linear interpersonal preferences (5). There are some reciprocity upper bound* $\overline{e}(\beta,\gamma)$ *and lower bound* $\underline{e}(\alpha,\gamma)$ *such that*

$$e(w)=\begin{cases}\overline{e} & \text{if } w>\overline{w}\\ e^{fair}(w) & \text{if } \underline{w}\leq w\leq\overline{w}\\ \underline{e} & \text{if } w<\underline{w}\end{cases}$$

where $\overline{w}\equiv\left(e^{fair}\right)^{-1}(\overline{e})$ *and* $\underline{w}\equiv\left(e^{fair}\right)^{-1}(\underline{e})$*. Moreover,* $\overline{e}(\beta,\gamma)$ *is increasing in* β *and* γ*, while* $\underline{e}(\alpha,\gamma)$ *is decreasing in* α *and increasing in* γ*.*

Proof. Note that for a given w, the worker's utility function is concave in e. The worker's marginal utility gain from shirking on effort is the gain in his material payoff, $c'(e)$. The marginal utility cost, $\beta(c'(e)+1)+\gamma$, is that the transaction becomes (advantageously) unfair and hurts the firm. The reciprocity upper bound is defined by $c'(\overline{e})=\beta(c'(\overline{e})+1)+\gamma$. The worker's marginal utility gain from exerting extra effort is the altruism-weighted gain in the firm's material payoff, γ. The marginal utility cost, $c'(e)+\alpha(1+c'(e))$, is that the transaction becomes (disadvantageously) unfair. The reciprocity lower bound is defined by $\gamma=c'(\underline{e})+\alpha(1+c'(\underline{e}))$. For wages between $\underline{w}$ and $\overline{w}$, the marginal utility cost of deviating from the fair level of effort exceeds the marginal utility benefit. It is easy to check that $\overline{e}(\beta,\gamma)$ is increasing in β and γ, while $\underline{e}(\alpha,\gamma)$ is decreasing in α and increasing in γ. ■

The greater is the worker's altruism γ, the higher will be both the reciprocity lower bound and upper bound. The reciprocity lower bound depends on the worker's aversion to disadvantageous unfairness α, while the reciprocity upper bound depends on the worker's aversion to advantageous unfairness β.

For wages between $\underline{w}$ and $\overline{w}$, the worker behaves in accordance with the equal-split fairness rule, $e^{\text{fair}}(w)$. There is exactly one wage offer that induces a materially-efficient wage-effort pair. Call the wage w^{eff} and the corresponding effort $e^{\text{eff}}=e^{\text{fair}}\left(w^{\text{eff}}\right)$.

7.2.2 Firm's optimal wage offer

If e^{eff} lies between the reciprocity lower bound and the reciprocity upper bound, then it is possible for the firm to induce e^{eff}. Doing so may be optimal, but there is another possibility. If the altruism parameter γ is large, the firm may instead earn higher profit from inducing the reciprocity lower bound $\underline{e}$. Since the firm can get the worker to exert $\underline{e}$ for any wage smaller than $\underline{w}$, the firm would want to offer the lowest wage such that the worker still accepts employment. In this case of "too

much altruism," the worker earns exactly his outside option level of utility, and the equilibrium is not in general materially-efficient.

However, if the worker is sufficiently averse to both advantageous and disadvantageous unfairness, then the equilibrium is the Pareto efficient one.

Theorem 3 *(Rotten Firm theorem for piece-wise linear interpersonal preferences) Suppose the worker has piece-wise linear interpersonal preferences (5). Fix* γ, $v(\cdot)$, *and* $c(\cdot)$. *There exist* $\widehat{\widehat{\alpha}} < 1$ *and* $\widehat{\widehat{\beta}} < 1$ *such that if* $\alpha \geq \widehat{\widehat{\alpha}}$ *and* $\beta \geq \widehat{\widehat{\beta}}$, *then the equilibrium wage-effort pair is Pareto efficient.*

