Fairness and Efficiency: The Rotten Firm Theorem

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

July 20, 2006

Abstract

When contracting is not possible, a preference for *fair* exchange can generate *efficient* exchange, fully exhausting the potential gains from trade – or no exchange at all, leaving all gains from trade unexploited. A profit-maximizing firm offers a wage to a fair-minded worker, who then chooses how much effort to exert. The Rotten Firm theorem says: if the worker cares sufficiently about fairness and the firm employs the worker, the equilibrium transaction is Pareto efficient. However, even when gains from trade were possible, the firm may not to hire the worker because every "fair" transaction (acceptable to the worker) could be less profitable than the firm's outside option. The theory explains a puzzle: why firms offer profit-sharing plans to non-management employees.

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

Keywords: fairness, social preferences, gift-exchange, efficiency wages, profit-sharing

*A previous version of this paper circulated under the title, "A Theory of Fairness in Labor Markets." I am grateful to Philippe Aghion, Attila Ambrus, Antonia Atanassova, George Baker, Gary Becker, Lynn Benjamin, James Choi, Steve Coate, Noam Elkies, Florian Englmaier, Erik Eyster, Dan Friedman, John Friedman, Roland Fryer, Drew Fudenberg, Alexander Gelber, Jerry Green, Jonathan Hall, Oliver Hart, Daniel Hojman, Richard Holden, Caroline Hoxby, Erzo Luttmer, Lisa Kahn, Lauren Kaiser, Emir Kamenica, Lawrence Katz, Fuhito Kojima, Ilyana Kuziemko, Sendhil Mullainathan, Karthik Muralidharan, Emi Nakamura, Natalija Novta, Ted O'Donoghue, Emily Oster, Giacomo Ponzetto, Jesse Shapiro, Monica Singhal, Jón Steinsson, Jeremy Tobacman, Stephen Weinberg, Richard Zeckhauser, seminar participants at Cornell, Haas School of Business, Harvard University, LSE, MIT, University of Maryland at College Park, University of California at Santa Cruz, and especially Edward Glaeser, David Laibson, Matthew Rabin, and Andrei Shleifer for valuable comments and advice. 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; 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: benjamin@fas.harvard.edu.

1 Introduction

When are potential gains from trade realized? When both parties' actions can be bound by an enforceable contract, such a contract enables full realization of the potential gains from trade. The Coase theorem implies that the parties will agree to a Pareto efficient transaction (Coase 1960). In practice, though, it may be difficult to observe (or verify) whether the parties actually carry through on their promised actions. In general, this imperfect information causes the optimal contract to fall short of efficiency (Grossman & Hart 1983). If the exchange will be repeated, then reputational concerns can enable the parties to transact, even in the absence of contracting (e.g., Bull 1987, MacLeod & Malcomson 1989). However, there may be a multiplicity of equilibria (some inefficient), and efficient equilibria may require extensive information about past behavior.

In some contexts, social preferences such as altruism may promote efficiency. The Rotten Kid theorem says that if the head of a household is altruistic, then even purely self-regarding members of the household will act so as to maximize family income (Becker 1974; Bergstrom 1989). The role of altruism outside the family may be more limited. By contrast, a preference for fair transactions has been argued to influence behavior in a wide range of market settings (Akerlof 1982; Kahneman, Knetsch, & Thaler 1986; Akerlof & Yellen 1990; Mas 2005; but see Gneezy & List 2006 and List 2006).

In simple social allocation problems, a preference for fairness – a desire for equal payoffs across individuals – has no direct connection with social efficiency and in fact often conflicts with it. It is well-known that a preference for fairness often causes individuals to choose social allocations that are *inefficient*, but more fair (e.g., Bazerman, Loewenstein, & White 1992; Charness & Rabin 2002; Fisman, Kariv, & Markovits 2005b). Yet a preference for fair transactions appears to explain why exchange occurs at all in laboratory markets, where anonymous, random matching rules out reputational mechanisms and where contracting is impossible (e.g., Fehr, Kirchsteiger, & Riedl 1993; Fehr & Falk 1999). In these markets, where purely self-regarding individuals would leave all potential gains from trade unexploited, a preference for fairness promotes efficiency. However, it remains an open question to what extent a preference for fair transactions can "substitute for" the availability of binding contracts.

In this paper, I explore when a preference for *fair* exchange can generate *efficient* exchange. For concreteness, I study the interaction between a profit-maximizing firm, who pays a wage, and a fair-minded worker, who provides effort. The main result is the Rotten Firm theorem: if the worker cares sufficiently about fairness, then even though the firm is purely self-regarding, the equilibrium

transaction is Pareto efficient. Like when contracts are available, potential gains from trade will be *fully* exhausted when the worker has a strong enough preference for fairness – but only *if* trade occurs at all. The efficiency-promoting properties of a preference for fairness are the result of the strategic interaction between the parties. However, unlike when contracts are available, it may be that the firm prefers not to employ the worker, even though gains from trade were possible. Therefore, a preference for fairness substitutes perfectly for the availability of binding contracts under some circumstances, but not at all under others.

The central intuitions can be illustrated with a simple example. A firm offers a wage w to a worker. Then the worker chooses how much effort to exert e. The firm's profit, or "material payoff," $\pi^F(w, e)$ is decreasing in the wage and increasing in effort. The worker's "material payoff" $\pi^W(w, e)$ – the self-regarding payoff that would describe the worker's preferences if the worker were purely selfish – is increasing in the wage and decreasing in effort.

Later I will discuss the worker's preference for fair transactions in detail and how it influences his behavior. For now, take as given that the most fair choice of effort $e^{\text{fair}}(w)$ satisfies the following "fairness rule":

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

$\widehat{\pi}^W$ and $\widehat{\pi}^F$ are exogenous "reference payoffs." These describe the payoff levels for each party that would transpire from the "reference transaction," a relevant precedent likely influenced by current and past market rates, the worker's recent personal labor market experience, and the terms enjoyed by other workers (Kahneman, Knetsch, & Thaler 1986). According to (fair), if the worker's actual material payoff from the exchange exactly equals his reference payoff, then it is most fair for the firm's material payoff to equal its reference payoff. If the firm gets a greater payoff than its reference payoff, then the most fair thing is for the worker also to get a greater payoff than his reference payoff. If the worker cares sufficiently about fairness, then it will turn out that the worker will choose his effort level in accordance with (fair) in equilibrium. This fair level of effort is increasing in the wage: $\frac{de^{\text{fair}}(w)}{dw} = -\frac{\partial \pi^W\left(w, e^{\text{fair}}\right)/\partial w \; - \; \partial \pi^F\left(w, e^{\text{fair}}\right)/\partial w}{\partial \pi^W(w, e^{\text{fair}})/\partial e \; - \; \partial \pi^F(w, e^{\text{fair}})/\partial e} > 0$.

Taking into account how the worker's effort choice responds to the wage, the employer's profit-maximizing wage offer satisfies the first-order condition $\frac{\partial \pi^F\left(w, e^{\text{fair}}\right)}{\partial w} + \frac{\partial \pi^F\left(w, e^{\text{fair}}\right)}{\partial e}\frac{de^{\text{fair}}(w)}{dw} = 0$. Substituting and rearranging, it follows that at the equilibrium wage and effort,

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

That is, the firm's marginal rate of substitution between wage and effort equals the worker's marginal rate of substitution, the usual condition for Pareto efficiency.

Why is the equilibrium efficient? Since the worker chooses effort according to (fair), the firm and worker both receive a positive share of the marginal gains from trade that accrue from an incremental change in wage; their payoffs increase or decrease in tandem. The firm maximizes its own profit by realizing the maximum possible gains from trade. Consequently, if the firm employs the worker, the transaction will be Pareto efficient.

Why might the firm prefer not to employ the worker, even though gains from trade were possible? Whenever the worker's concern for fairness allows exchange to occur, the equilibrium requires the firm to share the rents from employment with the worker. Because there are potential gains from trade, the worker would like to commit to an effort level that makes both the firm and worker better off than their outside options. But because no contract is enforceable, when the time comes to choose effort, the worker will choose according to (fair). It may be that every "fair" transaction is less profitable than the firm's outside option.

In the paper, I also discuss what happens when the worker cares less about fairness. In that case, the equilibrium transaction is not Pareto efficient. However, the stronger the worker's concern for fairness, the more efficient the exchange will be if it occurs *and* the more likely it is that it will occur.

The rest of the paper develops these ideas more carefully. A description of the worker's fair-minded preferences is central to a complete analysis. Section 2 introduces a model of a preference for fair transactions that may be relevant in a wide range of market settings. The reference transaction sets a benchmark for what is fair. Deviations from these payoffs are judged most unfair when one party gets more than her reference payoff, and the other party gets less than his. The preferences represent an adaptation of Fehr & Schmidt's (1999) model of "inequity-aversion," as well as an elaboration on and formalization of Kahneman, Knetsch, & Thaler's (1986) "dual entitlement principle." By making explicit the reference transaction, the theory helps clarify how to apply Fehr & Schmidt's (1999) model in market settings.

In the example sketched above, I asserted that (eff) implies that the equilibrium is Pareto efficient. This is true if the worker's preferences are represented by $\pi^W(w, e)$, but what about when the worker's preferences include a concern for fair exchange? Section 3 characterizes the set of Pareto efficient transactions when the worker's preferences are those described in Section 2. Even though a fair-minded worker's preferences are more complicated than simply $\pi^W(w, e)$, I show that (eff) is a necessary condition for Pareto efficiency. Combined with an additional condition that is satisfied in equilibrium, (eff) is sufficient for Pareto efficiency. Section 3 also describes what it means for there to be potential gains from trade when the worker is fair-minded.

To set a benchmark for the subsequent analysis, Section 4 briefly discusses what happens when the firm can offer the fair-minded worker an enforceable contract. In that case, exchange occurs if and only if there are potential gains from trade. The firm gets the entire value of these rents, and the worker earns only his outside option level of utility.

For various degrees of fair-mindedness, Section 5 characterizes the employment equilibrium of the model when contracts are not available. Section 5 also states the Rotten Firm theorem. Appendix A calibrates the model using existing laboratory evidence on social preferences.

Section 6 shows that the firm might not employ the worker despite potential gains from trade, but trade occurs more often the stronger is the worker's concern for fairness. In the potentially interesting special case where the market terms of exchange set the reference transaction in addition to the firm's and worker's outside options, and if the worker's concern for fairness is sufficiently strong, then exchange occurs whenever there are potential gains from trade. To the extent that market rates serve as the exclusive benchmark for judging fairness, sufficient fair-mindedness is a "perfect substitute" for enforceable contracts (in terms of efficiency).

While the main part of the analysis focuses on efficiency, Section 7 addresses the testable implications of the model. Workers' preference for fair transactions explains the puzzle of why so many firms offer profit-sharing plans to non-management employees. Widespread profit-sharing is a puzzle for standard incentive theory because most employees have only a negligible impact on profit. However, profit-sharing makes employees feel fairly paid, regardless of how profitable the firm ends up being. The model predicts that, even though the optimal wage is increasing in profit, equilibrium effort may be *decreasing* in profit. The theory also explains several other labor market regularities: rent-sharing, the relative insensitivity of wages to external market conditions, and the fact that reference transactions (relevant precedents) matter in wage negotiations. The theory predicts that these regularities should cluster in occupations where important aspects of output are non-contractible.

Section 8 explicitly contrasts the implications of a preference for fairness with those of altruism. Rotten Firm and Rotten Kid theorems are analogous in many ways. However, classic Rotten Kid theorems assume transferable utility (Bergstrom 1989), which is not generally satisfied in situations of exchange. Moreover, for labor market transactions, it seems likely that fairness is a more salient motivation for effort choice than altruism toward the firm. If a worker were altruistic, then a profit-maximizing firm would (counterfactually) make the wage *decreasing* in profit. A utilitarian preference for efficiency (e.g., Charness & Rabin 2002) is formally equivalent to altruism, so the same results apply. Section 9 concludes. Appendix B discusses robustness to alternative ways of

modeling a preference for fair transactions and presents an analog to the Rotten Firm theorem. Formal proofs are in Appendix C.

This paper relates to a growing literature that incorporates inequity-averse preferences into contract theory (see Englmaier 2004 for a review). For example, Fehr, Klein, & Schmidt (2001) theoretically and experimentally study incomplete contracting and adverse selection with fair-minded and selfish firms and workers. Fehr & Schmidt (2004) study multi-tasking, again with heterogeneous firms and workers. By contrast, the present paper focuses on understanding a very simple model of exchange in the absence of contracting when a transactor cares about fairness. In addition, the model of fairness preferences presented here provides a microfoundation for the "fair wage-effort" theory of efficiency wages. The original presentation of this theory posited a positive relationship between effort and wage without a formal specification of the worker's preferences (Akerlof & Yellen 1990). The present paper can address the efficiency of the employment transaction because it builds the theory from an underlying preference for fairness.

2 A Preference For Fair Transactions

Two players, a firm and a worker, play a sequential trading game. The firm offers a wage $w \in \mathbb{R}$ to the worker. Then the worker chooses effort $e \in \mathbb{R}$.[1] For simplicity, the firm's profit, or **material payoff function**, is additively-separable in wage and effort:

$$\pi^F(w, e) = y(e) - w,$$

where $y(\cdot)$ is a concave production function ($y' > 0$, $y'' < 0$). The firm maximizes its material payoff. The worker's material payoff function,

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

reflects the concave personal benefits of higher compensation ($v' > 0$, $v'' < 0$) and convex costs of exerting greater effort ($c' > 0$, $c'' > 0$). For technical convenience, I assume $\lim_{e \to -\infty} y'(e) = \infty$, $\lim_{e \to \infty} y'(e) = 0$, $\lim_{w \to -\infty} v'(w) = \infty$, and $\lim_{w \to \infty} v'(w) = 0$. The worker does not necessarily maximize only his material payoff. The worker maximizes utility U (described below), which may include social considerations beyond his own material payoff. Rather than offer a wage to the worker, the firm could receive an outside option level of profit $\overline{\pi^F}$. Rather than accept a wage offer,

[1] Wage and effort are taken to be unbounded for technical convenience. The results go through if they are restricted to bounded intervals, as long as the equilibrium is interior to the intervals.

the worker could get outside option utility $\overline{U}$.[2] If trade occurs, call the wage-effort pair (w, e) a **transaction**. The solution concept is subgame-perfect Nash equilibrium.

In a typical principal-agent problem, output is partly random, and the firm can make the wage a function of output. By contrast, here I assume that output is a deterministic function of effort, and the firm cannot make the wage depend on output (or effort). I make these assumptions for two reasons. First, the setup provides a reasonable approximation to real-world settings where there are components of output that are costly or impossible to contract on. For example, it can be difficult to write an enforceable contract that requires an employee to suggest creative, new production technologies or be friendly to customers. When hiring a doctor or lawyer, it can be hard even to determine the quality of output.

Second, if the worker were purely selfish, with utility function $U = \pi^W(w, e)$, there would be no exchange. The firm would prefer its outside option because it could not guarantee that the worker would exert any effort. Because selfishness leads to no-trade under these assumptions, they serve to make as clear as possible the implications of the worker's preference for fair transactions. The equilibrium terms of the transaction – and the fact that exchange occurs at all – are entirely driven by the worker's concern for fairness. In the remainder of this section, I describe how the worker judges the fairness of a transaction and how the worker's preferences combine self-interest with a concern for fairness.

What does it mean for a transaction to be more fair or less fair than another transaction? Two concepts are central in describing how people judge the fairness of an exchange. First, transactions are judged in comparison to some "relevant precedent" that for sets a benchmark for fair exchange (Kahneman, Knetsch, & Thaler 1986). Call this benchmark the **reference transaction**, denoted $(\widehat{w}, \widehat{e})$. Exactly what the reference transaction is does not matter for whether exchange is efficient (when trade occurs), so I take it as exogenous. In practice, current and past market rates, the worker's recent labor market experience, and the terms enjoyed by other workers in the same firm probably all influence the reference transaction. Notice that, even though there is only one worker in the model, there is implicitly a role for other workers. In judging the fairness of his own terms of trade with the firm, the worker may compare his terms with other workers'.

The worker's **reference payoff** $\widehat{\pi}^W \equiv \pi^W(\widehat{w}, \widehat{e})$ is the material payoff the worker would receive if the reference transaction occurred. Analogously, $\widehat{\pi}^F \equiv \pi^F(\widehat{w}, \widehat{e})$ is the firm's reference payoff. The worker judges an alternative transaction (w, e) in terms of how well the parties do from the

[2] In principle, $\overline{U}$ incorporates not only the worker's material payoff from his next best alternative employment, but also how the worker feels about the fairness of that alternative. Taking $\overline{U}$ as constant is a simplifying assumption. It means that the worker's and firm's choices do not affect how the worker evaluates his alternative.

alternative transaction, relative to how well they would have done from the reference transaction. To be precise, call the difference between the worker's material payoff from the alternative transaction and from the reference transaction, $\widetilde{\pi}^W \equiv \pi^W(w,e) - \widehat{\pi}^W$, the worker's **surplus payoff**. Analogously, let $\widetilde{\pi}^F \equiv \pi^F(w,e) - \widehat{\pi}^F$ denote the firm's surplus payoff. Note that the surplus payoffs are both equal to zero when the reference transaction actually occurs.

The second key idea is that a transaction is maximally fair only if the worker's surplus payoff equals the firm's surplus payoff. Specifically, suppose a transaction gives rise to surplus payoffs $\widetilde{\pi}^W$ and $\widetilde{\pi}^F$. The following **fairness function** describes the worker's assessment of how fair the transaction is:

$$f\left(\widetilde{\pi}^W, \widetilde{\pi}^F\right) = -\gamma \max\left\{\widetilde{\pi}^W - \widetilde{\pi}^F, 0\right\} - (1-\gamma)\max\left\{\widetilde{\pi}^F - \widetilde{\pi}^W, 0\right\}, \tag{1}$$

where $0 < \gamma < 1$. This function takes a maximum value of zero when $\widetilde{\pi}^W = \widetilde{\pi}^F$. A transaction that gives equal surplus payoffs, such as the reference transaction, is a **fair transaction**. The second term of (1) captures **disadvantageous unfairness** for the worker: it is unfair when the firm's surplus exceeds the worker's. It is also unfair when the worker's surplus exceeds the firm's. The first term captures this **advantageous unfairness** for the worker. The extent to which the worker may perceive disadvantageous unfairness as worse than advantageous unfairness is parameterized by γ. Evidence suggests that individuals generally perceive disadvantageous unfairness as worse (e.g., Fehr & Schmidt 1999). For the proceeding analysis, I allow any $0 < \gamma < 1$. Figure 1 displays some of the key concepts in the space of material payoffs. The "material payoff possibility set," the set of material payoff pairs $\left(\pi^W, \pi^F\right)$ that are attainable by some transaction, is convex due to the assumptions on the material payoff functions. The "equal-surplus line" traces out the material payoff pairs that correspond to fair transactions.

There is much evidence that these principles are at work in individuals' fairness judgments. For example, Kahneman, Knetsch, & Thaler (1986) presented 195 respondents with two scenarios. In the first:

> A small company employs several workers and has been paying them average wages. There is severe unemployment in the area and the company could easily replace its current employees with good workers at a lower wage. The company has been making money. The owners reduce the current workers' wages by 5 percent.

In this case, 77% judged the wage cut as unfair. The second scenario was the same, except:

...The company has been losing money. The owners reduce the current workers' wages by 5 percent.

In the second case, only 32% regarded the wage reduction as unfair. In both cases, the reference transaction is the worker's current wage (and presumably current effort level). In the first scenario, cutting the wage (presumably holding effort constant) is considered unfair because the firm gains while the worker loses. The firm's surplus payoff exceeds the worker's, generating disadvantageous unfairness for the worker. In the second scenario, by contrast, the firm is described as having a lower payoff than its reference payoff ("losing money"). Keeping the surplus payoffs equal actually requires cutting wages.

The same principles apply in non-labor-market contexts. In product markets, the reference transaction is a price paid to the firm and a quality delivered to the consumer. When the cost of apples to the supermarket increases, consumers consider it fair for the firm to raise the price of apples, ensuring that the seller and buyer share in the reduction of gains from trade. By contrast, when costs remain constant, raising the price is particularly unfair because the seller gains at the buyer's expense. Fairness requires a diner to pay a higher (or lower) than customary tip to a waiter who provides better (or worse, respectively) service than typical.

I assume that the worker cares about the fairness of his transaction with the firm, in addition to his own material payoff. For convenience, the worker's utility is additively-separable in these two components:

$$U = \pi^W + \phi f\left(\pi^W - \widehat{\pi}^W, \pi^F - \widehat{\pi}^F\right). \tag{2}$$

The weight the worker puts on fairness relative to his purely self-regarding payoff is parameterized by $\phi \geq 0$. The special case $\phi = 0$ corresponds to the common assumption that the worker is entirely selfish. The utility function represents an adaptation of Fehr & Schmidt's (1999) model of "inequity-aversion" to non-laboratory contexts (see also Loewenstein, Thompson, & Bazerman 1989). The model here makes the reference transaction explicit and allows for more than one commodity (both wage and effort). It also can be understood as an elaboration on and formalization of Kahneman, Knetsch, & Thaler's (1986) "dual entitlement theory."

This specification of preferences omits other notions of fairness, such as procedural fairness (e.g., Frey, Benz, & Stutzer 2004; Cohen-Charash & Spector 2001) and reciprocating kind intentions (Rabin 1993; Levine 1994). However, the above utility function tractably captures crucial features of a concern for fair transactions. Moreover, in many cases of interest, the qualitative results

from this model are likely to be similar to those from more complex theories.[3] The specification also omits a utilitarian motivation that affects the way many experimental participants allocate resources across individuals (e.g., Charness & Rabin 2002; Engelmann & Strobel 2004; Fisman, Kariv, & Markovits 2005a). I contrast a utilitarian motivation with fairness in Section 8, and Appendix B explores a case where the worker has both motivations.[4]

An important feature of the utility function (2) for the analysis that follows is that it is kinked at every transaction that equates the parties' surplus payoffs. This kink will imply that, under some conditions, the worker's optimal level of effort will follow a "fairness rule" of equating the parties' surplus payoffs. The kink could be viewed as an approximation to a highly-curved but smooth function. In Appendix B, I discuss how the kink is crucial for the equilibrium to be fully efficient, and I show how a smooth fairness function that approximates a kink generates near-efficiency. However, the kink accurately captures the behavior of many participants in laboratory experiments. A disproportionate number of participants choose to split monetary rewards exactly equally between themselves and others, even though they presumably care about their own material payoff in addition to caring about fairness (e.g., Andreoni & Miller 2002; Fisman, Kariv, & Markovits 2005a). Similarly, in real-world settings, people often adhere to rules of fairness (such as 50-50 splits) as though not trading off fairness with other considerations.

The analysis that follows will show that a preference for fair transactions leads to efficient exchange when the weight on fairness ϕ is sufficiently large. However, it is important to recognize at the outset that there is no direct connection between fairness and efficiency. In fact, in non-strategic settings, a preference for fairness typically causes individuals to choose *inefficient* (but more equal) allocations across individuals. 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 "Pareto-damaging" choices when real money is at stake, though less commonly (e.g., Fisman, Kariv, & Markovits 2005b;

[3] For example, in much of what follows, the key implication of the above utility function is that the worker reciprocates a higher wage with greater effort. The same implication would generally follow from a theory of reciprocal kindness.

[4] The experimental evidence typically implicates both a preference for fairness and a utilitarian motivation (e.g., Charness & Rabin 2002; Engelmann & Strobel 2004). As long as the worker cares enough about fairness, the results of this paper go through. For example, the results are essentially unchanged if $f\left(\widetilde{\pi}^W, \widetilde{\pi}^F\right) = \min\left\{\widetilde{\pi}^W, \widetilde{\pi}^F\right\}$ (see Appendix B). Since the kink in the fairness function drives the results in that case (see below), it makes the exposition clearer to put aside the utilitarian motivation for purposes of analysis.

Charness & Rabin 2002).[5] Yet it is precisely the individuals with large ϕ who are *most* likely to choose inefficient allocations.

3 Pareto Efficiency and Potential Gains from Trade

The analysis that follows addresses questions about when potential gains from trade are exploited and to what extent. However, whether there are potential gains from trade and whether exchange is efficient depends on the parties' preferences. This section clarifies what these concepts mean when the worker has a preference for fair transactions.

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

Definition 1 *A transaction (w, e) is **Pareto efficient** if there is no other transaction (w', e') such that $\pi^F(w', e') \geq \pi^F(w, e)$ and $U(w', e') \geq U(w, e)$, at least one inequality strict.*

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 $\phi = 0$ and $U(w, e) \equiv \pi^W(w, e)$. A transaction that *would be* Pareto efficient if the worker were selfish rather than fair-minded is called materially-efficient.

Definition 2 *A transaction (w, e) is **materially-efficient** if there is no other transaction (w', e') such that $\pi^F(w', e') \geq \pi^F(w, e)$ and $\pi^W(w', e') \geq \pi^W(w, e)$, at least one inequality strict.*

A transaction is materially-efficient if and only if it equates the transactors' (selfish) marginal rates of substitution,

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

Figure 1 shows that the frontier of materially-efficient transactions is downward-sloping in wage-effort space.

What conditions characterize Pareto efficiency when a worker has a preference for fair transactions? In that case, the worker's utility function (2) is much more complicated than just $\pi^W(w, e)$. The worker's marginal rate of substitution is not in general equal to $\frac{\partial \pi^W(w,e)}{\partial w} / \frac{\partial \pi^W(w,e)}{\partial e}$. Consequently, (eff) cannot in general be expected to describe Pareto efficient transactions.

[5] To be precise, individuals' choices are "materially-inefficient" (as defined in Section 3). The choices are not actually Pareto inefficient because even though the outcome is worse for the other party, it is preferred by the individual making the choice.

Nonetheless, it turns out that (eff) remains a necessary condition for Pareto efficiency. To see why, recall that the worker's utility function (2) equals his material payoff $\pi^W(w,e)$, minus a correction for the unfairness of the transaction. If a transaction (w,e) does not satisfy (eff), then there are other transactions that improve both parties' material payoffs. It will always be possible to find some transaction (w',e') that increases both parties' material payoffs equally, so that (w',e') is just as fair as (w,e). But if (w',e') gives the worker a higher material payoff and is no less fair than (w,e), then the worker's utility must be higher under (w',e') than (w,e). This logic shows that a transaction that does not satisfy (eff) cannot be Pareto efficient.

Proposition 1 *If a transaction (w,e) is Pareto efficient, then it satisfies (eff). If a transaction (w,e) satisfies (eff) and $\pi^W(w,e) - \widehat{\pi}^W \leq \pi^F(w,e) - \widehat{\pi}^F$, then it is Pareto efficient.*

The second part of Proposition 1 gives a condition a transaction could satisfy that, when combined with material-efficiency, is sufficient for Pareto efficiency. The condition is that the transaction is disadvantageously unfair to the worker: the worker's surplus payoff is smaller than the firm's surplus payoff. Recall that along the frontier of materially-efficient transactions, any exchange that gives a higher material payoff to the firm gives a lower material payoff to the worker. If a materially-efficient transaction were *advantageously* unfair to the worker, then it might not be Pareto efficient. An alternative transaction on the material-efficiency frontier that gives slightly higher payoff to the firm and slightly lower payoff to the worker might make both parties better off. The firm earns greater profit, and the worker may prefer the more fair transaction, even though it gives lower material payoff. This is not the case for a materially-efficient transaction that is *disadvantageously* unfair to the worker. An alternative transaction on the material-efficiency frontier that gives slightly higher payoff to the firm and slightly lower payoff to the worker will make the firm better off and the worker worse off. The worker will be worse off under this alternative transaction because it is less fair, in addition to giving him a lower material payoff. Figure 1 illustrates that the set of transactions that is Pareto efficient is contained in the set of transactions that is materially-efficient.

I have discussed whether a particular transaction is Pareto efficient – whether it exhausts all of the gains that are possible from exchange. A distinct question is whether there are any gains from trade to be had at all, relative to not trading. In an encounter between a worker and a firm, there are potential gains from trade if there is some possible exchange that would make both parties better off than their outside options.

Definition 3 ***There are potential gains from trade*** *if there is some* (w, e) *such that* $\pi^F(w, e) > \overline{\pi^F}$ *and* $U(w, e) > \overline{U}$.

Figure 2 illustrates a case where there are potential gains from trade. The worker's indifference curves have a tilted-V shape, with a kink on the equal-surplus line. The figure shows one of these (drawn darkly), the worker's outside option indifference curve. The firm's outside option indifference curve is a horizontal line $\pi^F = \overline{\pi^F}$. The darkly-shaded region corresponds to the set of material payoff pairs (attainable by some transaction) that make both parties better off than their outside options.

Of course, whether there are potential gains from trade depends on the worker's preferences. Recall that the worker's utility function (2) has a parameter ϕ describing how much weight the worker puts on the fairness function (1). A transaction's utility is smaller than its material payoff to the extent that the transaction is unfair. Holding constant the material payoff functions, any given transaction is at least as attractive to a worker who puts less weight on fairness.

Proposition 2 *If there are potential gains from trade when the worker's weight on fairness is* ϕ*, then there are potential gains from trade when the worker's weight on fairness is any* $\phi' < \phi$.

Figure 2 shows that reducing the weight on fairness enlarges the set of transactions that are perceived as mutually beneficial to the firm and the worker. The lightly-drawn outside option indifference curve corresponds to a higher weight on fairness than the darkly-drawn one. The lightly-shaded area shows the additional material payoff pairs that now make both parties better off than their outside options.

On the other hand, the stronger the weight on fairness, the smaller the set of transactions that are acceptable to the worker because any unfairness carries greater weight. Because the primary question of interest is what happens when there are potential gains from trade, it will be convenient to assume in the analysis that follows that there are potential gains from trade *for all* ϕ. This assumption is largely technical. It means that there is some fair transaction (on the equal-surplus line) that both parties prefer to their outside options. It ensures that exchange is possible when ϕ is large but does not imply that it occurs.

4 Exchange With Enforceable Contracts

As a benchmark for the equilibrium without contracts, this section examines what happens when the firm can offer a take-it-or-leave-it contract to the worker. To be precise, instead of taking its

outside option level of profit $\overline{\pi^F}$, the firm could offer a wage-effort pair (w, e) to the worker. In that case, the worker can either accept this contract and be committed to exert effort e or decline and get outside option utility $\overline{U}$.

The most profitable contract for the firm must be Pareto efficient. This is because if the worker would accept a contract that were not Pareto efficient, then there is some alternative contract that makes both parties strictly better off (they could split the unexploited gains from trade). The worker would of course accept that contract, and the firm would prefer it to the initial contract.

Along the Pareto efficient frontier, any contract that gives higher material payoff to the firm gives lower utility to the worker. Therefore, the most profitable contract for the firm gives the worker the lowest level of utility, subject to the worker preferring the contract to his outside option. The black point in Figure 2 shows that the equilibrium payoffs occur at the intersection of the material-efficiency frontier and the worker's outside option indifference curve.

Proposition 3 *Suppose the firm can offer the worker an enforceable contract. If exchange occurs, it is Pareto efficient, and the worker gets exactly his outside option level of utility. Exchange occurs if and only if there are potential gains from trade.*

The last part of Proposition 3 states that gains from trade are always exploited. The only circumstance when exchange does not occur is when the most profitable contract is less profitable than the firm's outside option. But if that is true, then there cannot be any potential gains from trade. In Figure 2, this would mean that the $\pi^F = \overline{\pi^F}$ line lies above the equilibrium point, so the darkly-shaded region would vanish.

The equilibrium transaction when enforceable contracts are not available will differ in several ways. First, exchange will be Pareto efficient only if the worker's concern for fair transactions is strong enough. In that case, the worker will get strictly more than his outside option level of utility. Second, because of this rent-sharing, exchange may not occur even when there are potential gains from trade. The next two sections analyze these two differences in turn.

5 Exchange Without Enforceable Contracts

When enforceable contracts are not available, the firm must rely on the worker's sense of fairness to provide effort. This section solves for the equilibrium employment transaction in two parts. First, I characterize how the worker's choice of effort responds to the firm's wage offer. Second, I derive the firm's optimal wage offer.

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

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

where f is the fairness function (1). The most fair choice of effort $e^{\text{fair}}(w)$ satisfies the "fairness rule"

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

equating the worker's surplus payoff with the firm's surplus payoff. It follows that

$$\frac{de^{\text{fair}}(w)}{dw} = -\frac{\partial \pi^W\left(w, e^{\text{fair}}\right)/\partial w \; - \; \partial \pi^F\left(w, e^{\text{fair}}\right)/\partial w}{\partial \pi^W\left(w, e^{\text{fair}}\right)/\partial e \; - \; \partial \pi^F\left(w, e^{\text{fair}}\right)/\partial e} > 0.$$

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 surplus 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 both his concern for a fair transaction and his own material payoff:

$$e(w) \equiv \arg\max_{e} \pi^W(w,e) + \phi f\left(\pi^W(w,e) - \widehat{\pi}^W, \pi^F(w,e) - \widehat{\pi}^F\right).$$

The worker's most-preferred choice of effort turns out to be equal to the most fair level of effort, 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 $\overline{e}$.

Lemma 1 *Suppose the firm offers the worker a wage w, and the worker accepts employment. Then there is some reciprocity upper bound $\overline{e}$ such that*

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

where $\overline{w} \equiv \left(e^{fair}\right)^{-1}(\overline{e})$.

It might seem surprising that for a range of wage offers, the worker chooses exactly the most fair level of effort, even though the worker's preferences trade off fairness with material payoff. The reason is that the worker's utility function is kinked at the most fair level of effort. The most fair level of effort is the solution to the worker's utility maximization problem when the wage offer is low. However, as the wage increases, the marginal cost of reciprocating with higher effort (in terms of material payoff) grows relative to marginal benefit of higher effort (in terms of fairness). At some

point, when the wage offer is sufficiently large, the marginal utility costs of higher effort exceed the marginal utility benefits. Hence the worker will not exert more effort than some upper bound $\overline{e}$, no matter how high the wage.

Of course, the level of this upper bound depends on the degree of the worker's concern for fair transactions ϕ. If the worker is entirely selfish (that is, if $\phi = 0$), then $\overline{e} = -\infty$, and the worker will never exert effort. When $\phi > 0$, then $\overline{e} > -\infty$, and the worker would potentially reciprocate a higher wage with greater effort. The higher is ϕ, the higher is $\overline{e}$.

For a given wage, the worker's effort level depends on the reference transaction. For example, suppose the reference transaction is more favorable to the firm – the reference wage is lower or the the reference effort is higher. In that case, the worker's reference payoff is lower, and the firm's reference payoff is higher. For any given wage, the worker will perceive it to be fair to exert a higher level of effort.

Lemma 1 microfounds the positive dependence of effort on the wage assumed in Akerlof & Yellen's (1990) fair wage-effort theory of efficiency wages. The original theory simply posited that effort responds positively to the wage (Akerlof 1982; Akerlof & Yellen 1990). Being explicit about the preferences that underlay that relationship makes clear that limited fairness puts an upper bound on the extent of reciprocity and allows this paper to address questions of efficiency.

Akerlof (1982) and Akerlof & Yellen (1990) review sociological evidence that effort responds positively to the wage. Mas (2005) provides the most convincing field evidence that effort responds to plausibly exogenous wage changes. He examined compensation disputes between police offers and their city employers that were resolved by final-offer arbitration. In final-offer arbitration, each side submits a salary proposal, and the arbitrator must pick one of the proposals. Conditional on information in the proposals, the arbitrator's ruling is theoretically random, and Mas presents evidence consistent with that view. He also controls for observable differences between municipalities. He finds that when final-offer arbitration rules in favor of the police officers, exogenously raising pay, police initiate more and higher-quality arrests. When instead final-offer arbitration rules against the police officers (and in favor of their city employers), the number and quality of arrests declines.[6] Experimental economists have also found that higher wage offers induce greater

[6] 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.

Gneezy & List (2006) show that workers' reciprocity of greater effort for higher wages may be short-lived in some settings (data-entry and fundraising). See also List (2006), who shows that *sellers* in product markets may not exhibit a preference for fair transactions in some market settings. He finds that the same baseball card dealers who behave fairly when selling baseball cards in a laboratory experiment behave much more selfishly when selling baseball cards in a naturalistic market.

effort in laboratory labor markets with one-shot, anonymous interaction (e.g., Fehr, Kirchsteiger, & Riedl 1993; Fehr, Kirchsteiger, & Riedl 1998; Fehr & Falk 1999).

In a field experiment, Pritchard, Dunnette, & Jorgenson (1972) demonstrated that manipulating the reference transaction can affect workers' effort. The experimenters hired college students for a week to work at a company, and, holding the actual wage *constant*, made some workers feel "overpaid" (or "equitably-paid" or "underpaid," respectively) by telling them they were receiving higher pay than usual (or the same or lower pay, respectively).[7] Overpaid workers produced more output than equitably-paid workers, who produced more than underpaid workers. Also consistent with a preference for fair transactions but hard to reconcile with purely selfish motivations, equitably-paid workers reported more overall job satisfaction than underpaid or overpaid workers (see also Austin & Walster 1974).

Because effort responds positively to the wage, a profit-maximizing firm has an incentive to offer a high wage. The firm's first-order condition is

$$\frac{\partial \pi^F\left(w, e\left(w\right)\right)}{\partial w} + \frac{\partial \pi^F\left(w, e\left(w\right)\right)}{\partial e}\frac{de\left(w\right)}{dw} = 0.$$

Suppose the worker always chose the most fair level of effort. In that case, substituting for $\frac{de^{\text{fair}}(w)}{dw}$ and rearranging the firm's first-order condition, the equilibrium wage offer w^* solves

$$\frac{\partial \pi^F\left(w^*, e^{\text{fair}}\left(w^*\right)\right)/\partial w}{\partial \pi^F\left(w^*, e^{\text{fair}}\left(w^*\right)\right)/\partial e} = \frac{\partial \pi^W\left(w^*, e^{\text{fair}}\left(w^*\right)\right)/\partial w}{\partial \pi^W\left(w^*, e^{\text{fair}}\left(w^*\right)\right)/\partial e}. \tag{eff}$$

This is indeed the equilibrium if the associated effort level $e^{\text{fair}}\left(w^*\right)$ does not exceed the maximum effort $\overline{e}$ that the worker will actually exert in reciprocity for a high wage. Recall that this maximum will be greater the greater is the worker's concern for fair transactions. When this concern is strong enough, the equilibrium wage and effort satisfy (fair) and (eff). Hence the exchange is Pareto efficient.

Theorem 1 *(Rotten Firm theorem) Suppose there are potential gains from trade for all ϕ. Consider the subgame where the firm employs the worker. There exists $0 < \widehat{\widehat{\phi}} < \infty$ such that if $\phi \geq \widehat{\widehat{\phi}}$, the equilibrium wage-effort pair is Pareto efficient.*

Efficiency is the result of the *strategic* interaction between a purely selfish firm and a sufficiently fair-minded worker. Regardless of the firm's wage offer, the worker's choice of effort will ensure that the transaction is fair. Consequently, for a small change in the wage, the firm and the worker each

[7] The experimenters also made workers feel "overpaid," "equitably-paid," or "underpaid" by increasing, holding constant, or decreasing pay, respectively, halfway through the week.

get a positive share of the incremental change in the surplus from exchange. The firm therefore maximizes its own payoff by maximizing the total gains from trade.

As discussed in Section 2, a concern for fairness *per se* does not necessarily lead to efficiency. In a non-strategic setting where a fair-minded individual is choosing among alternative social allocations, a concern for fairness may lead to inefficiency. Efficiency in this strategic setting is an equilibrium phenomenon. At any other wage offer, the resulting wage-effort pair would be inefficient. The equilibrium occurs at the *only* wage-effort pair satisfying (fair) that is Pareto efficient.