Proof. $e^{\text{fair}}(w) - w$ is strictly concave in w and maximized at wage w^{eff}. Define $\widehat{\widehat{\beta}} < 1$ by $\overline{e}\left(\widehat{\widehat{\beta}}, \gamma\right) = e^{\text{eff}}$. Define $\widehat{\widehat{\alpha}}'$ by $\underline{e}\left(\widehat{\widehat{\alpha}}', \gamma\right) = e^{\text{eff}}$. We will need $\alpha \geq \widehat{\widehat{\alpha}}'$ in order for e^{eff} to be a feasible effort level for the firm to induce. But we will also need that the firm earns greater profit from the fair and efficient transaction than from the lowest wage that induces $\underline{e}$. Define $w^{\min}(\alpha)$ as the lowest wage such that $U\left(w^{\min}, \underline{e}(\alpha, \gamma)\right) = \overline{U}$. Since $w^{\min}(\alpha)$ is increasing in α, and $\underline{e}(a, \gamma)$ is decreasing in α, there is some $\widehat{\widehat{\alpha}}''$ beyond which the firm's profit from the low wage, $\underline{e}(\alpha, \gamma) - w^{\min}(\alpha)$, is smaller than the firm's profit from the fair and efficient wage, $e^{\text{eff}} - w^{\text{eff}}$. Let $\widehat{\widehat{\alpha}} = \max\left\{\widehat{\widehat{\alpha}}', \widehat{\widehat{\alpha}}''\right\}$. If $\alpha \geq \widehat{\widehat{\alpha}}$ and $\beta \geq \widehat{\widehat{\beta}}$, then the equilibrium transaction is $\left(w^{\text{eff}}, e^{\text{eff}}\right)$. ■

It is important that the worker be sufficiently averse to advantageous unfairness so that he is not tempted to shirk at the equilibrium transaction. It is important that the worker be sufficiently averse to disadvantageous unfairness for two reasons: so that he is not tempted to exert too much effort (out of altruism) at the equilibrium transaction, and so that the firm is not tempted to exploit his altruism by offering a very low wage (as in Figure 5a).[7]

7.3 Calibrating the Theorem

The Rotten Firm theorem states that gift-exchange is efficient if the worker's interpersonal preferences are sufficiently-kinked. But how much is sufficiently? In real-world settings, for what proportion of individuals is it true that $\alpha \geq \widehat{\widehat{\alpha}}$ and $\beta \geq \widehat{\widehat{\beta}}$? Using existing parameter estimates, this subsection provides a very rough calibration of the piecewise-linear interpersonal preferences from the previous subsection.

[7] Note that in the pure inequity-aversion model ($\gamma = 0$), $\underline{e} = -\infty$, so $\widehat{\widehat{\alpha}} = \widehat{\widehat{\alpha}}'$ (the firm is never tempted to offer a wage that gives the worker exactly $\overline{U}$). By contrast, if the worker's altruism γ is large relative to his aversion to disadvantageous unfairness α, then the firm will exploit the worker's altruism by offering a low wage. The equilibrium will be generically inefficient, and the worker will earn utility $\overline{U}$. Since altruism toward the firm is equivalent to a direct concern for material efficiency, these results underscore the fact that the Rotten Firm theorem is *not* a consequence of the worker caring directly about material efficiency.

In general, calibrating the model requires knowledge of the worker's cost-of-effort function $c(\cdot)$ in addition to the interpersonal preference parameters α, β, and γ.[8] To eliminate the need to know $c(\cdot)$, I assume the worker's material payoff function is quasi-linear in money[9] (i.e., $v(w) = w$):

$$\pi^W(w, e) = w - c(e).$$

As a result, $c'\left(e^{\text{eff}}\right) = 1$ at the efficient level of effort. Furthermore, since Fehr & Schmidt's (1999) model assumes that $\gamma = 0$ and since Charness & Rabin (2002, Table VI, line 5) estimate that $\gamma \approx 0$, I will assume $\gamma = 0$ for the purposes of this calculation.

Under these assumptions, the threshold level of aversion to disadvantageous inequality is $\widehat{\widehat{\alpha}} = -\frac{1}{2}$, and the threshold level of aversion to advantageous inequality is $\widehat{\widehat{\beta}} = \frac{1}{2}$. Charness & Rabin (2002, Table VI, line 5) estimate the population average $\alpha \approx 0$ and $\beta \approx 0.4$ (both with standard errors around 0.02). Although these estimates imply that $\beta < \frac{1}{2}$ on average, they are consistent with $\beta \geq \frac{1}{2}$ for a sizeable minority of the population since there is substantial behavioral heterogeneity.

Based on a different set of experiments, Fehr & Schmidt (1999, Table III and p.864) assume that $\alpha \geq 0$ and estimate β. As a rough calibration to fit behavior in a wide variety of experimental games, they argue that

$$\beta = \begin{cases} 0 & \text{for about 30\% of individuals} \\ 0.25 & \text{for about 30\% of individuals} \\ 0.6 & \text{for about 40\% of individuals} \end{cases}$$

These estimates suggest that $\beta \geq \frac{1}{2}$ for about 40% of the subject population. However, it is important to keep in mind that this calculation is extremely rough.[10]

7.4 Interpreting the Rotten Firm Theorem

On one view, the Rotten Firm theorem demonstrates that a concern for fairness can fully substitute for the availability of binding contracts. The reasonableness of that view depends on whether people actually have sufficiently-kinked interpersonal preferences. The calibration in the last subsection suggests that *if* people have kinked preferences, those preferences are sufficiently-kinked for a substantial minority of individuals.[11] But do individuals actually have kinked interpersonal

[8] In particular, $\widehat{\widehat{\alpha}}' = \frac{\gamma - c'\left(e^{\text{eff}}\right)}{1 + c'\left(e^{\text{eff}}\right)}$ and $\widehat{\widehat{\beta}} = \frac{c'\left(e^{\text{eff}}\right) - \gamma}{c'\left(e^{\text{eff}}\right) + 1}$.