Figure 3a illustrates the equilibrium. The tilted-V shapes are the worker's indifference curves when the weight on fairness is high. The three black downward-sloping curves correspond to three different wage offers $w_0 < w_1 < w_2$ the firm could make. The $w = w_0$ "wage curve," for example, shows the possible material payoff pairs $\left(\pi^W(w_0, e), \pi^F(w_0, e)\right)$ that could occur for different choices of effort e. For a given level of effort, a higher wage reduces the firm's material payoff and increases the worker's. For that reason, the $w = w_2$ wage curve is the same shape as the $w = w_1$ wage curve, except shifted downward and to the right. The upper envelope of all possible wage curves is the material-efficiency frontier. At the lowest wage offer w_0, the light point on the kink of the outside option indifference curve shows the effort the worker would choose. Wage curves corresponding to higher wage offers would intersect with the kink of a better indifference curve for the worker. The arrow shows how the material payoffs crawl up the equal-surplus line as the firm offers a higher and higher wage. At the $w = w_1$ wage curve, the material payoffs lie on the material-efficiency curve. For a wage higher than that (such as w_2), wage and effort are inefficiently high, and the material payoffs crawl back down the equal-surplus line. Notice that any wage offer that gives the firm a higher material payoff also gives the worker a higher material payoff and greater utility. The wage offer that gives the highest material payoff to the firm is $w = w_1$. The resulting equilibrium is Pareto efficient.

The Rotten Firm theorem does not depend on the assumption that fair transactions equate the surplus payoffs. What is crucial is that the maximally-fair choice of effort assigns to the firm a positive share of the marginal gains from trade (even if that share is not 50%). That is what makes it in the firm's interest to maximize the total gains from trade.[8]

[8] That is, what is important is that if two transactions are fair, then *both* parties must have higher surplus payoffs at one fair transaction compared to the other. Qualitatively, the results of this paper would be essentially the same if the fairness function were

$$f\left(\widetilde{\pi}^W, \widetilde{\pi}^F\right) = -\beta \max\left\{\widetilde{\pi}^W - g\left(\widetilde{\pi}^F\right), 0\right\} - (1-\beta)\max\left\{g\left(\widetilde{\pi}^F\right) - \widetilde{\pi}^W, 0\right\},$$

for some strictly increasing function g. This corresponds to a kind of non-linear inequity-aversion, advocated by Fehr & Schmidt (1999). See also Appendix B.

Unfortunately, if the worker's concern for fairness is not sufficiently large, then the worker will not be willing to reciprocate enough for the Rotten Kid theorem to hold. In that case, the firm's optimization problem has a "corner solution" at the reciprocity upper bound. The following proposition generalizes the Rotten Firm theorem, characterizing the equilibrium for any degree of fair-mindedness.

Proposition 4 *Suppose there are potential gains from trade for all ϕ. Consider the subgame where the firm employs the worker. There exist $0 < \widehat{\phi} < \widehat{\widehat{\phi}} < \infty$ such that:*

1. *If $\phi \in (0, \widehat{\phi}]$, then at the equilibrium transaction, the worker's utility $U(w^*, e^*)$ exactly equals his outside option level of utility $\overline{U}$.*

2. *If $\phi \in (\widehat{\phi}, \infty)$, then at the equilibrium transaction, the worker's utility $U(w^*, e^*)$ strictly exceeds his outside option level of utility $\overline{U}$. Moreover, (fair) is satisfied. Finally, if $\phi \in [\widehat{\widehat{\phi}}, \infty)$, then (eff) is also satisfied, and the equilibrium transaction is Pareto efficient.*

Figure 3b shows what happens when the worker's weight on fairness is in the middle range $\widehat{\phi} < \phi < \widehat{\widehat{\phi}}$. With this reduced concern for fairness, the lower part of the worker's V-shaped indifference curves are closer to vertical. Once again, consider three wage offers $w_0 < w_1' < w_2$ the firm could make (where $w_1' < w_1$). The arrow shows how the material payoffs crawl up the equal-surplus line as the firm offers a wage between w_0 and w_1'. The firm would like to offer an even higher wage (such as w_1) and induce higher effort, but the best it can do is get the reciprocity upper bound $\overline{e}$. Since effort remains at $\overline{e}$, a higher wage offer than w_1' results in a higher material payoff for the worker but a lower material payoff for the firm. That is why the arrow moves southeast off of the equal-surplus line for wage offers above w_1'. The firm maximizes its material payoff by offering wage $w = w_1' = \overline{w}$, the lowest wage that induces effort $\overline{e}$. Notice that in this case, the equilibrium is not Pareto efficient. However, whenever $\phi > \widehat{\phi}$, the equilibrium lies on the equal-surplus line, so (fair) is satisfied. Moreover, the worker ends up on a strictly better indifference curve than $U = \overline{U}$. Although the equilibrium is not fully efficient in this case, the firm and worker share the rents from exchange.

If the weight on fairness is in the lowest range, $\phi < \widehat{\phi}$, then the reciprocity upper bound $\overline{e}$ is extremely low. As a result, even if the firm offered wage $w = w_0$, the worker's effort $\overline{e}$ would put the material payoffs southeast of the equal-surplus line. The transaction $(w_0, \overline{e})$ makes the worker worse off than his outside option. In order to employ the worker, the firm will offer the lowest wage it can that the worker will accept, and the worker will exert effort $\overline{e}$. In equilibrium, the worker

will receive utility $\overline{U}$. Of course, although the proposition has focused on the subgame where the firm employs the worker, the firm will be unlikely to employ the worker in this case because it will have to offer a high wage in exchange for low effort. In this case where $\phi < \widehat{\phi}$, the equilibrium is qualitatively similar to when the worker's preferences are purely self-regarding.

The reference transaction influences the equilibrium wage and effort. If the reference transaction is more favorable to the firm (lower reference wage or higher reference effort), then the worker exerts greater effort for a given wage. The equilibrium wage will be lower, and if ϕ is large enough that the worker is not already exerting maximum effort, then equilibrium effort will be higher. Similarly, when the reference transaction is more favorable to the worker, the equilibrium transaction is also more favorable to the worker.

The Rotten Firm theorem says that the equilibrium transaction is Pareto efficient when the worker is sufficiently fair-minded, but how much is "sufficiently"? Appendix A presents rough calibrations based on existing laboratory evidence on social preferences. The question is, in real-world settings, for what proportion of individuals is it true that $\phi \geq \widehat{\widehat{\phi}}$? Unfortunately, the threshold $\widehat{\widehat{\phi}}$ (as well as $\widehat{\phi}$) in general depends not only on the worker's utility function, but also on other features of the exchange environment: the reference transaction and the functions $v(\cdot)$, $y(\cdot)$, and $c(\cdot)$, some of which are difficult to measure empirically. However, it turns out that if $v(w) = w$, then $\widehat{\widehat{\phi}}$ depends *only* on γ, a parameter of the utility function that has been estimated in existing work. Very rough calculations based on these existing estimates (Fehr & Schmidt 1999; Charness & Rabin 2002) suggest that $\phi \geq \widehat{\widehat{\phi}}$ for a sizeable minority of the population, perhaps 40%.

Unfortunately, it is not possible to be precise about the likelihood that $\phi > \widehat{\phi}$ without many additional assumptions, but Appendix A derives several qualitative results. $\phi > \widehat{\phi}$ will be true more often when the worker's outside option utility is smaller and when the reference transaction (and hence utility in equilibrium) is more favorable to the worker. In both cases, $U \geq \overline{U}$ is less likely to be a binding constraint for the firm's wage offer. Also, $\phi > \widehat{\phi}$ will be true more often when $y'(e)$ is large and $c'(e)$ is small at the fair wage-effort pair (w^-, e^-) that gives the worker his outside option utility. In that case, it is unlikely that the reciprocity upper bound $\overline{e}$ is as low as e^-. If the firm offers an incrementally larger wage than w^-, fairness requires very little additional effort from the worker beyond e^- because the firm benefits a lot from even a little bit more, and exerting that additional effort does not reduce the worker's material payoff by much.

The Rotten Firm theorem (and Proposition 4) might seem to suggest that, in terms of efficiency, a preference for fair transactions can substitute perfectly (or at least very well) for the availability of contracts. But the result presupposes that exchange occurs at all. The next section examines

this issue of when exchange occurs.

6 When Does Exchange Occur?

When does a profit-maximizing firm choose to employ a fair-minded worker? Exchange occurs when the firm earns higher profit from employing the worker than from its outside option. Potential gains from trade are a necessary condition. If there are no potential gains from trade, then by definition there are no wage-effort pairs that both the worker and the firm would prefer to their outside options.

However, potential gains from trade are not a sufficient condition for exchange to occur. The fact that there are potential gains from trade means that it is possible for employment to make both the firm and worker better off. But the employment transaction may leave the firm worse off than its outside option because the firm must share the rents from employment with the worker.

Proposition 5 *If exchange occurs, then there are potential gains from trade. Now suppose there are potential gains from trade for all ϕ. If exchange occurs when the worker's concern for fair transactions is ϕ', then exchange occurs when the worker's concern for fair transactions is $\phi'' > \phi'$.*

The last part of Proposition 5 states that the firm is more likely to employ the worker the stronger is the worker's concern for fair transactions.[9] As Figure 3 suggests, the more fair-minded the worker, the higher the firm's profit at the equilibrium employment transaction. In Figure 3b, a higher weight on fairness ϕ causes the equilibrium to lay farther northeast on the equal-surplus line, closer to the material-efficiency frontier, giving higher profit to the firm and higher utility to the worker. Therefore, a stronger concern for fairness both makes exchange more likely and makes it more efficient when it occurs.

Even when the firm prefers not to employ the worker, the worker would prefer to be employed. Because of the unexploited gains from trade, the worker would like to commit to accepting lower pay and exerting greater effort, if only the firm would hire him. That is, the worker would prefer to commit to a transaction that gives him a lower-than-equilibrium material payoff and that is disadvantageously unfair. But commitment is impossible because enforceable contracts are not available. After being hired, the worker would renege on his commitment and enforce a fair transaction by exerting less effort.

[9] The assumption that there are potential gains from trade for all ϕ is stronger than necessary. At the cost of substantially lengthening the proof, the same conclusion would follow from assuming that there are potential gains from trade for ϕ''.

The lower the reference wage and the higher the reference level of effort, the higher the firm's material payoff in equilibrium (and the lower the worker's). Consequently, a reference transaction that is more favorable to the firm makes it more likely that the firm will hire the worker.

There is a special circumstance under which sufficient fair-mindedness *does* perfectly substitute for the availability of enforceable contracts, in the sense that potential gains from trade will *always* be fully exploited. That is when the current market rate determines the worker's reference transaction for judging fairness, in addition to determining the firm's and worker's outside options. That is, the reference transaction gives the firm the same level of profit as its outside option and gives the worker the same level of utility as his outside option:

$$\pi^F(\widehat{w}, \widehat{e}) = \overline{\pi^F}$$

and

$$\pi^W(\widehat{w}, \widehat{e}) + \phi f\left(\pi^W(\widehat{w}, \widehat{e}) - \widehat{\pi}^W, \pi^F(\widehat{w}, \widehat{e}) - \widehat{\pi}^F\right) = \pi^W(\widehat{w}, \widehat{e}) = \overline{U}.$$

In that case, the fact that there are potential gains from trade implies that there are fair transactions that give the firm greater profit than its outside option.

Proposition 6 *If $\widehat{\pi}^F = \overline{\pi^F}$ and $\widehat{\pi}^W = \overline{U}$ and ϕ is sufficiently large and there are potential gains from trade, then exchange occurs.*

This result is illustrated in Figure 4. Because $\widehat{\pi}^F = \overline{\pi^F}$ and $\widehat{\pi}^W = \overline{U}$, the equal-surplus line, the worker's outside option indifference curve, and the $\pi^F = \overline{\pi^F}$ all intersect at a point. If there are potential gains from trade, then it must be that the material-efficiency curve intersects the equal-surplus line northeast of that point. Since the equilibrium transaction occurs at the intersection of the equal-surplus line and the material-efficiency curve, it gives higher payoffs to the worker and firm than their outside options. Hence exchange occurs.

This result is potentially of interest because market rates provide a natural benchmark against which to judge alternative transactions. To the extent that the market alone pins down the reference transaction, the efficiency-promoting properties of a preference for fairness are all the more striking. In reality, however, many factors appear to influence the reference transaction besides current market rates, such as past transactions between the same worker and firm or the treatment enjoyed by other workers, whose terms of employment may have been set under different market conditions (Kahneman, Knetsch, & Thaler 1986).

7 Profit-Sharing

The analysis so far has focused on the power of a preference for fair transactions to enable efficient exchange. However, efficiency implications are difficult to test. This section focuses on observable predictions that may be useful in assessing the theory empirically.

Attention to detail, friendliness to other employees, and friendliness to customers are aspects of output on which it can be difficult or costly to write an enforceable contract. The theory is primarily applicable to jobs where, on the margin, these non-contractible components of output are important to the firm. Many white collar, middle-management positions fall into this category. By contrast, the theory does not apply to jobs where the firm can easily contract on output, like assembly-line work. An obvious but important prediction is that implications of fairness-motivated effort – like "internal labor markets," reference transaction effects, profit-sharing, and rent-sharing, all discussed below – should co-occur in jobs that feature major non-contractible components of output.

Several predictions come out of the analysis already discussed. In contracting or bargaining theories, the worker's and the firm's outside options play an important role in wage determination. A worker whose outside option level of utility is higher must be paid more. By contrast, in the model from previous sections, if the worker's concern for fairness is strong enough, the outside options only matter for whether or not the firm employs the worker. The outside options are *irrelevant* in determining the equilibrium wage (and effort). In fact, wages often do seem to be relatively insensitive to "external labor market" conditions and much more sensitive to conditions internal to the firm, like profit levels (e.g., Beaudry & DiNardo 1991; Baker, Gibbs, & Holmstrom 1994).

The theory also predicts that the worker's reference transaction will be highly relevant for the equilibrium wage and effort. This is consistent with the fact that in wage negotiations, unions and management often expend substantial resources arguing over which terms of trade should serve as the relevant benchmark (e.g., Babcock & Loewenstein 1997).

In the remainder of this section, I show how a worker's preference for fair transactions explains a puzzle: the widespread use of profit-sharing plans. Over one-fifth of U.S. private-sector employees have their salaries supplemented by profit-sharing plans, gain-sharing plans, or stock options (Kruse et al 2003). These compensation schemes make the worker's wage a function of the firm's profit. Profit-sharing plans are strongly endorsed by both firms and workers in surveys (Weitzman & Kruse 1990). Compensation professionals and firms with profit-sharing plans believe these plans improve

productivity (e.g., Lawler 1987; Ehrenberg & Milkovich 1987), and existing evidence suggests they do (Weitzman & Kruse 1990; Kruse 1993). Of course, pay-for-performance is the classic formula for motivating effort. But it only makes sense as an incentive contract for top management, where the level of profit serves as an informative signal about effort. The puzzle is widespread profit-sharing for non-management employees, whose individual effect on the firm's profit is small. If these workers were selfish, they would free-ride off of the effort of others. Workers may monitor each other to enforce high effort, but there is a strong incentive also to free-ride on monitoring (Prendergast 1999).

The worker's preference for fairness explains why a firm would offer a profit-sharing compensation scheme. The worker perceives his "fair wage" to depend on how much the firm benefits from the worker's effort. When the firm offers a compensation scheme to the worker at the beginning of the year, there is uncertainty about what the firm's profit will be and hence about the worker's contribution to profit. By making the wage an increasing function of profit, the firm can ensure that the worker will feel fairly paid. Profit-sharing ensures that the worker will exert an efficient level of effort, regardless of how profitable the firm turns out to be.

I illustrate this logic with a simple extension of the previous analysis. The worker's material payoff function, $\pi^W(w,e) = v(w) - c(e)$, and preferences U remain the same as before. The firm's profit from hiring the worker now has a component that is unknown at the time the firm offers a compensation package to the worker. The shock to profit has mean zero and is drawn from a continuous distribution on $[\underline{\varepsilon}, \overline{\varepsilon}]$. For simplicity, I assume that the firm's material payoff function, denoted Π^F, is additively-separable across the deterministic and random components:[10]

$$\Pi^F(w,e) = \pi^F(w,e) + \varepsilon = y(e) - w + \varepsilon.$$

The firm offers a compensation scheme $w(\varepsilon)$ to the worker to maximize expected profit. If the worker accepts, ε is realized, and then the worker chooses effort. To keep the analysis as simple as possible, I restrict attention to what happens when both the firm and worker prefer employment to their deterministic outside options, $\overline{\pi^F}$ and $\overline{U}$, respectively.

Because the worker chooses effort after observing the shock, the most fair choice of effort $e^{\text{fair}}(w,\varepsilon)$ satisfies a simple modification of the deterministic "fairness rule":

$$\pi^W\left(w, e^{\text{fair}}\right) - \widehat{\pi}^W = \pi^F\left(w, e^{\text{fair}}\right) - \widehat{\pi}^F + \varepsilon. \qquad \text{(FAIR)}$$

[10] I could instead assume that $\Pi^F(w,e) = \varepsilon y(e) - w$ so that the shock affects the worker's productivity. The results would be largely similar, with equilibrium profit-sharing. The main difference is that the efficient level of effort would be increasing in the shock. This would create a force for the firm to offer an even higher wage in higher-ε states. This would change Proposition 7 in the following way: $e^{*\prime}(\varepsilon)$ would be of ambiguous sign for $\phi \in (\overline{\overline{\phi}}, \infty)$.

Fixing ε, the most fair choice of effort is increasing in the wage. Fixing the wage at a particular value $w(\varepsilon) = w$, the most fair choice of effort is decreasing in ε. If the firm gains more from any particular level of effort, then the worker perceives it as fair to exert less effort. As before, the worker's actual effort choice equals the most fair level of effort, up to some reciprocity upper bound $\overline{e}$ (which is independent of ε).

There is a unique profit-maximizing wage schedule $w^*(\varepsilon)$ that locks in the optimal wage for each possible realization of ε. As in the deterministic case, the character of the equilibrium depends on how strong is the worker's preference for fair transactions relative to thresholds $0 < \overline{\phi} < \overline{\overline{\phi}} < \infty$. As long as the worker cares enough about fairness, the profit-maximizing compensation schedule involves sharing the increment to profit between the worker and the firm: $0 < w^{*\prime}(\varepsilon) < 1$.

Proposition 7 *Suppose there are potential gains from trade for all (ϕ, ε). Consider the subgame where the firm employs the worker. There exist $0 < \overline{\phi} < \overline{\overline{\phi}} < \infty$ such that: If $\phi \in (\overline{\phi}, \infty)$, then the equilibrium wage schedule features profit-sharing: $0 < w^{*\prime}(\varepsilon) < 1$. If $\phi \in (\overline{\overline{\phi}}, \infty)$, then $e^{*\prime}(\varepsilon) < 0$, and the equilibrium transaction is Pareto efficient for any realization of ε.*

The profit-sharing result is illustrated in Figure 5a. For any given ε, the worker's effort choice $e(w)$ traces out an upward-sloping relationship between wage and effort, up to the reciprocity upper bound $\overline{e}$. The material-efficiency condition (eff) is a downward-sloping relationship between wage and effort. If the worker's concern for fairness is strong enough, then the intersection occurs before the effort function $e(w)$ becomes vertical at $e = \overline{e}$. The equilibrium transaction occurs at the intersection. A greater shock to profit corresponds to a leftward shift of the upward-sloping part of the effort function $e(w)$; the worker provides less effort for any given wage. The new equilibrium has a higher wage and less effort. The Rotten Firm theorem extends to this stochastic model. The equilibrium is Pareto efficient for every realization of ε.

The result is similar when the worker has a middling concern for fairnesss, $\phi \in (\overline{\phi}, \overline{\overline{\phi}})$, but the maximum level of effort $\overline{e}$ that the worker will exert is smaller. If the realization of ε is large enough, the equilibrium may be Pareto efficient and qualitatively like Figure 5a. Alternatively, for lower realizations of ε, the equilibrium can be qualitatively different. In that case, as Figure 5b shows, the intersection of (eff) and $e(w)$ cannot be the equilibrium because it involves the worker exerting only effort $\overline{e}$ while the firm offers a higher wage than necessary for $\overline{e}$. Instead, the equilibrium occurs at the lowest wage necessary to get the worker to exert $\overline{e}$. A greater shock to profit still corresponds to a leftward shift of the upward-sloping part of the effort function $e(w)$. Now the firm must offer a higher wage to maintain the same level of effort $\overline{e}$. Regardless of the worker's degree

of concern for fairness, realized payoffs – the firm's material payoff, the worker's material payoff, and the worker's utility – are increasing in the shock to profit.

I have interpreted the model in terms of profit-sharing: a particular firm making workers' compensation a function of realized profit. However, the same model can be applied to explain rent-sharing, the fact that more profitable firms pay higher wages to apparently identical workers, in contradiction to a competitive labor market model (e.g., Blanchflower, Oswald, & Sanfey 1996; Abowd, Kramarz, & Margolis 1999). To see how the above model explains rent-sharing, reinterpret ε as the *cross-sectional* distribution of profits across firms. Since $w^{*\prime}(\varepsilon) > 0$, the model predicts that more profitable firms pay higher wages to identical workers.

Of course, rent-sharing is often the result of a wage negotiation, where the negotiation outcome reflects relative bargaining power and the outside options of the worker and the firm. However, much of the documented rent-sharing occurs in non-unionized firms (e.g., Dickens & Katz 1987; Blanchflower, Oswald, & Sanfey 1996). Furthermore, when confronted with various possible explanations for their wage policies, managers generally endorse the idea that workers' effort responds to perceived fairness over other possibilities, such as implicit risk-sharing contracts or insider-outsider bargaining theory (e.g., Campbell & Kamlani 1997). The fact that rent-sharing also arises in anonymous, one-shot laboratory labor markets, where bargaining cannot occur, makes plausible the idea that fairness may be responsible for at least some of rent-sharing observed empirically (Fehr, Kirchsteiger, & Riedl 1993; Fehr, Kirchsteiger, & Riedl 1998; Fehr & Falk 1999; Brown, Falk, & Fehr 2004).

8 Rotten Firm vs Rotten Kid Theorems

Put aside the worker-firm relationship for a moment, and consider family dynamics. There are two players, a child and a parent. The child's initial income is y^c, and the parent's is $y^p >> y^c$. The child can take an action that increases family income, $y^p + y^c$, but reduces his own personal income y^c. Then the parent transfers some amount t of family income to the child. The child's material payoff function is $\pi^c(y^c + t)$, and the parent's is $\pi^p(y^p - t)$, both increasing functions. The child is purely selfish (a "rotten kid") and maximizes π^c. The parent is altruistic and maximizes $U(\pi^p, \pi^c)$, increasing in both arguments. The Rotten Kid theorem says that, even though the child is selfish, the child will take actions that maximize family income $y^p + y^c$, regardless of the effect on his own income y^c (Becker 1974). That is, the equilibrium is Pareto efficient.

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.

The intuitions are also analogous. The Rotten Kid theorem follows from the fact that the parent's transfer will ensure that the child's consumption (after the transfer) is increasing in family income. That gives the child an incentive to maximize family income. The Rotten Firm theorem follows from the fact that the worker's choice of the fair level of effort will ensure that the firm's profit is increasing in the surplus from exchange. The firm therefore seeks to maximize overall gains from trade. Both results entail that the solution to the second-mover's problem is interior rather than constrained. The Rotten Kid theorem requires that the parent's income is large enough relative to the child's so that the parent is not constrained in implementing her most-preferred transfer. The Rotten Firm theorem requires that the worker is sufficiently fair-minded so that the maximum effort the worker will offer to reciprocate is not binding, and the worker prefers to exert the fair level of effort.

However, there is an important difference. The Rotten Firm theorem holds quite generally in situations of exchange. By contrast, Bergstrom (1989) has shown that "transferable utility [i.e., material payoff]" is a key assumption underlying the Rotten Kid theorem and its extensions. As in Bergstrom's (1989) counterexamples, the existence of a second commodity introduces a failure of the "transferable utility" assumption. Since trade almost always involves at least two commodities, the Rotten Kid theorem will not typically apply in settings of exchange. If the worker is altruistic toward the firm, the transaction is not in general efficient. This section explores the implications of altruism toward the firm. The analysis helps to clarify the distinct nature of a preference for fair transactions.

In parallel with the last section, the firm offers a state-contingent wage to the worker, then a shock to profit is realized, and then the worker chooses effort. The firm's material payoff is $\Pi^F(w,e) = \pi^F(w,e) + \varepsilon$. The firm maximizes its expected material payoff. The worker's material payoff is $\pi^W(w,e)$, but the worker maximizes utility, which includes social considerations in addition to his own self-interest. In particular, assume now that the worker is altruistic toward the firm, putting positive weight on the firm's payoff. To keep things simple, suppose the worker's utility is additively-separable in his own and the firm's material payoff:

$$U = \pi^W + a\Pi^F, \tag{3}$$

where $0 \leq a \leq 1$ parameterizes the degree of altruism. The firm's and worker's outside options are

$\overline{\pi^F}$ and $\overline{U}$, respectively.[11]

With these preferences, a necessary condition for Pareto efficiency is material efficiency,

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

In this respect, altruistic preferences are like fair-minded preferences.

However, with altruistic preferences, the worker's equilibrium effort choice e^* satisfies

$$\frac{\partial \pi^W(w,e^*)}{\partial e} + a\frac{\partial \pi^F(w,e^*)}{\partial e} = 0. \tag{4}$$

With the additively-separable utility function (3), the worker's effort is independent of the wage and the shock to the firm's profit.

The firm therefore pays the lowest wage it can get away with, the smallest w^* satisfying

$$\pi^W(w^*,e^*) + a\left(\pi^F(w^*,e^*) + \varepsilon\right) = \overline{U}. \tag{5}$$

This makes the worker just willing to accept the firm's offer rather than his outside option.

The equilibrium wage-effort pair (w^*, e^*) satisfies (4) and (5). As long as this gives the firm a higher material payoff than its outside option, employment occurs in equilibrium. But notice that, except by lucky coincidence, there is no particular reason why (w^*, e^*) would satisfy (eff). In general, the equilibrium transaction is not Pareto efficient. Moreover, it can easily happen that no exchange occurs, even though there are potential gains from trade.

With an altruistic worker, the equilibrium wage is *decreasing* in the shock to the firm's profit, the opposite of profit-sharing (and rent-sharing). The reason is that an altruistic worker has higher utility when the firm's profit is higher. The firm can offer a lower wage and still have the worker accept employment.

Proposition 8 *Suppose the worker is altruistic and there are potential gains from trade for all ϕ. Consider the subgame where the firm employs the worker. Generically, the equilibrium transaction is not Pareto efficient. In equilibrium, the worker earns exactly his outside option level of utility, and $w^{*\prime}(\varepsilon) < 0$.*

This counterfactual implication supports the view that, in most cases, workers' altruism toward the firm is unlikely to play a major role in motivating effort.

[11] The assumption that $\overline{\pi^F}$ and $\overline{U}$ are constant keeps the model parallel with previous sections. However, it may be less realistic in the context of altruistic preferences. It means that the worker's utility is insensitive to the firm's payoff when the worker is not employed by the firm.

Notice that the altruistic utility function (3) can be written equivalently as

$$U = (1 - \alpha)\,\pi^W + a\left(\pi^W + \Pi^F\right).$$

In this formulation, the worker puts weight on his own material payoff and weight on a utilitarian social welfare function. In the experimental economics literature, this is often called a preference for "efficiency" (e.g., Charness & Rabin 2002; Engelmann & Strobel 2004). Altruism and a preference for efficiency are formally identical. Therefore, the results for altruism carry over directly. A preference for efficiency does not necessarily lead to efficient exchange.

When describing the "invisible hand" that allocates resources efficiently in competitive markets, Adam Smith (1776) wryly commented, "By pursuing his own interest [an individual] frequently promotes that of society more effectually than when he really intends to promote it." The results of this section point to a similar irony: By pursuing fair transactions, an individual may promote social efficiency more effectively than he would if he acted directly on a preference for efficiency.

9 Conclusions

A preference for fair transactions promotes efficient exchange – when it does not rule out exchange altogether. The fact that this preference enables efficient exchange might explain why it evolved biologically or culturally. The hypothesis that workers have such a preference is consistent with a variety of observations about labor markets, such as "internal labor markets," reference transaction effects, profit-sharing, and rent-sharing.

The same preference for fair transactions may also explain a wide variety of observations about product markets, housing markets, and marketing tactics. When costs increase, firms typically raise prices and often inform customers of the reason. A cost increase makes a price increase fair because the firm and consumer share the loss in material payoff. By contrast, firms often voluntarily maintain prices below market-clearing during temporary periods of high demand, leading to long lines or stockouts (e.g., Rotemberg 2005; Olmstead & Rhode 1985; Dacy & Kunreuther 1969). Raising prices would be perceived as unfair because the firm's profit would increase at the consumer's expense. Rent increases on new tenants are much more common than rent increases on existing tenants (Genesove 1999; Kahneman, Knetsch, & Thaler 1986). Rent increases on existing tenants will seem unfair to the extent that the previous rent serves as the reference transaction. Free samples not only inform potential new customers, but also make some feel obligated to purchase, even if they did not much like the product (Cialdini 1984).

Throughout the paper, I have assumed that only employees have a preference for fair transactions. The overall pull toward "fair" outcomes would be even stronger if employers also had a preference for fairness, at least if the worker and the firm share the same reference transaction. In fact, however, disagreements about the reference transaction are often important in real-world interactions. For example, when there are several reasonable precedents, negotiators seem able to convince themselves that the one most favorable to themselves is appropriate. This self-serving bias often causes negotiations to break down (Babcock & Loewenstein 1997). Similar problems could arise if only one party cares about fairness, but the selfish party does not know what the fair-minded party considers to be the reference transaction.

If a worker's reference transaction is influenced by what other workers at the same firm are paid, then the firm must take this fairness externality into account when setting a wage policy. If past pay influences a worker's reference transaction, then a preference for fair transactions will affect wage dynamics. The analysis in this paper has taken the reference transaction as exogenous. However, understanding the broader effects of fairness concerns on market equilibrium will require specifying how the reference transaction is determined.

10 Appendix A: Calibration

How fair-minded must a worker be for the equilibrium to be Pareto efficient, or at least to involve sharing the rents from exchange between the firm and the worker? This appendix provides a rough calibration of the model from Section 5 using estimates of preference parameters from behavior in laboratory experiments.

Recall from Proposition 4 that there exist $0 < \widehat{\phi} < \widehat{\widehat{\phi}} < \infty$ such that if $\phi > \widehat{\phi}$, then the worker earns strictly more than his outside option utility; and if $\phi \geq \widehat{\widehat{\phi}}$, then the equilibrium is Pareto efficient. By contrast, if $\phi \leq \widehat{\phi}$, then the worker earns exactly his outside option utility, and the equilibrium is qualitatively similar to what would arise from the standard, purely selfish model. The question is: In an actual situation of economic exchange, how likely is it that ϕ is larger than the thresholds $\widehat{\phi}$ and $\widehat{\widehat{\phi}}$?

From the proof of Proposition 4, the threshold $\widehat{\widehat{\phi}}$ is defined by

$$\widehat{\widehat{\phi}} \equiv \frac{1}{\gamma\left(\frac{y'(e^+)}{c'(e^+)}+1\right)},$$

where e^+ corresponds to effort level such that some transaction (w^+, e^+) gives the material payoff pair at the intersection of the equal-surplus line and the material efficiency frontier. In general, $\widehat{\widehat{\phi}}$ depends on $\frac{y'(e^+)}{c'(e^+)}$ (or equivalently $\frac{1}{v'(w^+)}$), which will vary from one setting to another. It may be challenging to measure relevant aspects of the exchange environment, such as the cost-of-effort function. However, calibration of $\widehat{\widehat{\phi}}$ becomes straightforward if the material payoff functions are quasi-linear in the wage: $v(w) = w$, so that $\pi^W = w - c(e)$ (and, as before, $\pi^F = y(e) - w$). In that case, there is a unique efficient level of effort satisfying $y'\left(e^{\text{eff}}\right) = c'\left(e^{\text{eff}}\right)$, which is independent of the wage. Since e^+ is the effort level at a materially-efficient transaction, it follows that $e^+ = e^{\text{eff}}$, so

$$\widehat{\widehat{\phi}} = \frac{1}{2\gamma}.$$

Hence $\widehat{\widehat{\phi}}$ depends *only* on the worker's preferences, not on any specific features of the situation like $y(\cdot)$, $c(\cdot)$, or the reference transaction. In this case, it is possible to assess how likely it is that $\phi \geq \widehat{\widehat{\phi}}$ by finding estimates of ϕ and γ from existing experimental work.

Recall that the worker's utility function is

$$U = \pi^W - \phi\gamma \max\left\{\widetilde{\pi}^W - \widetilde{\pi}^F, 0\right\} - \phi(1-\gamma)\max\left\{\widetilde{\pi}^F - \widetilde{\pi}^W, 0\right\}. \tag{6}$$

In laboratory experiments, participants are paid monetary amounts, say x^W and x^F, respectively. Assuming that in the laboratory, $\widetilde{\pi}^W \approx x^W$ and $\widetilde{\pi}^F \approx x^F$, the utility function (6) represents a

reparameterization of the utility function estimated from laboratory behavior by Fehr & Schmidt (1999):

$$U_{FS} = x^W - \beta \max\left\{x^W - x^F, 0\right\} - \alpha \max\left\{x^F - x^W, 0\right\}.$$

In particular, $\phi = \alpha + \beta$ and $\gamma = \frac{\beta}{\alpha+\beta}$. As a rough calibration to fit behavior in a wide variety of experimental games, Fehr & Schmidt (1999, Table III and p.864) argue that

$(\alpha, \beta) = (0, 0)$ (which implies $\left(\phi, \frac{1}{2\gamma}\right) = (0, \infty)$)	for about 30% of individuals
$(\alpha, \beta) = (0.5, 0.25)$ (which implies $\left(\phi, \frac{1}{2\gamma}\right) = (0.75, 1.5)$)	for about 30% of individuals
$(\alpha, \beta) = (1, 0.6)$ (which implies $\left(\phi, \frac{1}{2\gamma}\right) = (1.6, 1.33)$)	for about 30% of individuals
$(\alpha, \beta) = (4, 0.6)$ (which implies $\left(\phi, \frac{1}{2\gamma}\right) = (4.6, 3.83)$)	for about 10% of individuals

These estimates suggest that $\phi \geq \widehat{\widehat{\phi}}$ for about 40% of the subject population. However, it is important to keep in mind that this calculation is extremely rough.[12]

Based on a different set of experiments, Charness & Rabin (2002, Table VI, line 5) estimate the population average $(\alpha, \beta) \approx (0, 0.4)$, so $\left(\phi, \frac{1}{2\gamma}\right) \approx (0.4, 0.5)$. Although these point estimates imply that $\phi < \widehat{\widehat{\phi}}$ on average, they are consistent with $\phi \geq \widehat{\widehat{\phi}}$ for a sizeable minority of the population since there is substantial behavioral heterogeneity.

It is not straightforward to estimate $\widehat{\phi}$ without additional assumptions about the economic environment, but several qualitative conclusions are possible. From the proof of Proposition 4, the threshold $\widehat{\phi}$ is defined by

$$\widehat{\phi} \equiv \frac{1}{\gamma\left(\frac{y'(e^-)}{c'(e^-)} + 1\right)},$$

where e^- corresponds to smallest effort level such that some transaction (w^-, e^-) gives the material payoff pair at the intersection of the equal-surplus line and the worker's outside option indifference curve. A larger outside option utility $\overline{U}$ for the worker, or a more favorable reference transaction for the firm, implies that e^- is larger and therefore $\widehat{\phi}$ is larger. Clearly, the larger is $\frac{y'(e^-)}{c'(e^-)}$, the smaller is $\widehat{\phi}$. Roughly speaking, this is more likely to be true at small e^- if the production function is not very concave (y'' is small) and the cost-of-function is highly convex (c'' is large).

[12] For example, Shaked (2005) emphasizes that Fehr & Schmidt have little basis for estimating the *joint* distribution of α and β.

11 Appendix B: A Smooth Preference For Fair Transactions

In this appendix, I discuss the robustness of the Rotten Firm theorem to alternative ways of modeling a preference for fair transactions, and I present an analog to the Rotten Firm theorem. The fairness function (1) in the main text has a kink at each "fair" transaction (that equates the firm's and worker's surplus payoffs). Here I instead assume a smooth functional form. The property of the fairness function that captures the worker's preference for fairness is the fact that the firm's surplus payoff and the worker's surplus payoff enter in the worker's utility function as *complements*. The stronger the complementarity, the stronger the worker's concern for fair transactions, and the closer the equilibrium is to Pareto efficient. A kinked fairness function exhibits perfect complementarity and leads to full efficiency.

The model here also differs from the text in that it incorporates altruism (or, equivalently, a utilitarian preference for efficient transactions). Hence the analysis serves to demonstrate that the key intuitions are robust to a realistic mixture of these motivations, as long as the concern for fairness is sufficiently strong.

The setup is the same as Section 5. The firm offers a wage to the worker, and the worker chooses effort. I assume there are potential gains from trade and focus on what happens when exchange occurs.

The fairness function depends on the worker's and firm's surplus payoffs:

$$f\left(\widetilde{\pi}^W,\widetilde{\pi}^F\right)=\frac{-\ln\left((1-\theta)\exp\left\{-\kappa\widetilde{\pi}^W\right\}+\theta\exp\left\{-\kappa\widetilde{\pi}^F\right\}\right)}{\kappa}, \tag{7}$$

where $0\leq\theta\leq 1$ and $\kappa>0$. I call this a "constant-coefficient-of-complementarity" functional form because the parameter $\kappa=\frac{\frac{\partial^2 f}{\partial\widetilde{\pi}^W\partial\widetilde{\pi}^F}}{\frac{\partial f}{\partial\widetilde{\pi}^W}\frac{\partial f}{\partial\widetilde{\pi}^F}}>0$ measures the degree of complementarity.[13] It is more appropriate than a constant-elasticity-of-substitution function because it allows the surplus payoffs to take negative values.[14] The limit $\kappa\to 0$ is altruistic, $f\left(\widetilde{\pi}^W,\widetilde{\pi}^F\right)=(1-\theta)\,\widetilde{\pi}^W+\theta\widetilde{\pi}^F$. The limit $\kappa\to\infty$, $f\left(\widetilde{\pi}^W,\widetilde{\pi}^F\right)=\min\left\{\widetilde{\pi}^W,\widetilde{\pi}^F\right\}$, is a type of perfect complementarity that includes a utilitarian concern for social efficiency.[15] For intermediate values of κ, (7) exhibits complementarity as well as altruistic (or efficiency) concerns. As before, I assume the worker's utility is additively-

[13] I am grateful to Michael Ostrovsky for suggesting this functional form to me.

[14] The constant-elasticity-of-substitution (CES) and constant-coefficient-of-complementarity (CCC) functions are closely related. Consider a CES function, $f^{\text{CES}}\left(\widetilde{\pi}^W,\widetilde{\pi}^F\right)=\left((1-\psi)\left(\widetilde{\pi}^W\right)^\rho+\psi\left(\widetilde{\pi}^F\right)^\rho\right)^{\frac{1}{\rho}}$ (defined for $\widetilde{\pi}^W,\widetilde{\pi}^F\geq 0$) and a CCC function (7) (defined for all $\widetilde{\pi}^W,\widetilde{\pi}^F$). Notice that $f^{\text{CCC}}\left(\ln\widetilde{\pi}^W,\ln\widetilde{\pi}^F\right)=\ln f^{\text{CES}}\left(\widetilde{\pi}^W,\widetilde{\pi}^F\right)$, where $\theta=\psi$ and $\kappa=-\rho$. Various properties of the CCC function follow. For example, since $f^{\text{CES}}\left(\lambda\widetilde{\pi}^W,\lambda\widetilde{\pi}^F\right)=\lambda f^{\text{CES}}\left(\widetilde{\pi}^W,\widetilde{\pi}^F\right)$, it follows that $f^{\text{CCC}}\left(\widetilde{\pi}^W+\lambda,\widetilde{\pi}^F+\lambda\right)=f^{\text{CCC}}\left(\widetilde{\pi}^W,\widetilde{\pi}^F\right)+\lambda$.