[9] The exact same calibration would follow if I assumed instead that the worker's material payoff function were quasi-linear in effort: $c(e) = e$, which would also imply $c'\left(e^{\text{eff}}\right) = 1$.

[10] For example, see Shaked (2005).

[11] Of course, that conclusion depends on the assumption (maintained throughout) that the worker's interpersonal preferences are common knowledge, an assumption whose reasonableness depends on context. Weakening that assumption in the formal analysis would require additional structure on the preferences to allow for interpersonal preferences under uncertainty.

preferences?

In laboratory settings where it is clear how to split the surplus equally, many individuals choose to do so (Fehr & Schmidt 1999, Charness & Rabin 2002), consistent with interpersonal preferences that are kinked around an "equal-split fairness rule." However, fairness-rule behavior may be less common in field settings than in the laboratory because the relevant fairness rule may be less clear. Even if a worker's preferences exhibit first-order unfairness aversion, uncertainty about the firm's profit function may smooth out the kinks. In those cases, the results in this paper imply that a concern for fairness cannot generate fully efficient exchange.

On the other hand, market wage and effort levels, other workers' terms of employment, and prior experience may set reference points for what is fair (Kahneman, Knetsch, & Thaler 1986). Similarly, a contract (even if unenforceable) may play an important role in setting expectations about what is fair (Hart & Moore 2006). Although not explicitly modeled in this paper, such reference points probably underlie fairness rules in many real-world settings. Hence the extent to which fairness preferences enable efficient exchange may depend on the salience of a relevant fairness rule.

8 Relationship Between the Rotten Firm and the Rotten Kid Theorems

The Rotten Kid theorem gives conditions under which one player's altruistic interpersonal preferences leads to an efficient equilibrium outcome. There are many parallels between the Rotten Kid setup and the Rotten Firm setup. In both cases, the first-mover can take an action that helps the second-mover at a cost to himself. The second-mover can then transfer resources back to the first-mover. The first-mover is purely selfish, while the second-mover has preferences that depend on both her own payoff and the first-mover's. However, there are also important differences between the Rotten Kid theorem and the Rotten Firm theorem. The Rotten Firm theorem requires sufficiently-kinked interpersonal preferences, while the Rotten Kid theorem relies on material payoff functions that are quasi-linear in effort:

Theorem 4 *Suppose U is jointly-monotonic and weakly quasi-concave, and suppose π^F enters U as a normal good. Let $\left(w^{eff}, e^{eff}\right)$ be a materially-efficient transaction with associated material payoffs $\left(\pi^{W*}, \pi^{F*}\right)$, and let $(\widehat{w}, \widehat{e})$ be the transaction such that $\pi^F(\widehat{w}, \widehat{e}) = \pi^{F*}$ and $U\left(\pi^W(\widehat{w}, \widehat{e}), \pi^F(\widehat{w}, \widehat{e})\right) = \overline{U}$.*

1. *(Rotten Kid Theorem) If the material payoff functions are quasi-linear in effort (i.e., $c(e) = e$, contrary to the maintained assumption that $c(\cdot)$ is convex) then the equilibrium transaction is Pareto efficient.*

2. *(Rotten Firm Theorem) If the worker's interpersonal preferences are sufficiently-kinked in the sense that*

$$\frac{\lim_{\pi^W \downarrow \pi^{W*}} \partial U\left(\pi^W, \pi^{F*}\right)/\partial \pi^W}{\lim_{\pi^F \uparrow \pi^{F*}} \partial U\left(\pi^{W*}, \pi^F\right)/\partial \pi^F} \leq \frac{1}{c'\left(\widehat{e}\right)} < \frac{1}{c'\left(e^{\mathit{eff}}\right)} \leq \frac{\lim_{\pi^W \uparrow \pi^{W*}} \partial U\left(\pi^W, \pi^{F*}\right)/\partial \pi^W}{\lim_{\pi^F \downarrow \pi^{F*}} \partial U\left(\pi^{W*}, \pi^F\right)/\partial \pi^F},$$

then $\left(w^{\mathit{eff}}, e^{\mathit{eff}}\right)$ is the equilibrium transaction, and it is Pareto efficient.