[15] In fact, the min function corresponds to a particular mixture of the fairness function (1) from the main text with

separable in his own material payoff and fairness:

$$U = (1-\varphi)\,\pi^W + \varphi f\left(\pi^W - \widehat{\pi}^W, \pi^F - \widehat{\pi}^F\right), \tag{8}$$

where $0 < \varphi < 1$.[16]

The extent to which the equilibrium transaction is efficient depends on how strong the worker's concern for fair transactions is. Assuming the worker puts enough weight on fairness, the equilibrium is closer to Pareto efficient the stronger the complementarity κ in the worker's preferences.

Proposition 9 *Suppose there are potential gains from trade for all* (φ, κ)*. Consider the subgame where the firm employs the worker. There exists* $0 < \widehat{\widehat{\varphi}} < 1$ *such that if* $\varphi > \widehat{\widehat{\varphi}}$*, then in the limit* $\kappa \to \infty$*, the equilibrium transaction is Pareto efficient.*

This limit result implies that the Rotten Firm theorem holds not only with the fairness function (1) in the main text but also with the "Rawlsian" fairness function $f\left(\widetilde{\pi}^W, \widetilde{\pi}^F\right) = \min\left\{\widetilde{\pi}^W, \widetilde{\pi}^F\right\}$, advocated by some authors (Yaari & Bar-Hillel 1984; Charness & Rabin 2002).

---

a utilitarian preference for efficiency. To see this, notice that

$$\begin{aligned} f\left(\widetilde{\pi}^W, \widetilde{\pi}^F\right) &= -\beta \max\left\{\widetilde{\pi}^W - \widetilde{\pi}^F, 0\right\} - (1-\beta)\max\left\{\widetilde{\pi}^F - \widetilde{\pi}^W, 0\right\} + \beta\widetilde{\pi}^W + (1-\beta)\,\widetilde{\pi}^F \\ &= \beta \min\left\{\widetilde{\pi}^F - \widetilde{\pi}^W, 0\right\} + (1-\beta)\min\left\{\widetilde{\pi}^W - \widetilde{\pi}^F, 0\right\} + \beta\widetilde{\pi}^W + (1-\beta)\,\widetilde{\pi}^F \\ &= \beta \min\left\{\widetilde{\pi}^F, \widetilde{\pi}^W\right\} + (1-\beta)\min\left\{\widetilde{\pi}^W, \widetilde{\pi}^F\right\} \\ &= \min\left\{\widetilde{\pi}^F, \widetilde{\pi}^W\right\}. \end{aligned}$$

[16] The utility function in the main text (2) is parameterized differently, with $U = \pi^W + \phi f$. The fairness function (1) has a maximum value of zero, so the level of utility along the equal-surplus line does not depend on ϕ. By contrast, the fairness function (7) has no maximum value. By parameterizing utility as $U = (1-\varphi)\,\pi^W + \varphi f$, I can ensure that the level of utility along the equal-surplus line does not depend on φ in the limit $\kappa \to \infty$. If the utility function did not have this property, then it would make little sense to take the worker's outside option $\overline{U}$ as constant.

12 Appendix C: Proofs

12.1 Proof of Proposition 1

Note that the worker's utility function can be written

$$U = \begin{cases} \widehat{\pi}^W + (1-\phi\gamma)\,\widetilde{\pi}^W + \phi\gamma\widetilde{\pi}^F & \text{if } \widetilde{\pi}^W \geq \widetilde{\pi}^F \\ \widehat{\pi}^W + (1+\phi\,(1-\gamma))\,\widetilde{\pi}^W - \phi\,(1-\gamma)\,\widetilde{\pi}^F & \text{if } \widetilde{\pi}^W < \widetilde{\pi}^F \end{cases}.$$

In $\left(\widetilde{\pi}^W, \widetilde{\pi}^F\right)$-space, call $\widetilde{\pi}^W = \widetilde{\pi}^F$ the **equal-surplus line**. Above the equal-surplus line, U is increasing in $\widetilde{\pi}^W$ and decreasing in $\widetilde{\pi}^F$. Below the equal-surplus line, U is increasing in $\widetilde{\pi}^F$ and may be either increasing or decreasing in $\widetilde{\pi}^W$, depending on the sign of $(1-\phi\gamma)$. On the equal-surplus line, $U = \widetilde{\pi}^F = \widetilde{\pi}^W$. Call $\left\{\left(\pi^W(w,e), \pi^F(w,e)\right)\right\}_{(w,e)}$ the **material payoff possibility set**, and call the boundary (which satisfies (eff)) the **material-efficiency frontier**. Note that the material payoff possibility set is convex, and the material-efficiency frontier is downward-sloping.

To demonstrate the first part of the proposition, assume that (w,e) does not satisfy (eff), so that a neighborhood of $\left(\pi^W(w,e), \pi^F(w,e)\right)$ lies on the interior of the material payoff possibility set. There are three cases. If $\left(\widetilde{\pi}^W(w,e), \widetilde{\pi}^F(w,e)\right)$ lies below the equal-surplus line, then we can find $\left(\widetilde{\pi}^{W\prime}, \widetilde{\pi}^{F\prime}\right) = \left(\widetilde{\pi}^W(w',e'), \widetilde{\pi}^F(w',e')\right)$ such that $\widetilde{\pi}^{W\prime} > \widetilde{\pi}^W(w,e)$, $\widetilde{\pi}^{F\prime} = \widetilde{\pi}^F(w,e)$ and $\widetilde{\pi}^{W\prime} > \widetilde{\pi}^{F\prime}$. This implies that $U(w',e') > U(w,e)$ and $\widetilde{\pi}^F(w',e') = \widetilde{\pi}^F(w,e)$, so (w,e) is Pareto dominated. If $\left(\widetilde{\pi}^W(w,e), \widetilde{\pi}^F(w,e)\right)$ lies on the equal-surplus line, then (w,e) must be Pareto dominated by some other transaction (w',e') on the equal-surplus line with $\widetilde{\pi}^F(w',e') > \widetilde{\pi}^F(w,e)$ (and hence also $U(w',e') > U(w,e)$). Finally, if $\left(\widetilde{\pi}^W(w,e), \widetilde{\pi}^F(w,e)\right)$ lies above the equal-surplus line, then we can find $\left(\widetilde{\pi}^{W\prime}, \widetilde{\pi}^{F\prime}\right) = \left(\widetilde{\pi}^W(w',e'), \widetilde{\pi}^F(w',e')\right)$ such that $\widetilde{\pi}^{F\prime} > \widetilde{\pi}^F(w,e)$, $\widetilde{\pi}^{W\prime} = \widetilde{\pi}^W(w,e)$ and $\widetilde{\pi}^{W\prime} < \widetilde{\pi}^{F\prime}$. This implies that $U(w',e') > U(w,e)$ and $\widetilde{\pi}^F(w',e') = \widetilde{\pi}^F(w,e)$, so (w,e) is Pareto dominated.

For the second part, assume that (w,e) satisfies (eff) and $\widetilde{\pi}^W(w,e) < \widetilde{\pi}^F(w,e)$ but that (w,e) is not Pareto efficient. Then there is some (w',e') such that $\widetilde{\pi}^F(w',e') \geq \widetilde{\pi}^F(w,e)$ and $U(w',e') \geq U(w,e)$, with at least one inequality strict. Since $\widetilde{\pi}^F(w',e') \geq \widetilde{\pi}^F(w,e)$ and the material-efficiency frontier is downward-sloping, $\left(\widetilde{\pi}^W(w',e'), \widetilde{\pi}^F(w',e')\right)$ must lie above the equal-surplus line, where U is increasing in $\widetilde{\pi}^W$ and decreasing in $\widetilde{\pi}^F$. This and $U(w',e') \geq U(w,e)$ imply that $\widetilde{\pi}^W(w',e') \geq \widetilde{\pi}^W(w,e)$, with strict inequality if $\widetilde{\pi}^F(w',e') = \widetilde{\pi}^F(w,e)$, which contradicts the assumption that (w,e) satisfies (eff).

12.2 Proof of Proposition 2

Assume there are potential gains from trade when the worker's weight on fairness is ϕ. Then there is some (w,e) such that $\pi^F(w,e) \geq \overline{\pi^F}$ and $U(w,e) \geq \overline{U}$. When the worker's weight on fairness is $\phi < \phi'$, his utility from the transaction (w,e) is

$$\begin{aligned} & \pi^W(w,e) - \phi\gamma \max\left\{\widetilde{\pi}^W - \widetilde{\pi}^F, 0\right\} - \phi(1-\gamma)\max\left\{\widetilde{\pi}^F - \widetilde{\pi}^W, 0\right\} \\ \leq\ & \pi^W(w,e) - \phi'\gamma \max\left\{\widetilde{\pi}^W - \widetilde{\pi}^F, 0\right\} - \phi'(1-\gamma)\max\left\{\widetilde{\pi}^F - \widetilde{\pi}^W, 0\right\} \end{aligned}$$

which is the worker's utility from the transaction when his weight on fairness is ϕ'. So when the worker's weight on fairness is ϕ', $\pi^F(w,e) \geq \overline{\pi^F}$ and $U(w,e) \geq \overline{U}$, and there are potential gains from trade.

12.3 Proof of Proposition 3

Assume that exchange (w,e) occurs and is not Pareto efficient. Then for some (w',e'), $\widetilde{\pi}^F(w',e') \geq \widetilde{\pi}^F(w,e)$ and $U(w',e') \geq U(w,e)$, with at least one inequality strict. If $\widetilde{\pi}^F(w',e') > \widetilde{\pi}^F(w,e)$ and $U(w',e') \geq U(w,e)$, the firm would strictly prefer (and the worker would accept) (w',e') to (w,e), so exchange cannot occur at (w,e). We are left with the case $\widetilde{\pi}^F(w',e') = \widetilde{\pi}^F(w,e)$ and $U(w',e') > U(w,e)$. It is easy to check that both $\widetilde{\pi}^F$ and U are locally non-satiated in (w,e)-space, which implies that for some (w'',e'') close to (w',e'), $\widetilde{\pi}^F(w'',e'') > \widetilde{\pi}^F(w,e)$ and $U(w'',e'') \geq U(w,e)$. Again, the firm would strictly prefer (and the worker would accept) (w'',e'') to (w,e), so exchange cannot occur at (w,e). By contradiction, any exchange that occurs is Pareto efficient. Next, assume towards contradiction that exchange occurs with $U(w,e) > \overline{U}$. Again using local non-satiation, for some (w',e') close to (w,e), $\widetilde{\pi}^F(w',e') > \widetilde{\pi}^F(w,e)$ and $U(w',e') \geq \overline{U}$. Thus the firm would strictly prefer (and the worker would accept) (w',e') to (w,e), and exchange does not occur at (w,e). The claim that exchange occurs if and only if there are potential gains from trade is obvious.

12.4 Proof of Lemma 1

The worker's utility function can be written

$$U(w,e) = \begin{cases} \widehat{\pi}^W + (1-\phi\gamma)\,\widetilde{\pi}^W(w,e) + \phi\gamma\widetilde{\pi}^F(w,e) & \text{if } e \leq e^{\text{fair}}(w) \\ \widehat{\pi}^W + (1+\phi(1-\gamma))\,\widetilde{\pi}^W(w,e) - \phi(1-\gamma)\,\widetilde{\pi}^F(w,e) & \text{if } e > e^{\text{fair}}(w) \end{cases}.$$

It is easy to check that $U(w,e)$ is strictly concave in e and that $\lim_{e\to\pm\infty} U(w,e) = -\infty$. Thus the problem $e(w) = \arg\max_e U(w,e)$ has a unique solution. $U(w,e)$ is strictly decreasing in e on

$e > e^{\text{fair}}(w)$, so $e(w) \leq e^{\text{fair}}(w)$. Define $\overline{e}$ by

$$-(1-\phi\gamma)\,c'(\overline{e}) + \phi\gamma y'(\overline{e}) = 0.$$

If $\overline{e} < e^{\text{fair}}(w)$, then clearly $e(w) = \overline{e}$. Otherwise, the only candidate for $e(w)$ is $e^{\text{fair}}(w)$. Since $e^{\text{fair}}(w)$ is increasing in w, the lemma follows. Note that $\lim_{w\to\pm\infty} e^{\text{fair}}(w) = \pm\infty$, so $\overline{w}$ is well-defined.

12.5 Proof of Theorem 1

A special case of Proposition 4.

12.6 Proof of Proposition 4

Define $\left(\widetilde{\pi}^{-}, \widetilde{\pi}^{-}\right)$ to be the intersection (in terms of surplus payoffs) between the equal-surplus line and the set $\left\{\left(\pi^W, \pi^F\right) : U\left(\pi^W, \pi^F\right) = \overline{U}\right\}$ which gives the worker his outside option level of utility (see Figure 2). Define $\left(\widetilde{\pi}^{+}, \widetilde{\pi}^{+}\right)$ to be the intersection of the equal-surplus line and the material-efficiency frontier (see Figure 1). Note that $\widetilde{\pi}^{-}$ and $\widetilde{\pi}^{+}$ are independent of ϕ.

It follows from Lemma 1 that any equilibrium transaction must give a material payoff pair $\left(\pi^W, \pi^F\right)$ that lies on or below the equal-surplus line. We now proceed in 9 steps.

1. $\widetilde{\pi}^{+} > \widetilde{\pi}^{-}$. Assume otherwise, that $\widetilde{\pi}^{+} \leq \widetilde{\pi}^{-}$. Consider first the case where $\widetilde{\pi}^{+} < \widetilde{\pi}^{-}$. Then we may easily check that for $1-\phi\gamma < 0$, any surplus payoff pair $\left(\widetilde{\pi}^{W}, \widetilde{\pi}^{F}\right)$ with $U\left(\widetilde{\pi}^{W}, \widetilde{\pi}^{F}\right) \geq \overline{U}$ has $\left(\widetilde{\pi}^{W}, \widetilde{\pi}^{F}\right) \geq \left(\widetilde{\pi}^{-}, \widetilde{\pi}^{-}\right)$. But any attainable material payoff pair $\left(\widetilde{\pi}^{W\prime}, \widetilde{\pi}^{F\prime}\right)$ lies on or under the material-efficiency frontier, and so has $\widetilde{\pi}^{W\prime} \leq \widetilde{\pi}^{+} < \widetilde{\pi}^{-}$ or $\widetilde{\pi}^{F\prime} \leq \widetilde{\pi}^{+} < \widetilde{\pi}^{-}$. It follows that the material payoff possibility set has null intersection with the set of material payoff pairs that give the worker at least his outside option utility. Thus there are no potential gains from trade for sufficiently large ϕ. We may attack the case $\widetilde{\pi}^{+} = \widetilde{\pi}^{-}$ similarly to show that when $1-\phi\gamma < 0$, the material payoff possibility set with the set of material payoff pairs that give the worker at least his outside option utility is the singleton $\left(\widetilde{\pi}^{-}, \widetilde{\pi}^{-}\right)$. Thus the worker cannot receive more than his outside option utility level, and there are no potential gains from trade.

2. *Any material payoff pair $\left(\pi^W, \pi^F\right)$ that lies on the material-efficiency frontier in $\left(\pi^W, \pi^F\right)$-space corresponds to exactly one transaction, while any material payoff pair that lies below the material-efficiency frontier can be generated by exactly two transactions.* To see this, fix $\pi^{F\prime}$

and define $e\left(\pi^{F\prime}, w\right)$ to be the unique effort level such that $\pi^F\left(w, e\left(\pi^{F\prime}, w\right)\right) = \pi^{F\prime}$. This gives us $e\left(\pi^{F\prime}, w\right) = y^{-1}\left(\pi^{F\prime} + w\right)$. Notice that

$$\pi^W\left(w, e\left(\pi^{F\prime}, w\right)\right) = v\left(w\right) - c\left(y^{-1}\left(\pi^{F\prime} + w\right)\right)$$

is concave with $\lim_{w\to\pm\infty} \pi^W\left(w, e\left(\pi^{F\prime}, w\right)\right) = -\infty$. The observation follows immediately.

3. *If some equilibrium material payoff pair lies below the equal-surplus line, then the worker earns his outside option level of utility, and the firm's material payoff is less than* $\widetilde{\pi}^-$. Note that Lemma 1 implies that any equilibrium transaction with a material payoff pair below the equal-surplus line has effort level $\overline{e}$. Define $w^{\text{fair}}\left(e\right) \equiv \left(e^{\text{fair}}\right)^{-1}\left(e\right)$. Assume towards a contradiction that some equilibrium material payoff pair $\left(\pi^W\left(w, \overline{e}\right), \pi^F\left(w, \overline{e}\right)\right)$ lies below the equal-surplus line, but the firm's surplus payoff exceeds $\widetilde{\pi}^-$. Then since $\widetilde{\pi}^W\left(w, \overline{e}\right) > \widetilde{\pi}^F\left(w, \overline{e}\right)$, we must have $\pi^F\left(w^{\text{fair}}\left(\overline{e}\right), \overline{e}\right) > \pi^F\left(w, \overline{e}\right)$, so the firm would strictly prefer $\left(w^{\text{fair}}\left(\overline{e}\right), \overline{e}\right)$ to $\left(w, \overline{e}\right)$. Since $\widetilde{\pi}^F\left(w, \overline{e}\right) > \widetilde{\pi}^-$ and $U\left(\widetilde{\pi}, \widetilde{\pi}\right)$ is strictly increasing in $\widetilde{\pi}$, we have $U\left(\widetilde{\pi}^W\left(w^{\text{fair}}\left(\overline{e}\right), \overline{e}\right), \widetilde{\pi}^F\left(w^{\text{fair}}\left(\overline{e}\right), \overline{e}\right)\right) > U\left(\widetilde{\pi}^-, \widetilde{\pi}^-\right) = \overline{U}$. Moreover since $\overline{e} = \arg\max_e U\left(w^{\text{fair}}\left(\overline{e}\right), e\right)$, the worker will accept $\left(w^{\text{fair}}\left(\overline{e}\right), e\right)$. This implies that $\left(w, \overline{e}\right)$ cannot be an equilibrium transaction.

   Next, assume towards a contradiction that an equilibrium material payoff pair $\left(\pi^W\left(w, \overline{e}\right), \pi^F\left(w, \overline{e}\right)\right)$ lies below the equal-surplus line, but that the worker exceeds his outside option level of utility. Choose $w' < w$ so that $U\left(w', \overline{e}\right) > \overline{U}$ and $\pi^F\left(w', \overline{e}\right) > \pi^F\left(w, \overline{e}\right)$ but $w' > w^{\text{fair}}\left(\overline{e}\right)$. The firm strictly prefers $\left(w', \overline{e}\right)$ to $\left(w, \overline{e}\right)$. Since $\overline{e} = \arg\max_e U\left(w', e\right)$, the worker accepts $\left(w', \overline{e}\right)$. This implies that $\left(w, \overline{e}\right)$ cannot be an equilibrium transaction. By contradiction, the observation holds.

4. *An equilibrium transaction* $\left(w', e'\right)$ *lies on the equal-surplus line if and only if there exists some (possibly distinct) transaction* $\left(w, e^{\text{fair}}\left(w\right)\right)$ *such that (1)* $e^{\text{fair}}\left(w\right) = \arg\max_e U\left(w, e\right)$*, and (2)* $\left(w, e^{\text{fair}}\left(w\right)\right)$ *gives surplus payoffs* $\left(\widetilde{\pi}^-, \widetilde{\pi}^-\right)$. We first prove the if direction. Assume there exists some $\left(w, e^{\text{fair}}\left(w\right)\right)$ that satisfies (1) and (2). Suppose $\left(w', e'\right)$ gives material payoffs lying below the equal-surplus line. Step 3 shows that the firm's equilibrium surplus payoff is less than $\widetilde{\pi}^-$. Thus the firm would strictly prefer $\left(w, e^{\text{fair}}\left(w\right)\right)$ to the equilibrium transaction. The worker would accept $\left(w, e^{\text{fair}}\left(w\right)\right)$ because $e^{\text{fair}}\left(w\right) = \arg\max_e U\left(w, e\right)$ and $U\left(\widetilde{\pi}^-, \widetilde{\pi}^-\right) = \overline{U}$. Thus the equilibrium transaction cannot be an equilibrium, a contradiction. We now prove the only if direction. Assume $\left(w', e'\right)$ lies on the equal-surplus line. Note that $e' = e^{\text{fair}}\left(w'\right)$. If $\left(w', e'\right)$ gives surplus payoffs $\left(\widetilde{\pi}^-, \widetilde{\pi}^-\right)$, then we are finished, so suppose it

does not. Since there are potential gains from trade, it must be that (w', e') gives surplus payoffs $\left(\widetilde{\pi}', \widetilde{\pi}'\right) > \left(\widetilde{\pi}^-, \widetilde{\pi}^-\right)$. Step 1 implies that for some $w < w'$, we have

$$\left(\widetilde{\pi}^W\left(w, e^{\text{fair}}(w)\right), \widetilde{\pi}^F\left(w, e^{\text{fair}}(w)\right)\right) = \left(\widetilde{\pi}^-, \widetilde{\pi}^-\right).$$

Also, since $e^{\text{fair}}(w') = \arg\max_e U(w', e)$, we have

$$-(1-\phi\gamma)\, c'\left(e^{\text{fair}}\left(w'\right)\right) + \phi\gamma y'\left(e^{\text{fair}}\left(w'\right)\right) \geq 0.$$

Since $e^{\text{fair}}(w) < e^{\text{fair}}(w')$,

$$-(1-\phi\gamma)\, c'\left(e^{\text{fair}}(w)\right) + \phi\gamma y'\left(e^{\text{fair}}(w)\right) \geq 0.$$

Thus $e^{\text{fair}}(w) = \arg\max_e U(w, e)$.

5. $0 < \widehat{\phi} < \infty$ *exists such that (1) if* $\phi \in \left(0, \widehat{\phi}\right)$*, any equilibrium payoff lies below the equal-surplus line, and (2) if* $\phi \in \left[\widehat{\phi}, \infty\right)$*, any equilibrium payoff lies on the equal-surplus line.* Let $\left(w^{\text{fair}}\left(e^-\right), e^-\right)$ and $\left(w^{\text{fair}}\left(e'\right), e'\right)$ be the two transactions corresponding to the surplus payoff pair $\left(\widetilde{\pi}^-, \widetilde{\pi}^-\right)$. Without loss of generality, assume $e^- \leq e'$. Define

$$\widehat{\phi} \equiv \frac{c'\left(e^-\right)}{\gamma\left(y'\left(e^-\right) + c'\left(e^-\right)\right)}.$$

Note that the conditions for step 4 are satisfied if and only if

$$-(1-\phi\gamma)\, c'\left(e^-\right) + \phi\gamma y'\left(e^-\right) \geq 0;$$

equivalently, only if $\phi \geq \widehat{\phi}$. The result follows immediately.

6. *If* $\phi \in \left(\widehat{\phi}, \infty\right)$*, then the worker's utility exceeds his outside option utility at the equilibrium transaction.* Note that since $\phi > \widehat{\phi}$, there is a transaction corresponding to the surplus payoff pair $\left(\widetilde{\pi}^-, \widetilde{\pi}^-\right)$ such that

$$-(1-\phi\gamma)\, c'(e) + \phi\gamma y'(e) > 0.$$

Denote this transaction by $\left(w^{\text{fair}}(e), e\right)$. Choose $e' > e$ such that

$$-(1-\phi\gamma)\, c'\left(e'\right) + \phi\gamma y'\left(e'\right) \geq 0.$$

Then $w^{\text{fair}}\left(e'\right) > w^{\text{fair}}(e)$, so

$$\widetilde{\pi}^F\left(w^{\text{fair}}\left(e'\right), e'\right) > \widetilde{\pi}^F\left(w^{\text{fair}}(e), e\right) = \widetilde{\pi}^-.$$

The firm strictly prefers $\left(w^{\text{fair}}\left(e'\right), e'\right)$ to $\left(w^{\text{fair}}\left(e\right), e\right)$. Our choice ensures that $e' = \arg\max_e U\left(w^{\text{fair}}\left(e'\right), e\right)$ and $U\left(w^{\text{fair}}\left(e'\right), e'\right) > \overline{U}$, so the worker accepts $\left(w^{\text{fair}}\left(e'\right), e'\right)$. Thus a fortiori the *equilibrium* payoff must earn the firm a surplus payoff strictly greater than $\widetilde{\pi}^{-}$. Since the equilibrium surplus payoff pair lies on the equal-surplus line, the worker's utility strictly exceeds $U\left(\widetilde{\pi}^{-}, \widetilde{\pi}^{-}\right) = \overline{U}$.

7. Define (w^{+}, e^{+}) to be the (unique) transaction that corresponds to $\left(\widetilde{\pi}^{+}, \widetilde{\pi}^{+}\right)$. *Then* (w^{+}, e^{+}) *is Pareto efficient, and* (w^{+}, e^{+}) *is the equilibrium transaction if and only if*

$$\phi \geq \frac{c'\left(e^{+}\right)}{\gamma\left(c'\left(e^{+}\right) + y'\left(e^{+}\right)\right)} \equiv \widehat{\widetilde{\phi}}.$$

It follows from Proposition 1 that the transaction (w^{+}, e^{+}) is Pareto efficient. From Lemma 1, any transaction that the worker accepts must give surplus payoffs that lie on or below the equal-surplus line. Thus the firm cannot do better than $\widetilde{\pi}^{+}$. Note that $e^{+} = e^{\text{fair}}\left(w^{+}\right)$. The firm will offer wage w^{+} whenever the employee responds with $e^{\text{fair}}\left(w^{+}\right)$. This occurs if and only if

$$-\left(1 - \phi\gamma\right) c'\left(e^{+}\right) + \phi\gamma y'\left(e^{+}\right) \geq 0,$$

or equivalently $\phi \geq \widehat{\widetilde{\phi}}$.

8. $\pi^{F}\left(w, e^{\text{fair}}\left(w\right)\right)$ *is strictly concave in* w *and has a unique maximum (which is* w^{+}*).* Straightforward calculations reveal that

$$\frac{de^{\text{fair}}\left(w\right)}{dw} = \frac{1 + v'\left(w\right)}{y'\left(e\right) + c'\left(e\right)}$$

and that

$$\frac{d^2 e^{\text{fair}}\left(w\right)}{dw^2} = \frac{v''\left(w\right)}{y'\left(e\right) + c'\left(e\right)} - \frac{1 + v'\left(w\right)}{\left(y'\left(e\right) + c'\left(e\right)\right)^2}\left(y''\left(e\right) + c''\left(e\right)\right)\frac{de^{\text{fair}}\left(w\right)}{dw}$$

so

$$\begin{aligned}
\frac{d^2\pi^{F}\left(w, e^{\text{fair}}\left(w\right)\right)}{dw^2} &= y''\left(e\right)\left(\frac{de^{\text{fair}}\left(w\right)}{dw}\right)^2 + y'\left(e\right)\frac{d^2 e^{\text{fair}}\left(w\right)}{dw^2} \\
&= \frac{y''\left(1 + v'\right)^2 c' + y' v''\left(r' + c'\right)^2 - y'\left(1 + c'\right)^2 c''}{\left(y' + c'\right)^3} \\
&< 0.
\end{aligned}$$

As shown in the text, the first-order condition of the firm's maximization problem implies that $\left(w^{*}, e^{\text{fair}}\left(w^{*}\right)\right)$ is materially-efficient. But that implies that the unique maximum is $w^{*} = w^{+}$.

9. $0 < \widehat{\phi} < \widehat{\widehat{\phi}} < \infty$. We already know that $0 < \widehat{\phi}, \widehat{\widehat{\phi}} < \infty$. Step 8 showed that $\widetilde{\pi}^F\left(w, e^{\text{fair}}(w)\right)$ is strictly concave in w and has a unique maximum at w^+, which implies that $\widetilde{\pi}^F\left(w^{\text{fair}}(e), e\right)$ is strictly quasiconcave in e and has a unique maximum of $\widetilde{\pi}^+$ at e^+. Since $\widetilde{\pi}^+ > \widetilde{\pi}^-$, it follows from the definition of e^- that $e^+ > e^-$. Recall $\widehat{\widehat{\phi}} \equiv \frac{c'(e^+)}{\gamma(c'(e^+)+y'(e^+))}$ and $\widehat{\phi} \equiv \frac{c'(e^-)}{\gamma(y'(e^-)+c'(e^-))}$, and note that $\frac{c'(e)}{\gamma(y'(e)+c'(e))} = \frac{1}{\gamma\left(\frac{y'(e)}{c'(e)}+1\right)}$ is strictly increasing in e. Hence $\widehat{\widehat{\phi}} > \widehat{\phi}$.

10. *We complete the proof by noting that the equilibrium transaction is unique.* We first consider the case $\phi \geq \widehat{\phi}$. Since we know the equilibrium material payoff pair lies on the equal-surplus line, the firm's profit-maximization problem becomes $\max_w \pi^F\left(w, e^{\text{fair}}(w)\right)$, subject to the constraint $-(1-\phi\gamma)c'\left(e^{\text{fair}}(w)\right) + \phi\gamma y'\left(e^{\text{fair}}(w)\right) \geq 0$. Recall from step 8 that $\pi^F\left(w, e^{\text{fair}}(w)\right)$ is concave in w and has a unique maximum, so we can restrict the maximization problem to a compact wage interval. This immediately implies that an equilibrium transaction exists. If the constraint is not binding, then step 8 has already shown uniqueness. If the constraint is binding, then it uniquely defines the equilibrium wage and effort level. (Note that this implies $e = \overline{e}(\phi)$ for $\phi \in (\widehat{\phi}, \widehat{\widehat{\phi}}]$.)

    Next we consider the case $\phi < \widehat{\phi}$. As we will discuss in further detail in the proof of Proposition 5, the equilibrium transaction is $(w(\phi), \overline{e}(\phi))$ where $w(\phi)$ satisfies (11). Noting that $\left(w\left(\widehat{\phi}\right), \overline{e}\left(\widehat{\phi}\right)\right)$ is the unique equilibrium transaction for $\widehat{\phi}$ and observing as in the proof of Proposition 5 that (11) satisfies Lipschitz conditions on $\left[w(\phi), w\left(\widehat{\phi}\right)\right] \times \left[\phi, \widehat{\phi}\right]$, we may invoke Picard's theorem and infer that the equilibrium transaction $(w(\phi), \overline{e}(\phi))$ is unique.

12.7 Proof of Proposition 5

If there are no potential gains from trade, then by definition, no transaction makes both firm and worker better off than their outside options. Exchange cannot occur. We prove the second part in three cases.

First, suppose $\widehat{\phi} \leq \phi' < \phi''$. Let (w, e) denote the equilibrium transaction that occurs at ϕ'. It is easy to check that (w, e) is a feasible contract for ϕ'', and it gives the same material payoff to the firm and the same utility to the worker. (The fairness function takes the value zero because the equilibrium material payoff pair lies on the equal-surplus line.) Since the equilibrium payoffs exceeded the outside option payoffs at ϕ', they do so at ϕ''.

Second, suppose $\phi' < \widehat{\phi}, \phi'' \geq \widehat{\phi}$. Then at ϕ', the equilibrium material payoff pair lies below the equal-surplus line. The worker's utility is $\overline{U}$, and the firm's surplus payoff is less than $\widetilde{\pi}^-$. At $\phi'' > \widehat{\phi}$, the equilibrium transaction gives the worker utility at least $\overline{U}$. Since $\left(\widetilde{\pi}^-, \widetilde{\pi}^-\right)$ is a feasible

surplus payoff pair that the worker would accept, the firm's surplus payoff in equilibrium must be at least $\widetilde{\pi}^{-}$.

Finally, suppose $\phi' < \phi'' < \widehat{\phi}$. Then at ϕ', the worker's utility is $\overline{U}$, and the firm's surplus payoff is less than $\widetilde{\pi}^{-}$. At ϕ'', we claim that there is a feasible transaction (w, e) where the worker receives his outside option utility. For a given $\phi < \widehat{\phi}$, this occurs if and only if (1) $e = \overline{e}(\phi)$ (since the transaction payoff pair must lie below the equal-surplus line), (2) and the transaction gives the worker utility $U = \overline{U}$ and gives the firm a higher material payoff than $\overline{\pi^F}$. We can write,

$$\overline{U} = \widehat{\pi}^W + (1 - \phi\gamma)\,\widetilde{\pi}^W\left(w, \overline{e}(\phi)\right) + \phi\gamma\widetilde{\pi}^F\left(w, \overline{e}(\phi)\right) \tag{9}$$

$$\overline{\pi^F} - \widehat{\pi}^F \leq \widetilde{\pi}^F\left(w, \overline{e}(\phi)\right) \leq \widetilde{\pi}^{-}. \tag{10}$$

Define $w(\phi)$ so that $(w(\phi), \overline{e}(\phi))$ satisfies (9). We assume that $w\left(\phi'\right)$ exists and $\widetilde{\pi}^F\left(w\left(\phi'\right), \overline{e}\left(\phi'\right)\right) \geq \overline{\pi^F} - \widehat{\pi}^F$, and we seek to show that $w\left(\phi''\right)$ exists and satisfies (10). Differentiating (9) with respect to ϕ gives

$$\begin{aligned} & (1 - \phi\gamma)\left(v'\left(w(\phi)\right)w'(\phi) - c'\left(\overline{e}(\phi)\right)\overline{e}'(\phi)\right) + \phi\gamma\left(y'\left(\overline{e}(\phi)\right)\overline{e}'(\phi) - w'(\phi)\right) \\ + \quad & \gamma\left(\widetilde{\pi}^F\left(w(\phi), \overline{e}(\phi)\right) - \widetilde{\pi}^W\left(w(\phi), \overline{e}(\phi)\right)\right) = 0. \end{aligned}$$

Recall that $\overline{e}'(\phi) > 0$ and $\widetilde{\pi}^W\left(w(\phi), \overline{e}(\phi)\right) - \widetilde{\pi}^F\left(w(\phi), \overline{e}(\phi)\right) > 0$. By definition of $\overline{e}(\phi)$, $(1 - \phi\gamma) = \frac{\phi\gamma y'(\overline{e}(\phi))}{c'(\overline{e}(\phi))}$. To see this, note that the "effort curve" $\left\{\left(\widetilde{\pi}^W\left(w', \overline{e}(\phi)\right), \widetilde{\pi}^F\left(w', \overline{e}(\phi)\right)\right)\right\}_{w'}$ must be tangent to the line (9) at $\left(\widetilde{\pi}^W\left(w, \overline{e}(\phi)\right), \widetilde{\pi}^F\left(w, \overline{e}(\phi)\right)\right)$. Substituting and rearranging, we obtain

$$\begin{aligned} w'(\phi) &= \frac{\overline{e}'(\phi)\left((1 - \phi\gamma)\,c'\left(\overline{e}(\phi)\right) - \phi\gamma y'\left(\overline{e}(\phi)\right)\right) + \gamma\left(\widetilde{\pi}^W\left(w(\phi), \overline{e}(\phi)\right) - \widetilde{\pi}^F\left(w(\phi), \overline{e}(\phi)\right)\right)}{(1 - \phi\gamma)\,v'\left(w(\phi)\right) - \phi\gamma} \\ &= \frac{\gamma\left(\widetilde{\pi}^W\left(w(\phi), \overline{e}(\phi)\right) - \widetilde{\pi}^F\left(w(\phi), \overline{e}(\phi)\right)\right)}{\phi\gamma\left(\frac{y'(\overline{e}(\phi))v'(w(\phi))}{c'(\overline{e}(\phi))} - 1\right)}. \end{aligned} \tag{11}$$

By definition of $\widehat{\phi}$, $w\left(\widehat{\phi}\right)$ exists and $\widetilde{\pi}^F\left(w\left(\widehat{\phi}\right), \overline{e}\left(\widehat{\phi}\right)\right) = \widetilde{\pi}^{-}$. Note that the denominator of (11) is zero if and only if (eff) is satisfied, which only occurs on the material-efficiency frontier. Thus Lipschitz conditions are satisfied in some neighbourhood of $\left(\widehat{\phi}, w\left(\widehat{\phi}\right)\right)$. Choosing this initial condition, we claim that the domain of the unique maximal solution to the Cauchy initial value problem (11) contains ϕ'. (That is, we claim that $w'(\phi)$ can be integrated from $\widehat{\phi}$ to ϕ', allowing us to find $w\left(\phi'\right)$ and $w\left(\phi''\right)$.) Assume otherwise, that the domain of the maximal solution is some $\left(\underline{\phi}, \overline{\phi}\right)$ with $\underline{\phi} \geq \phi'$. Let $w\left(\underline{\phi}\right) \equiv \lim_{\phi \to \underline{\phi}} w(\phi)$. Each $(w(\phi), \overline{e}(\phi))$ corresponds to a material payoff

pair which lies on the intersection of the line (9) and the material payoff possibility set. This intersection is a finite line segment. For $\phi \in \left(\underline{\phi}, \widehat{\phi}\right)$, the surplus payoff pair must thus lie within

$$\left\{\left(\widetilde{\pi}^W, \widetilde{\pi}^F\right) : \overline{U} = \widehat{\pi}^W + (1-\phi\gamma)\,\widetilde{\pi}^W\left(w, \overline{e}\left(\phi\right)\right) + \phi\gamma\widetilde{\pi}^F\left(w, \overline{e}\left(\phi\right)\right)\right\}_{\phi\in\left[\underline{\phi},\widehat{\phi}\right]}$$
$$\cap\left\{\left(\widetilde{\pi}^W\left(w,e\right), \widetilde{\pi}^F\left(w,e\right)\right)\right\}_{(w,e)\in\mathbb{R}^2}$$

which, being a union of such finite line segments over $\phi \in \left[\underline{\phi}, \widehat{\phi}\right]$, is a compact set. Since $\left(\widetilde{\pi}^W\left(w\left(\underline{\phi}\right), \overline{e}\left(\underline{\phi}\right)\right), \widetilde{\pi}^F\left(w\left(\underline{\phi}\right), \overline{e}\left(\underline{\phi}\right)\right)\right)$ must also lie in this set, $w\left(\underline{\phi}\right)$ is finite.

Since the maximal domain cannot be extended to $\underline{\phi}$, we must have $\frac{y'\left(\overline{e}\left(\underline{\phi}\right)\right)v'\left(w\left(\underline{\phi}\right)\right)}{c'\left(\overline{e}\left(\underline{\phi}\right)\right)} = 1$, so $\left(w\left(\underline{\phi}\right), \overline{e}\left(\underline{\phi}\right)\right)$ is materially-efficient. By contrast, we claim any transaction $(w, \overline{e}(\phi))$satisfying (9) cannot be materially-efficient. This is because the "effort curve" $\left\{\left(\widetilde{\pi}^W\left(w', \overline{e}\left(\phi\right)\right), \widetilde{\pi}^F\left(w', \overline{e}\left(\phi\right)\right)\right)\right\}_{w'}$ must be tangent to the line (9) at $\left(\widetilde{\pi}^W\left(w, \overline{e}\left(\phi\right)\right), \widetilde{\pi}^F\left(w, \overline{e}\left(\phi\right)\right)\right)$. However, if $(w, \overline{e}(\phi))$ were materially-efficient, the effort curve $\left\{\left(\widetilde{\pi}^W\left(w', \overline{e}\left(\phi\right)\right), \widetilde{\pi}^F\left(w', \overline{e}\left(\phi\right)\right)\right)\right\}_{w'}$ would also be tangent to the material-efficiency frontier at $\left(\widetilde{\pi}^W\left(w, \overline{e}\left(\phi\right)\right), \widetilde{\pi}^F\left(w, \overline{e}\left(\phi\right)\right)\right)$. The line (9) contains $\left(\widetilde{\pi}^-, \widetilde{\pi}^-\right)$, which lies strictly within the material payoff possibility set. (This is a consequence of the assumption that there are potential gains from trade for all ϕ, as we showed in Proposition 5). This implies, conversely to the separating hyperplane theorem, that the line (9) cannot be tangent to the material efficiency frontier at their intersection points. By contradiction, the previous claim holds. In particular, $\left(w\left(\phi\right), \overline{e}\left(\phi\right)\right)$ cannot be materially-efficient for $\phi \in \left(\underline{\phi}, \overline{\phi}\right)$. Since $\left(w\left(\underline{\phi}\right), \overline{e}\left(\underline{\phi}\right)\right)$ is a limit point of $\left\{\left(w\left(\phi\right), \overline{e}\left(\phi\right)\right)\right\}_{\phi\in\left(\underline{\phi},\overline{\phi}\right)}$, it must also satisfy (9). But this contradicts the observation that $\left(w\left(\underline{\phi}\right), \overline{e}\left(\underline{\phi}\right)\right)$ is materially-efficient. Thus $w\left(\phi'\right)$ and $w\left(\phi''\right)$ exist.