Proof. Part 2 is Theorem 2, so we prove Part 1. Since $c(e) = e$, the set of materially-efficient transactions is given by all $\left(w^{\text{eff}}, e^{\text{eff}}\right)$ satisfying $v'\left(w^{\text{eff}}\right) = 1$. Given wage offer w, the worker's optimal effort choice maximizes $U\left(\pi^W, \pi^F\right)$ subject to $\pi^W + \pi^F = v(w) - w$. Because π^F enters U as a normal good, we know that the value $\widetilde{\pi}^F$ that solves $\left(\widetilde{\pi}^W, \widetilde{\pi}^F\right) = \arg\max_{\{(\pi^W, \pi^F): \pi^W + \pi^F \leq v(w) - w\}} U\left(\pi^W, \pi^F\right)$ is strictly increasing in $v(w) - w$. Hence the firm's optimal wage offer is w^{eff}, and the resulting equilibrium transaction is the worker's favorite materially-efficient transaction (hence Pareto efficient). ■

To understand the Rotten Kid theorem, notice that if the material payoff functions are quasi-linear in effort, then the efficiency of a transaction is entirely determined by the level of the wage (the level of effort merely redistributes material payoff). Because π^F enters U as a normal good, the worker will choose effort such that both π^F and π^W are increasing in the total surplus to be divided. That is, just as in the Rotten Firm theorem, the worker chooses effort according to a fairness rule. As a result, it is in the firm's best interest to offer a wage that leads to a materially-efficient outcome. Figure 6 illustrates the Rotten Kid theorem.

Until now, the Rotten Kid theorem has been ruled out because of the assumption that $c(\cdot)$ is convex. If $c(\cdot)$ is convex and U is smooth, the analysis in Section 6 showed that the equilibrium is bounded away from efficiency.

Although both the Rotten Firm theorem and the Rotten Kid theorem provide possible reasons for gift-exchange to be efficient, the Rotten Kid theorem is less applicable in situations of exchange. The assumption that the material payoff functions are quasi-linear in *effort* is necessary for the Rotten Kid theorem but unrealistic in many cases. That assumption makes material payoffs "transferable," which Bergstrom (1989) has shown is a key assumption underlying the Rotten Kid theorem and its extensions. In the classic Rotten Kid setup, there is a single good (money), ensuring that material payoffs are transferable. As in Bergstrom's (1989) counterexamples, the existence

of a second commodity (in this case, effort) generally introduces a failure of the transferability assumption, unless the material payoff functions are quasi-linear in that commodity. Since trade almost always involves at least two commodities, the Rotten Kid theorem will not typically apply in settings of exchange. By contrast, the Rotten Firm theorem does not depend on quasi-linearity in either wage or effort.

More generally, the Rotten Firm theorem and the Rotten Kid theorem represent the *only* two scenarios where the equilibrium of the gift-exchange game could potentially be Pareto efficient. To see this, recall from the proof of Theorem 1 in Section 6 that in order for $\frac{d\pi^F(w,e(w))}{dw} = 0$ at a Pareto efficient transaction, it is necessary that $\frac{d\pi^W}{dw} = \frac{v'}{1-\frac{\widetilde{H}}{c''}} = 0$ at that transaction. There are only two ways for this to be true. Either $c'' = 0$, which corresponds to the Rotten Kid theorem, or $\widetilde{H} \to -\infty$, which corresponds to the Rotten Firm theorem.

9 Conclusions

This paper has analyzed the extent to which interpersonal preferences can generate efficient exchange in the simplest possible gift-exchange setting. If the worker chooses effort in accordance with a "fairness rule," then it is in the firm's interest to offer a wage that induces efficient effort. The results extend straightforwardly to a firm that employs multiple workers, as long as each worker is concerned about the fairness of his bilateral transaction with the firm. The fact that a preference for fair transactions promotes efficient exchange might explain why it evolved biologically or culturally.

The extent to which individuals in real-world markets behave according to a fairness rule may depend on whether there is a salient terms of trade that is considered fair. Kahneman, Knetsch, & Thaler (1986) have argued that transactions, precedents, norms, and contracts (perhaps informal and even legally unenforceable) likely play a key role in making clear what is fair (Kahneman, Knetsch, & Thaler 1986). The farmers who leave fresh produce on a table by the road post a price they expect to receive (Dawes & Thaler 1988), as does the "Bagel Guy" (Dubner & Levitt 2004). Tipping norms determine the appropriate tip for typical service quality (Conlin, Lynn, & O'Donoghue 2003). Hart & Moore (2006) argue that the most important role of a contract may be in setting a reference point for what is fair. Being explicit about these benchmarks for what is fair requires introducing reference points into the model of fairness preferences.

Incorporating reference points is also important for extending the model to richer and more realistic environments. First, when there are multiple workers employed by a firm, field evidence

suggests that, in judging the fairness of their transactions with the firm, workers compare their own pay and required effort with the pay and required effort of other workers (e.g., Bewley 1999). Second, I have analyzed a firm and worker in isolation, but what happens in labor market equilibrium? The market wage often serves as a point of reference for a worker when he judges whether his own wage is fair (Kahneman, Knetsch, & Thaler 1986). Third, I have assumed that only employees have a preference for fair transactions, but what about firms? The overall pull toward fair outcomes would be even stronger if employers also had a preference for fairness, as long as the worker and firm share the same benchmark for what is "fair." On the other hand, if the players have self-serving judgments about which reference point is relevant, then this disagreement can cause a breakdown of negotiations (Babcock & Loewenstein 1997).

This paper has shown that a sufficiently strong concern for fairness of the appropriate kind enables the full realization of gains from trade. Of course, the firm would prefer an enforceable contract because relying on fairness requires sharing the gains from trade. However, it appears that under some circumstances fairness preferences can sustain efficient exchange even in settings where contracts would be costly to write or to enforce.

References

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

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

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

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

[5] Robert J. Barro. Are government bonds net worth? *Journal of Political Economy*, 82(6):1095–1117, 1974.

[6] Pierpaolo Battigalli and Martin Dufwenberg. Dynamic psychological games. September 21 2005. University of Arizona Working Paper.

[7] 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.

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

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

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

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

[12] Clive Bull. The existence of self-enforcing implicit contracts. *Quarterly Journal of Economics*, 102(1):147–160, February 1987.

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

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

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

[16] Michael Conlin, Michael Lynn, and Ted O'Donoghue. The norm of restaurant tipping. *Journal of Economic Behavior and Organization*, 52:297–321, 2003.

[17] James C. Cox, Daniel Friedman, and Vjollca Sadiraj. Revealed altruism. *University of Arizona Working Paper*, November 2005.

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

[19] Stephen J. Dubner and Steven D. Levitt. What the bagel man saw. *New York Times Magazine*, June 6 2004.

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

[21] Ernst Fehr, Alexander Klein, and Klaus M. Schmidt. Fairness and contract design. *Econometrica*, Forthcoming, 2006.

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

[23] Ray Fisman, Shachar Kariv, and Daniel Markovits. Pareto-damaging behaviors. *UC Berkeley mimeo*, 2005.

[24] Robert Fortsythe, Joel L. Horowitz, and N.E. Savin. Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6:347–69, 1994.

[25] Uri Gneezy and John List. Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, forthcoming, 2006.

[26] 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.

[27] Oliver Hart and John Moore. Contracts as reference points. November 2006. Harvard University Working Paper.

[28] 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.

[29] Ehud Kalai. Proportional solutions to bargaining situations: Interpersonal utility comparisons. *Econometrica*, 45(7):1623–30, 1977.

[30] David K. Levine. Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, 1(3):593–731, 1998.

[31] B. MacLeod and J.M. Malcolmson. Implicit contracts, incentive compatibility, and involuntary unemployment. *Econometrica*, 57:312–22, 1989.

[32] Alexandre Mas. Pay, reference points, and police performance. March 2005. UC Berkeley mimeo.

[33] 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.

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

[35] John E. Roemer. *Theories of Distributive Justice*. Harvard University Press, Cambridge, MA, 1996.

[36] Uzi Segal and Avia Spivak. First order versus second order risk aversion. *Journal of Economic Theory*, 51:111–25, 1990.

[37] Avner Shaked. The rhetoric of inequity aversion. March 2005. Working paper.