One final step: the denominator of (11) must then be of one sign for $\phi \in \left(\underline{\phi}, \overline{\phi}\right)$. In particular, $\frac{y'\left(\overline{e}\left(\widehat{\phi}\right)\right)v'\left(w\left(\widehat{\phi}\right)\right)}{c'\left(\overline{e}\left(\phi\right)\right)} < 1$, so $\frac{y'\left(\overline{e}\left(\phi\right)\right)v'\left(w\left(\phi\right)\right)}{c'\left(\overline{e}\left(\phi\right)\right)} < 1$ throughout. This implies $w'\left(\phi\right) < 0$ for $\phi \in \left(\underline{\phi}, \overline{\phi}\right)$. $w'\left(\phi\right) < 0$ implies $\frac{d\widetilde{\pi}^F\left(w\left(\phi\right),\overline{e}\left(\phi\right)\right)}{d\phi} > 0$ (and incidentally $\frac{d\widetilde{\pi}^W\left(w\left(\phi\right),\overline{e}\left(\phi\right)\right)}{d\phi} < 0$). $\widetilde{\pi}^F\left(w\left(\widehat{\phi}\right), \overline{e}\left(\widehat{\phi}\right)\right) = \widetilde{\pi}^-$ and $\widetilde{\pi}^F\left(w\left(\phi'\right), \overline{e}\left(\phi'\right)\right) \geq \overline{\pi^F} - \widehat{\pi}^F$ then imply $\overline{\pi^F} - \widehat{\pi}^F \leq \widetilde{\pi}^F\left(w\left(\phi''\right), \overline{e}\left(\phi''\right)\right) \leq \widetilde{\pi}^-$. This completes the proof.

12.8 Proof of Proposition 6

Let $\phi \geq \frac{1}{\gamma}$, so $1 - \phi\gamma \leq 0$. Then since

$$U = \begin{cases} \widehat{\pi}^W + \widetilde{\pi}^W - \phi\gamma\left(\widetilde{\pi}^W - \widetilde{\pi}^F\right) & \text{if } \widetilde{\pi}^W \geq \widetilde{\pi}^F \\ \widehat{\pi}^W + \widetilde{\pi}^W - \phi\left(1-\gamma\right)\left(\widetilde{\pi}^F - \widetilde{\pi}^W\right) & \text{if } \widetilde{\pi}^W < \widetilde{\pi}^F \end{cases},$$

$U \geq \overline{U}$ only if $\widetilde{\pi}^W \geq 0$ and $\widetilde{\pi}^F \geq 0$. Also, $\widehat{\pi}^F = \overline{\pi^F}$ implies that the firm receives at least its outside option material payoff if and only if $\widetilde{\pi}^F \geq 0$. Thus a necessary condition for $\pi^F \geq \overline{\pi^F}$ and $U \geq \overline{U}$ is

that $\widetilde{\pi}^W \geq 0$ and $\widetilde{\pi}^F \geq 0$. This, combined with the assumption that there are potential gains from trade, implies that there exists some transaction that gives both the worker and the firm strictly positive surplus payoffs. Since the material payoff possibility set is convex and downward-sloping, the surplus payoff pair (η, η) for $\eta > 0$ small is attained by some transaction $\left(w, e^{\text{fair}}(w)\right)$. Since $1 - \phi\gamma \leq 0$, we have $-(1-\phi\gamma)c'(e') + \phi\gamma y'(e') \geq 0$, so $e^{\text{fair}}(w) = \arg\max_e U(w,e)$. Therefore by offering w, the firm could earn strictly more than its outside option. The equilibrium must be at least as good for the firm as offering w, so exchange occurs.

12.9 Proof of Proposition 7

We define $\widehat{\phi}$ and $\widehat{\widehat{\phi}}$ as in the proof of Proposition 5, but note that now $\widehat{\phi}$ and $\widehat{\widehat{\phi}}$ depend on the realized value of ε. We derive results in terms of $\widehat{\phi}(\varepsilon)$ and $\widehat{\widehat{\phi}}(\varepsilon)$, and at the end of the proof we will define $\overline{\phi}$ and $\overline{\overline{\phi}}$ appropriately.

First, consider the case $\phi \in (\widehat{\phi}(\varepsilon), \widehat{\widehat{\phi}}(\varepsilon)]$. In this case, step 10 of Proposition 5 demonstrates that we have $e^*(\varepsilon) = \overline{e}$, so $e^{*\prime}(\varepsilon) = 0$. $w^*(\varepsilon)$ is then defined by

$$v(w^*(\varepsilon)) - c(\overline{e}) - \widehat{\pi}^W = y(\overline{e}) - w^*(\varepsilon) + \varepsilon - \widehat{\pi}^F,$$

so

$$w^{*\prime}(\varepsilon) = \frac{1}{v'(w^*(\varepsilon)) + 1}.$$

It follows that $0 < w^{*\prime}(\varepsilon) < 1$.

Next, consider the case $\phi \in (\widehat{\widehat{\phi}}, 1]$. Then the equilibrium payoff lies at the intersection of the material-efficiency frontier and the equal-surplus line, which is defined by

$$\begin{aligned} v(w^*(\varepsilon)) - c(e^*(\varepsilon)) - \widehat{\pi}^W &= y(e^*(\varepsilon)) - w^*(\varepsilon) + \varepsilon - \widehat{\pi}^F, \\ \frac{c'(e^*(\varepsilon))}{y'(e^*(\varepsilon))} &= v'(w^*(\varepsilon)). \end{aligned}$$

Differentiating both equalities with respect to ε yields

$$w^{*\prime}(\varepsilon) = \frac{\left(y'(e^*(\varepsilon)) + c'(e^*(\varepsilon))\right)e^{*\prime}(\varepsilon) + 1}{v'(w^*(\varepsilon)) + 1}, \tag{12}$$

$$\frac{w^{*\prime}(\varepsilon)}{e^{*\prime}(\varepsilon)} = \frac{c''(e^*(\varepsilon)) - v'(w^*(\varepsilon))\, y''(e^*(\varepsilon))}{v''(w^*(\varepsilon))\, y'(e^*(\varepsilon))} < 0. \tag{13}$$

(13) implies that $w^{*\prime}(\varepsilon)$ and $e^{*\prime}(\varepsilon)$ must have opposite sign. But then (12) can only be satisfied if $w^{*\prime}(\varepsilon) > 0$, $e^{*\prime}(\varepsilon) < 0$. Since the numerator of (12) is smaller than 1 and the denominator is larger than 1, it also follows that $w^{*\prime}(\varepsilon) < 1$.

To complete the proof, let $\overline{\phi} \equiv \max_{\varepsilon \in [\underline{\varepsilon}, \overline{\varepsilon}]} \widehat{\phi}(\varepsilon)$ and $\overline{\overline{\phi}} \equiv \max_{\varepsilon \in [\underline{\varepsilon}, \overline{\varepsilon}]} \widehat{\widehat{\phi}}(\varepsilon)$. The result follows immediately.

12.10 Proof of Proposition 8

The worker's equilibrium effort choice uniquely satisfies

$$-c'(e^*) + ay'(e^*) = 0$$

and is independent of the firm's wage offer w and the shock to profit ε. The firm's problem is to maximize $\pi^F(w, e^*)$, which is equivalent to minimizing w, subject to the constraint

$$U(w, e^*) = \pi^W(w, e^*) + a\left(\pi^F(w, e^*) + \varepsilon\right) \geq \overline{U}.$$

Clearly the constraint is binding at any solution to the firm's problem, so the worker earns his outside option level of utility. Note that since $U(w, e^*)$ is concave in w, there are at most two distinct values of w at which the constraint binds. This implies that the equilibrium transaction $(w^*(\varepsilon), e^*)$ is unique. Implicit differentiation of the binding constraint with respect to ε yields, after rearranging,

$$w^{*\prime}(\varepsilon) = -\frac{a}{v'(w^*(\varepsilon)) + a} < 0.$$

Finally, we claim that there is exactly one value of ε for which $(w^*(\varepsilon), e^*)$ is Pareto efficient. To show this, first note that since any allocation (w, e) with

$$\left(\pi^W(w, e), \pi^F(w, e)\right) \geq \left(\pi^W(w^*(\varepsilon), e^*), \pi^F(w^*(\varepsilon), e^*)\right)$$

must also have

$$\left(U(w, e), \pi^F(w, e)\right) \geq \left(U(w^*(\varepsilon), e^*), \pi^F(w^*(\varepsilon), e^*)\right),$$

any Pareto efficient transaction must also be materially-efficient. But given e^*, the material-efficiency condition $\frac{v'(w)}{c'(e^*)} = \frac{1}{y'(e^*)}$ has exactly one solution in w. Since $w^{*\prime}(\varepsilon) < 0$, there is at most one value of ε for which the equilibrium transaction is materially-efficient. Generically, the equilibrium transaction is not Pareto efficient.

12.11 Proof of Proposition 9

We present a sketch of the proof, in 6 steps:

1. *If* $f\left(\widetilde{\pi}^W, \widetilde{\pi}^F\right) = \min\left\{\widetilde{\pi}^W, \widetilde{\pi}^F\right\}$, *then there exists* $0 < \widehat{\widehat{\varphi}} < 1$ *such that for* $\varphi \geq \widehat{\widehat{\varphi}}$, *the equilibrium payoff lies at the intersection of the equal-surplus line and the material-efficiency frontier.* The worker's utility function is

$$\begin{aligned} U\left(\pi^W, \pi^F\right) &= (1-\varphi)\,\pi^W + \varphi \min\left\{\widetilde{\pi}^W, \widetilde{\pi}^F\right\} \\ &= (1-\varphi)\,\widehat{\pi}^W + \widetilde{\pi}^W - \varphi\left(\max\left\{0, \widetilde{\pi}^F - \widetilde{\pi}^W\right\} + \max\left\{0, \widetilde{\pi}^W - \widetilde{\pi}^F\right\}\right) \end{aligned}$$

which is exactly equal (up to a constant) to (2) with $\beta = \frac{1}{2}$ and $\phi = 2\varphi$. Thus, when $\varphi > \frac{c'(e^+)}{c'(e^+)+y'(e^+)} = \widehat{\widehat{\varphi}}$, we may argue as in Proposition 4 that the equilibrium payoff lies at the intersection of the equal-surplus line and the material-efficiency frontier. Note that $\widehat{\widehat{\varphi}} = \widehat{\widehat{\phi}}$ for $\beta = \frac{1}{2}$.

2. *An equilibrium exists.* With utility function (2), it is straightforward to show that the worker's utility maximization problem has a unique solution $e(w)$. Since $\frac{\partial}{\partial w} e(w)$ is continuous in w, $e(w)$ is continuous in w. The firm maximizes $\pi^F(w, e(w))$ subject to the worker's outside option constraint

$$U\left(\pi^W(w, e(w)), \pi^F(w, e(w))\right) \geq \overline{U}.$$

Let $\left(\pi_0^W, \underline{\pi}^F\right)$ be some material payoff pair satisfying the worker's utility-maximization and outside-option constraints. Let $\Gamma = \left\{\left(\pi^W(w, e), \pi^F(w, e)\right)\right\}_{(w,e)}$ be the material payoff possibility set. We can show that

$$\left\{\left(\pi^W, \pi^F\right) : U\left(\pi^W, \pi^F\right) \geq \overline{U}\right\} \cap \left\{\left(\pi^W, \pi^F\right) : \pi^F \geq \underline{\pi}^F, \left(\pi^W, \pi^F\right) \in \Gamma\right\}$$

is bounded. This is the set of material payoff pairs satisfying the worker's outside option constraint such that profits are bounded below by $\underline{\pi}^F$.

With $P_{\pi^W}(\cdot)$ denoting the projection operator onto the π^W-axis, let

$$\underline{\pi}^W = \inf P_{\pi^W}\left(\left\{\left(\pi^W, \pi^F\right) : U\left(\pi^W, \pi^F\right) \geq \overline{U}\right\} \cap \left\{\left(\pi^W, \pi^F\right) : \pi^F \geq \underline{\pi}^F, \left(\pi^W, \pi^F\right) \in \Gamma\right\}\right),$$

so that no $\pi^W < \underline{\pi}^W$ is feasible for the firm when $\pi^F \geq \underline{\pi}^F$. Noting that the material payoff possibility set is convex, we can show that the set $\left\{(w, e) : \pi^F(w, e) \geq \underline{\pi}^F, \pi^W(w, e) \geq \underline{\pi}^W\right\}$ must be bounded. Note that Φ is bounded and non-empty, where

$$\Phi \equiv \left\{(w, e) : \pi^F(w, e) \geq \underline{\pi}^F, \pi^W(w, e) \geq \underline{\pi}^W, U(w, e) \geq \overline{U}\right\},$$

and let $\overline{P}_w(\Phi)$ be the closure of the projection of Φ onto the w-axis. Then without loss of generality, the firm maximizes a continuous function $\pi(w, e(w))$ over a compact set $\overline{P}_w(\Phi)$, for which a solution exists.

3. *Conditional on the firm's wage offer,* $\lim_{\kappa\to\infty}\left(\widetilde{\pi}^W(w, e(w;\kappa)) - \widetilde{\pi}^F(w, e(w;\kappa))\right) = 0$ *whenever* $\varphi \geq \frac{c'\left(e^{\text{fair}}(w)\right)}{c'(e^{\text{fair}}(w))+y'(e^{\text{fair}}(w))} \equiv \widehat{\widehat{\varphi}}$. As before, we define $e^{\text{fair}}(w)$ such that $\left(w, e^{\text{fair}}(w)\right)$ lies on the equal-surplus line. Fix w. Note that the worker chooses $\max_e U(w, e)$, which has first-order condition

$$\exp\left\{\kappa\left(\widetilde{\pi}^F - \widetilde{\pi}^W\right)\right\} = \varphi\frac{y'(e(w;\kappa))}{c'(e(w;\kappa))} - (1-\varphi).$$

Since w is fixed, we consider $e(w;\kappa)$ as a function of κ. As $\kappa \to \infty$, $e(w;\kappa)$ is bounded above. To see this, assume otherwise. We would have

$$\begin{aligned} \lim_{\kappa\to\infty} \exp\left\{\kappa\left(\widetilde{\pi}^F(w, e(w;\kappa)) - \widetilde{\pi}^W(w, e(w;\kappa))\right)\right\} &= \infty \text{ and} \\ \lim_{\kappa\to\infty} \frac{y'(e(w;\kappa))}{c'(e(w;\kappa))} &= 0, \end{aligned}$$

but this contradicts the worker's first-order condition. Similarly, as $\kappa \to \infty$, $e(w;\kappa)$ is bounded below. Hence $e(w;\kappa)$ is bounded. This implies

$$\lim_{\kappa\to\infty}\left(\widetilde{\pi}^F(w, e(w;\kappa)) - \widetilde{\pi}^W(w, e(w;\kappa))\right) \leq 0;$$

otherwise we would have

$$\lim_{\kappa\to\infty} \frac{y'(e(w;\kappa))}{c'(e(w;\kappa))} < \infty = \lim_{\kappa\to\infty} \exp\left\{\kappa\left(\widetilde{\pi}^F(w, e(w;\kappa)) - \widetilde{\pi}^W(w, e(w;\kappa))\right)\right\}.$$

Next we claim $\lim_{\kappa\to\infty} \widetilde{\pi}^F(w, e(w;\kappa)) - \widetilde{\pi}^W(w, e(w;\kappa)) \geq 0$. Assume otherwise; then $\lim_{\kappa\to\infty} e(w;\kappa) < e^{\text{fair}}(w)$ and

$$\begin{aligned} & \lim_{\kappa\to\infty} \varphi \frac{y'(e(w;\kappa))}{c'(e(w;\kappa))} - (1-\varphi) \\ = & \lim_{\kappa\to\infty} \exp\left\{\kappa\left(\widetilde{\pi}^F(w, e(w;\kappa)) - \widetilde{\pi}^W(w, e(w;\kappa))\right)\right\} \\ = & 0. \end{aligned}$$

Since

$$\begin{aligned} & \varphi \frac{y'(e(w;\kappa))}{c'(e(w;\kappa))} - (1-\varphi) \\ = & \varphi\left(\frac{y'(e(w;\kappa)) + c'(e(w;\kappa))}{c'(e(w;\kappa))}\right) - 1 \\ \geq & \frac{c'\left(e^{\text{fair}}(w)\right)}{c'(e^{\text{fair}}(w)) + y'(e^{\text{fair}}(w))}\left(\frac{y'(e(w;\kappa)) + c'(e(w;\kappa))}{c'(e(w;\kappa))}\right) - 1, \end{aligned}$$

we must have

$$\lim_{\kappa\to\infty} \frac{y'(e(w;\kappa)) + c'(e(w;\kappa))}{c'(e(w;\kappa))} < \frac{y'\left(e^{\text{fair}}(w)\right) + c'\left(e^{\text{fair}}(w)\right)}{c'(e^{\text{fair}}(w))},$$

so $\lim_{\kappa\to\infty} e(w;\kappa) \geq e^{\text{fair}}(w)$, a contradiction. Hence

$$\lim_{\kappa\to\infty}\left(\widetilde{\pi}^F(w, e(w;\kappa)) - \widetilde{\pi}^W(w, e(w;\kappa))\right) = 0.$$

4. We may show in a similar fashion that: *conditional on the firm's wage offer,*

$$\lim_{\kappa\to\infty}\left(\widetilde{\pi}^W(w, e(w;\kappa)) - \widetilde{\pi}^F(w, e(w;\kappa))\right) > 0$$

whenever $\varphi < \widehat{\widehat{\varphi}}$.

5. *Let U_κ denote the worker's utility function for parameter κ, and let $\varphi > \widehat{\widehat{\varphi}}$. As $\kappa \to \infty$, $U \to U_\kappa(w^+, e^+)$ at equilibrium, and any equilibrium transaction converges to (w^+, e^+).* Note that $U_\infty\left(\pi^W, \pi^F\right) = (1-\varphi)\,\pi^W + \varphi \min\left(\widetilde{\pi}^W, \widetilde{\pi}^F\right)$. From steps 3 and 4,

$$\lim_{\kappa\to\infty}\left(\widetilde{\pi}^W\left(w, e_\kappa(w)\right) - \widetilde{\pi}^F\left(w, e_\kappa(w)\right)\right) \geq 0,$$

with equality if and only if $\varphi \geq \widehat{\widehat{\varphi}}$. Since $\varphi > \widehat{\widehat{\varphi}}$, $\lim_{\kappa\to\infty} e_\kappa(w^+) = e^+$.