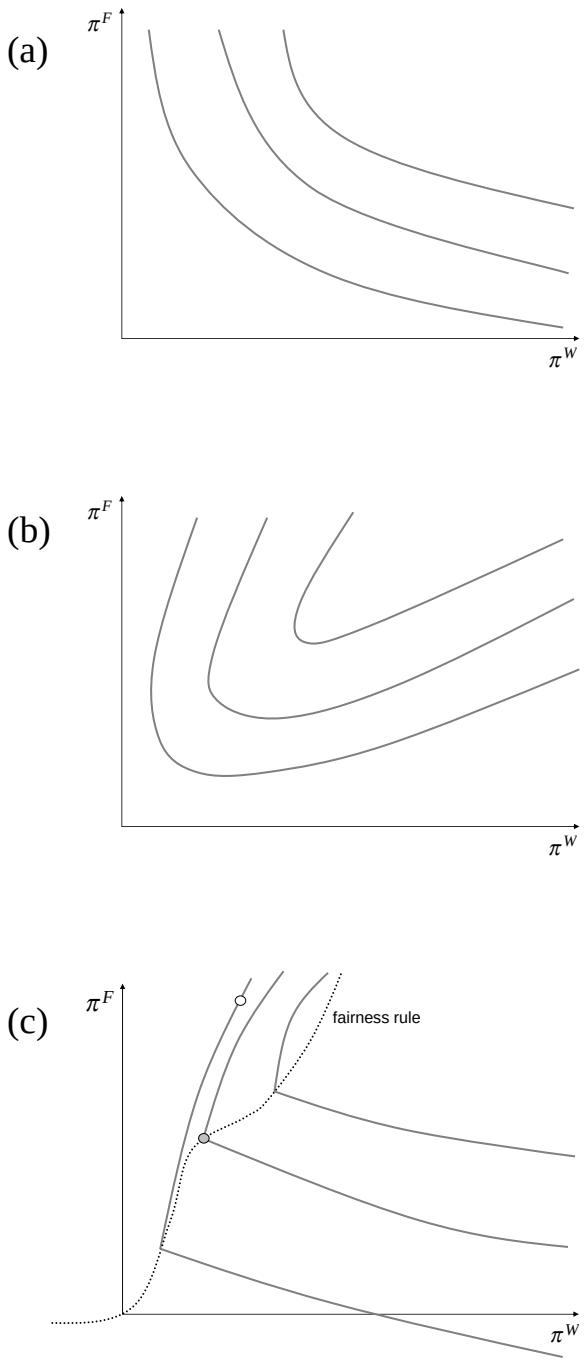

Figure 1. Indifference curves corresponding to interpersonal preferences that satisfy quasi-concavity and joint-monotonicity. Panel (a): Smooth preferences that are also monotonic. Panel (b): Smooth preferences that are not monotonic. Panel (c): Preferences that are first-order unfairness averse about a fairness rule. Due to the non-monotonicity, the darkly-shaded point is preferred to the white point that gives higher material payoffs to both players.

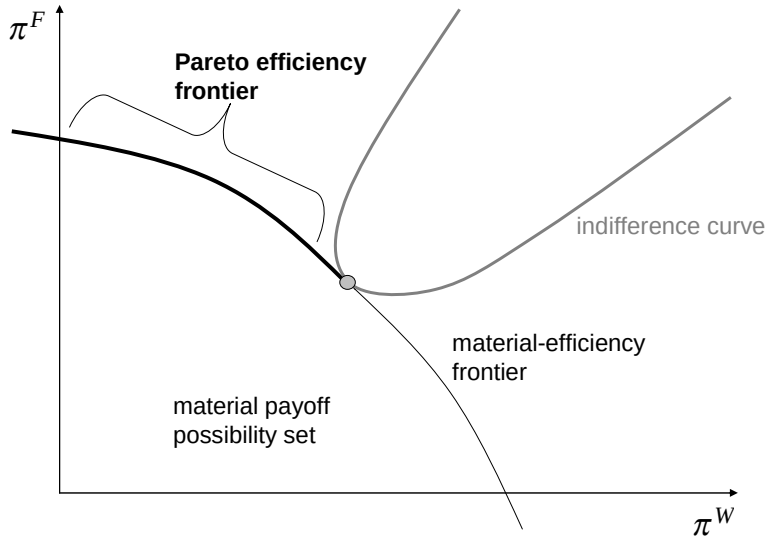

Figure 2. Relationship between Pareto efficiency and material-efficiency. The material payoff possibility set is convex, and the material-efficiency frontier is downward-sloping. The point corresponds to the worker's favorite materially-efficient transaction. The Pareto efficiency frontier is the subset of the material-efficiency frontier equal to material payoff pairs that give lower material payoff to the worker than the worker's favorite materially-efficient transaction does.

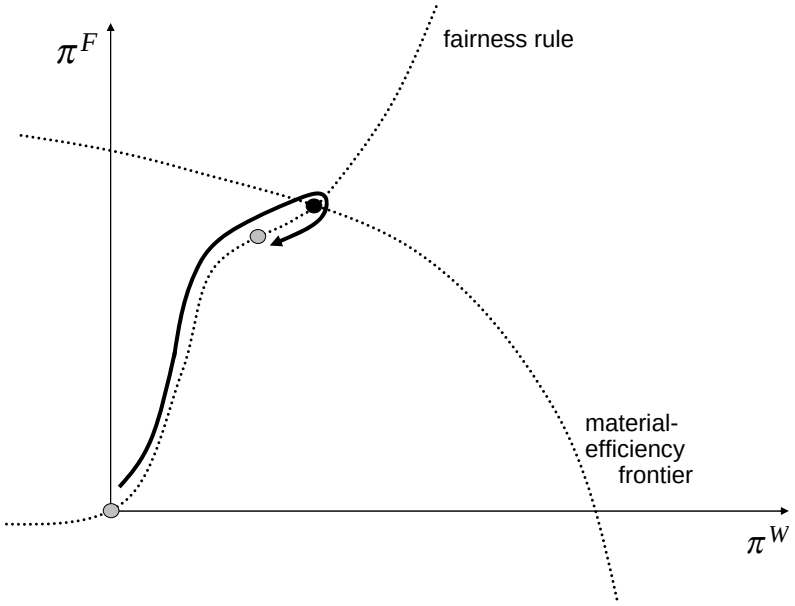

Figure 3. The logic of the Rotten Firm theorem. Because the worker obeys the fairness rule at all wage offers, the lightly-shaded points illustrate possible transactions that could occur, depending on the firm's wage offer. The arrow illustrates how material payoffs vary as the firm's wage offer increases. The black point is the equilibrium because it maximizes the firm's profit among the feasible transactions that satisfy the fairness rule.

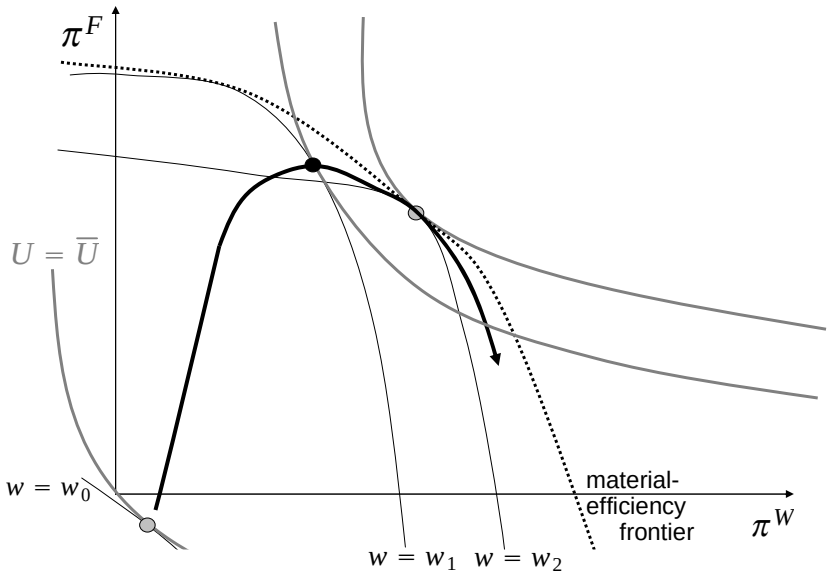

Figure 4. Equilibrium with smooth interpersonal preferences. The "wage curves" are, for given wage offers $w_0 < w_1 < w_2$, the possible material payoff pairs for varying effort levels. The lightly-shaded points show the effort the worker would choose at non-equilibrium wage offers. The arrow illustrates how material payoffs vary as the firm's wage offer increases. Note that the materially-efficient transaction along the path of the arrow is not an equilibrium because the firm can profitably deviate to a lower wage offer. The black point is the equilibrium.