We now assume towards a contradiction that for some sequence $\kappa_j \to \infty$, a corresponding sequence of equilibrium transactions $(w_j, e_j(w_j)) \to (w_\infty, e_\infty(w_\infty)) \neq (w^+, e^+)$. At the limit, firm's profit is $\pi^F(w_\infty, e_\infty)$. Note that for given j, $\left(\pi^W(w_j, e_j(w_j)), \pi^F(w_j, e_j(w_j))\right)$ must lie in

$$\left\{\left(\pi^W, \pi^F\right) : U_{\kappa_j}\left(\pi^W, \pi^F\right) \geq \overline{U}\right\} \cap \left\{\left(\pi^W, \pi^F\right) : \pi^F \geq \underline{\pi}^F, \left(\pi^W, \pi^F\right) \in \Gamma\right\},$$

so $\left(\pi^W(w_\infty, e_\infty), \pi^F(w_\infty, e_\infty)\right)$ must lie in

$$\left\{\left(\pi^W, \pi^F\right) : U_\infty\left(\pi^W, \pi^F\right) \geq \overline{U}\right\} \cap \left\{\left(\pi^W, \pi^F\right) : \pi^F \geq \underline{\pi}^F, \left(\pi^W, \pi^F\right) \in \Gamma\right\},$$

which is compact. Since the material payoff possibility set is closed, (w_∞, e_∞) is finite. The firm's profit converges to $\pi^F(w_\infty, e_\infty)$; from step 1, $\pi^W(w_\infty, e_\infty) = \pi^F(w_\infty, e_\infty)$. The maximum attainable firm payoff that lies on the equal-surplus line corresponds to $\pi^F(w^+, e^+)$. Thus, by assumption $\pi^F(w_\infty, e_\infty) < \pi^F(w^+, e^+)$; but $\pi^F(w^+, e_j(w^+)) \to_{j\to\infty} \pi^F(w^+, e^+)$, so $\lim_{j\to\infty} \pi^F(w^+, e_j(w^+)) - \pi^F(w_j, e_j(w_j)) > 0$, which contradicts our assumption that $(w_j, e_j(w_j))$ is an equilibrium transaction.

6. *The limit transaction is Pareto efficient.* The limit transaction (w^+, e^+) corresponds to the intersection between the equal-surplus line and the material-efficiency frontier. Keeping in mind our observation from step 3 that the worker's utility function at the limit is equivalent to (2), Proposition 1 implies that in the limit $\kappa \to \infty$, (w^+, e^+) is Pareto efficient.

References

[1] John M. Abowd, Francis Kramarz, and David N. Margolis. High wage workers and high wage firms. *Econometrica*, 67(2):251–333, March 1999.

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

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

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

[5] William Austin and Elaine Walster. Reactions to confirmations and disconfirmations of expectancies of equity and inequity. *Journal of Personality and Social Psychology*, 30(2):208–216, 1974.

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

[7] George Baker, Michael Gibbs, and Bengt Holmstrom. The wage policy of a firm. *Quarterly Journal of Economics*, 109(4):921–955, November 1994.

[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] Paul Beaudry and John DiNardo. The effect of implicit contracts on the movement of wages over the business cycle: Evidence from micro data. *Journal of Political Economy*, 99(4):665–688, August 1991.

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

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

[12] David G. Blanchflower, Andrew J. Oswald, and Peter Sanfey. Wages, profits, and rent-sharing. *Quarterly Journal of Economics*, 111(1):227–251, February 1996.

[13] Martin Brown, Armin Falk, and Ernst Fehr. Relational contracts and the nature of market interactions. *Econometrica*, 72(3):747–780, May 2004.

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

[15] Carl M. III Campbell and Kunal S. Kamlani. The reasons for wage rigidity: Evidence from a survey of firms. *Quarterly Journal of Economics*, 112(3):759–789, August 1997.

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

[17] Robert B. Cialdini. *Influence: The Psychology of Persuasion*. William Morrow and Company, Inc, 1984.

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

[19] Yochi Cohen-Charash and Paul E. Spector. The role of justice in organizations: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 86(2):278–321, November 2001.

[20] D.C. Dacy and H. Kunreuther. *The economics of natural disasters*. Free Press, New York, 1969.

[21] William T. Dickens and Lawrence F. Katz. Inter-industry wage differences and theories of wage determination. *NBER Working Paper No. 2271*, June 1987.

[22] Ronald G. Ehrenberg and George T. Milkovich. Compensation and firm performance. *NBER Working Paper No. 2145*, February 1987.

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

[24] Florian Englmaier. A survey on moral hazard, contracts and social preferences. In Bina Agarwal and Alessandro Vercelli, editors, *Psychology, Rationality and Economic Behavior: Challenging Standard Assumptions*. 2004.

[25] Ernst Fehr and Armin Falk. Wage rigidity in a competitive incomplete contract market. *Journal of Political Economy*, 107:106–134, 1999.

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

[28] Ernst Fehr, Alexander Klein, and Klaus M. Schmidt. Fairness, incentives and contractual incompleteness. *CEPR Discussion Paper No. 2790*, May 2001.

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

[30] Ernst Fehr and Klaus M. Schmidt. Fairness and incentives in a multi-task principal-agent model. *Scand. J. of Economics*, 106(3):453–474, 2004.

[31] Ray Fisman, Shachar Kariv, and Daniel Markovits. Distinguishing social preferences from preferences for altruism. *UC Berkeley manuscript*, 2005.

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

[33] Meredith C. Frey and Douglas K. Detterman. Scholastic assessment or g? the relationship between the Scholastic Assessment Test and general cognitive ability. *Psychological Science*, 15(6):373–378, 2004.

[34] David Genesove. The nominal rigidity of apartment rents. *NBER Working Paper No. 7137*, May 1999.

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

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

[37] 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.

[38] Douglas Kruse, Richard Freeman, Joseph Blasi, Robert Buchele, Adrin Scharf, Loren Rodgers, and Chris Mackin. Motivating employee-owners in esop firms: Human resource policies and company performance. *NBER Working Paper No. 10177*, December 2003.

[39] Edward E. III Lawler. Pay for performance: A motivational analysis. In Haig R. Nalbantian, editor, *Incentives, Cooperation, and Risk Sharing*, pages 69–86. Rowman Littlefield, Totowa, NJ, 1987.

[40] John List. The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions. *Journal of Political Economy*, forthcoming, 2006.

[41] George F. Loewenstein, Leigh Thompson, and Max H. Bazerman. Social utility and decision-making in interpersonal contexts. *Journal of Personality and Social Psychology*, 57(3):426–441, 1989.

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

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

[44] A.L. Olmstead and P. Rhode. Rationing without government: The west coast gas famine of 1920. *American Economic Review*, 75:1044–55, 1985.

[45] Canice Prendergast. The provision of incentives in firms. *Journal of Economic Literature*, 37(1):7–63, March 1999.

[46] 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.

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

[48] Julio J. Rotemberg. Fair pricing. March 2005. Harvard Business School mimeo.

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

[50] Adam Smith. *An Inquiry Into the Nature and Causes of the Wealth of Nations*. 1776. http://www.econlib.org/LIBRARY/Smith/smWN.html.

[51] Martin L. Weitzman and Douglas Kruse. Profit sharing and productivity. In Alan Blinder, editor, *Paying for Productivity*. Brookings Institution, Washington, D.C., 1990.

[52] M. E. Yaari and M. Bar-Hillel. On dividing justly. *Social Choice and Welfare*, 1:1–24, 1984.

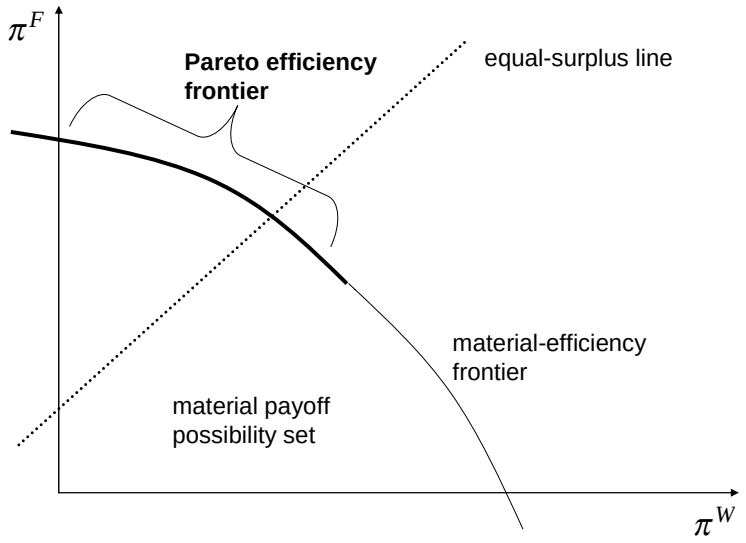

Figure 1. Some key concepts graphed in the space of material payoffs. The material payoff possibility set is convex, and the material-efficiency frontier is downward-sloping. The equal-surplus line is the set of material payoffs that equate the firm's and worker's surplus payoffs. The Pareto efficiency frontier is a subset of the material-efficiency frontier that includes at least all the material-efficient payoffs above the equal-surplus line.

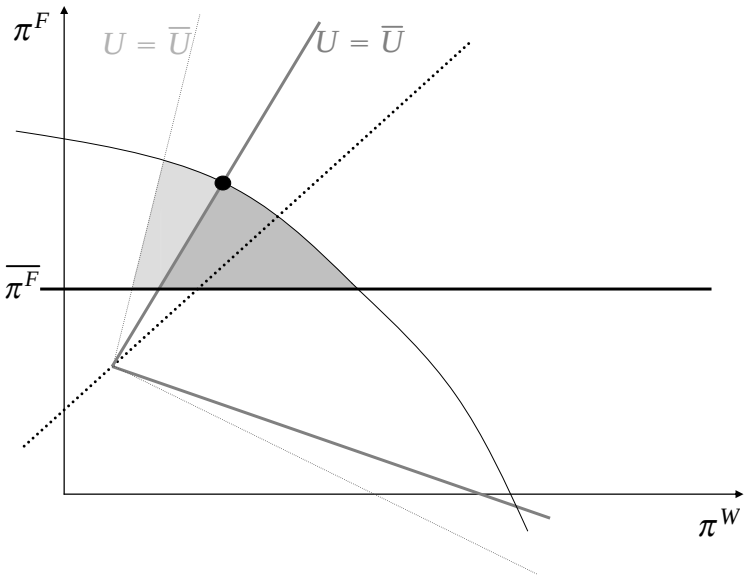

Figure 2. Potential gains from trade. The dark, tilted-V shape is the worker's outside option indifference curve. The horizontal line is the firm's outside option indifference curve. The darkly-shaded region is the set of attainable material payoff pairs that make both the worker and firm better off than their outside options. When the firm can offer the worker an enforceable contract, the equilibrium occurs at the black point, where the material-efficiency frontier intersects the worker's outside option indifference curve. The more lightly-drawn indifference curve corresponds to a lower weight on fairness in the worker's utility function. The lightly-shaded region is the set of additional attainable material payoff pairs that the worker and firm would now accept.

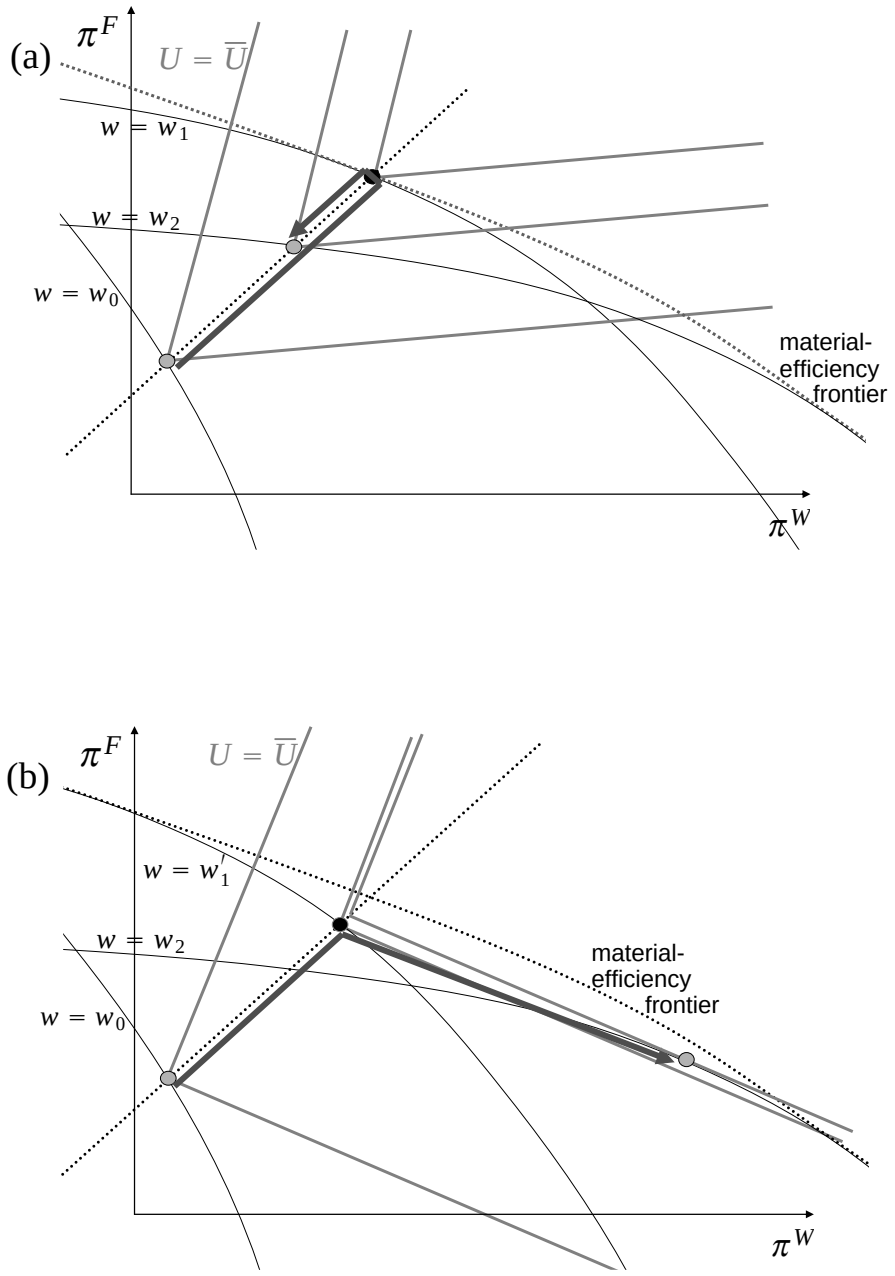

Figure 3. Equilibrium without enforceable contracts. The downward-sloping "wage curves" are, for given wage offers $w_0 < w_1' < w_1 < w_2$, the possible material payoff pairs for varying effort levels. The downward-sloping dotted line is the material-efficiency frontier. The lightly-shaded points show 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. Panel (a): The worker puts high weight on fairness. Panel (b): The worker puts lower weight on fairness.

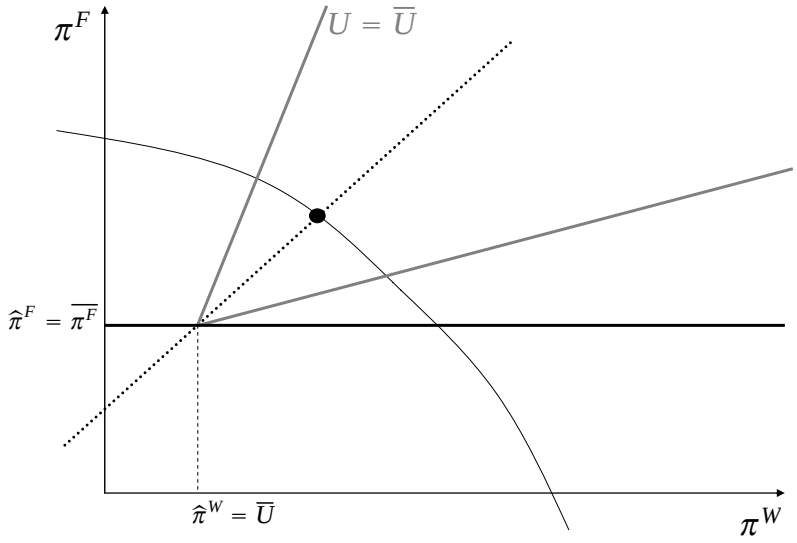

Figure 4. When the market terms of trade pin down the outside options and the reference transaction and when the worker cares sufficiently about fairness, then potential gains from trade imply that exchange will occur. The black point is the equilibrium.

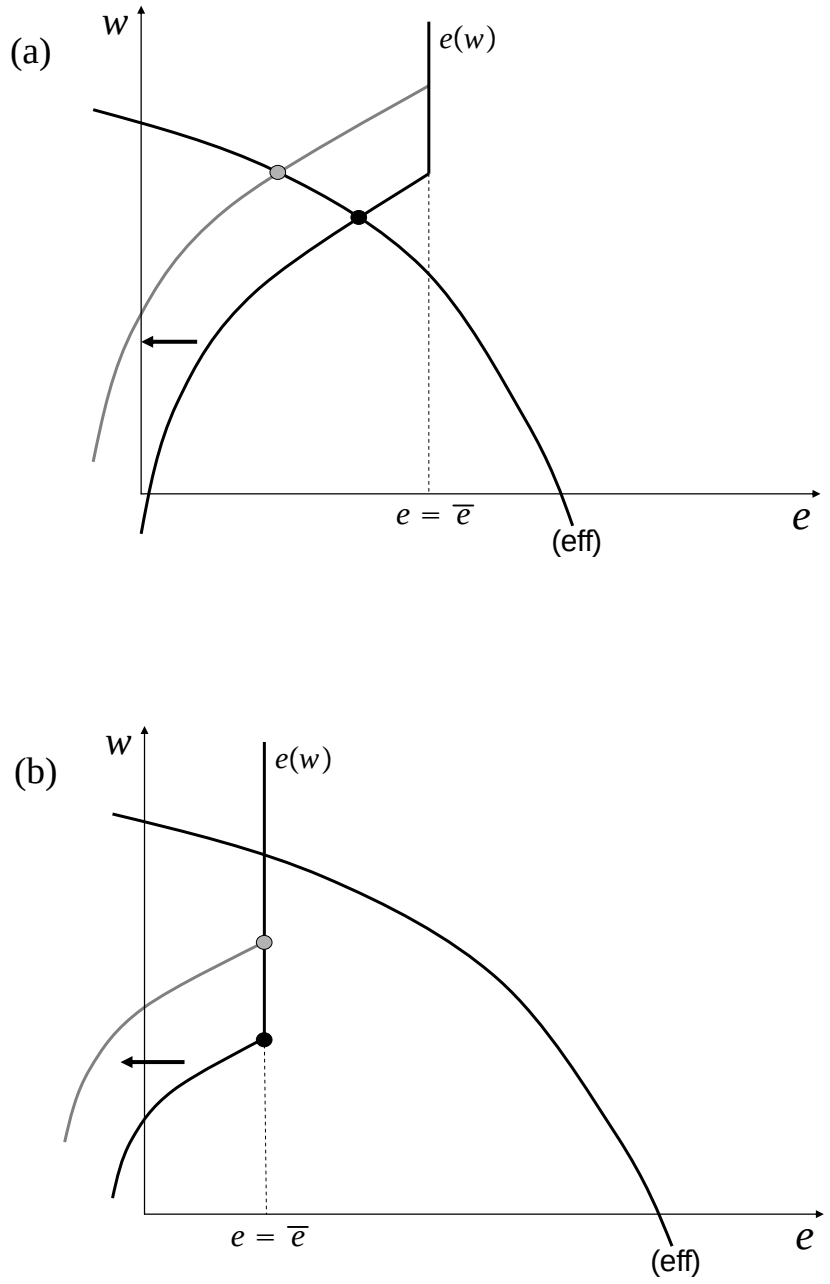

Figure 5. Profit-sharing. The material-efficiency condition (eff) is a downward-sloping relation in wage-effort space. The worker's effort choice $e(w)$ is increasing in the wage, up to some maximum level of effort. The black point is the equilibrium. A greater shock to profit reduces the "fair" level of effort for a given wage. The lightly-shaded point is the new equilibrium. Panel (a): When the worker cares sufficiently about fairness, the equilibrium occurs at the intersection of the two curves. Panel (b): When the worker cares less about fairness, the firm offers the lowest wage consistent with the effort maximum.