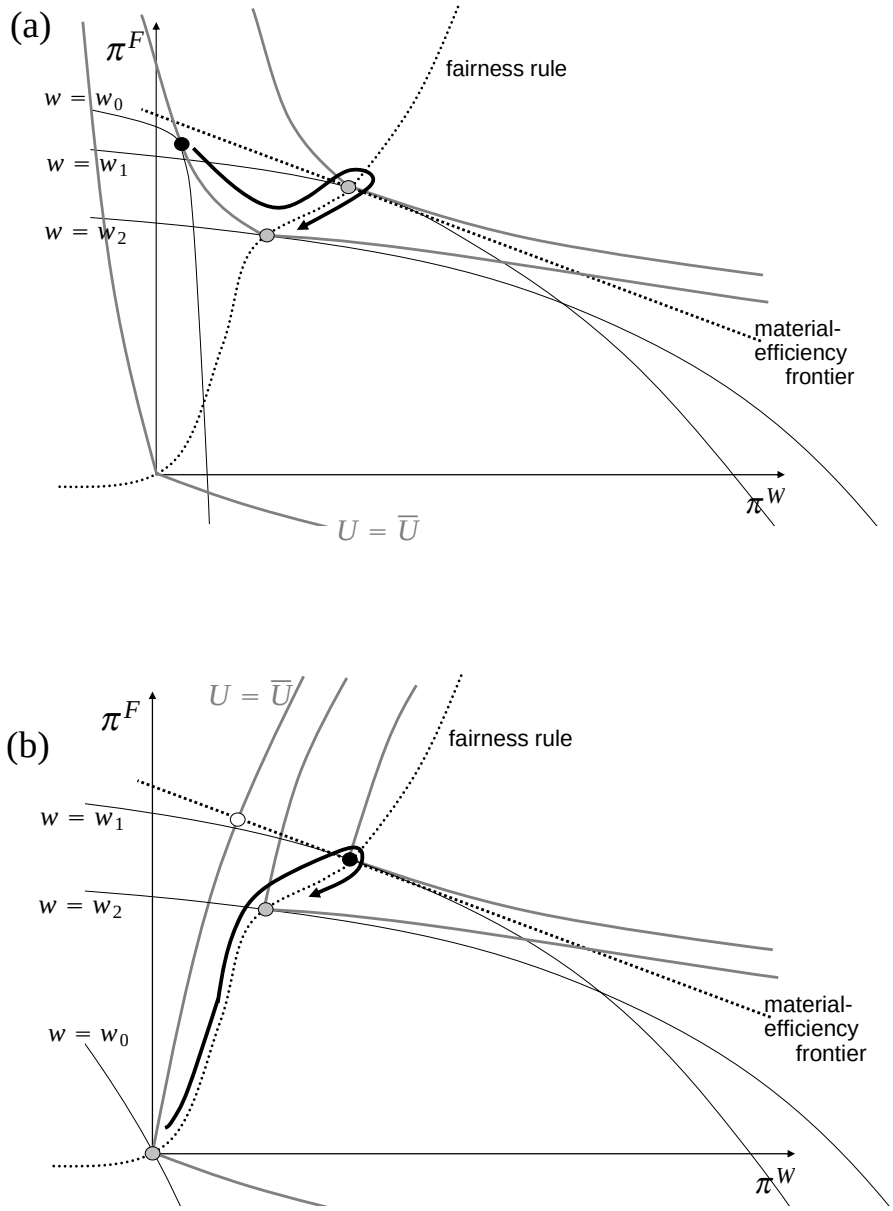

Figure 5. Equilibrium with interpersonal preferences that are first-order unfairness averse about a fairness rule. The "wage curves" are, for given wage offers $w_0 < w_1 < w_2$, the possible material payoff pairs for varying effort levels. The lightly-shaded points show the effort the worker would choose at non-equilibrium wage offers. The arrow illustrates how material payoffs vary as the firm's wage offer increases. The black point is the equilibrium. (a) "Too much" altruism: The firm maximizes profit by offering a low wage. (b) Rotten Firm theorem: Equilibrium with sufficiently-kinked interpersonal preferences is Pareto efficient. The white point is the outcome when the firm can offer the worker an enforceable contract.

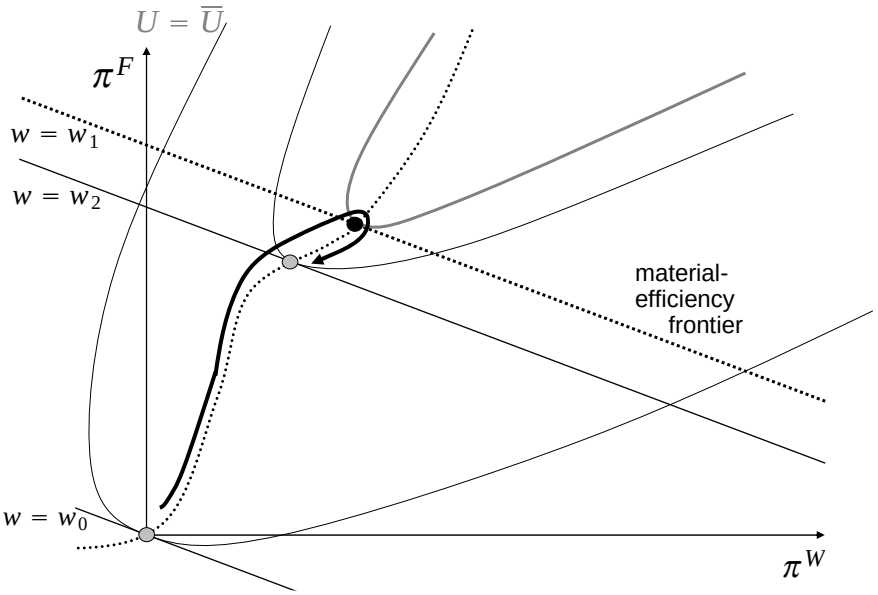

Figure 6. Rotten Kid theorem: If the firm's and worker's material payoff functions are quasi-linear in effort (making the wage curves parallel shifts of the material-efficiency frontier), and if the firm's material payoff enters the worker's utility as a normal good, then the equilibrium is efficient.
The "wage curves" are, for given wage offers $w_0 < w_1 < w_2$, the possible material payoff pairs for varying effort levels. The lightly-shaded points show the effort the worker would choose at non-equilibrium wage offers. The arrow illustrates how material payoffs vary as the firm's wage offer increases. The black point is the equilibrium.