DISCUSSION PAPER SERIES

No. 7667

ROTTEN KIDS WITH BAD INTENTIONS

Nick Netzer and Armin Schmutzler

INDUSTRIAL ORGANIZATION and
LABOUR ECONOMICS

Centre for Economic Policy Research

www.cepr.org

Available online at: www.cepr.org/pubs/dps/DP7667.asp

ISSN 0265-8003

ROTTEN KIDS WITH BAD INTENTIONS

Nick Netzer, Universität Zurich
Armin Schmutzler, Universität Zurich, ENCORE and CEPR

Discussion Paper No. 7667
January 2010

Centre for Economic Policy Research
53–56 Gt Sutton St, London EC1V 0DG, UK
Tel: (44 20) 7183 8801, Fax: (44 20) 7183 8820
Email: cepr@cepr.org, Website: www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programme in **INDUSTRIAL ORGANIZATION and LABOUR ECONOMICS**. Any opinions expressed here are those of the author(s) and not those of the Centre for Economic Policy Research. Research disseminated by CEPR may include views on policy, but the Centre itself takes no institutional policy positions.

The Centre for Economic Policy Research was established in 1983 as an educational charity, to promote independent analysis and public discussion of open economies and the relations among them. It is pluralist and non-partisan, bringing economic research to bear on the analysis of medium- and long-run policy questions.

These Discussion Papers often represent preliminary or incomplete work, circulated to encourage discussion and comment. Citation and use of such a paper should take account of its provisional character.

CEPR Discussion Paper No. 7667

January 2010

ABSTRACT

Rotten Kids with Bad Intentions

We examine a "Rotten Kid" model (Becker 1974) where a player with social preferences interacts with an egoistic player. We assume that social preferences are intention-based rather than outcome-based. In a very general multi-stage setting we show that any equilibrium must involve mutually unkind behavior of both players, endogenously generating negative emotions rather than positive altruism. In a large class of two-stage games that includes principal-agent and gift-giving games, this prevents equilibrium from being materially Pareto efficient. Compared to the subgame-perfect equilibrium without social preferences, efficiency is still generally increased. On the other hand, the materialistic player has lower whereas the reciprocal player has higher material payoffs, so that reciprocity does not increase equity: For sufficiently strong reciprocity concerns, the materialistic player ends up with a negligible share of the gains from trade.

JEL Classification: C72, D03, D86 and J01
Keywords: gift giving, moral hazard, psychological games and reciprocity

Nick Netzer
Socioeconomic Institute
University of Zurich
Blümlisalpstr. 10
CH-8006 Zurich
SWITZERLAND

Email: nick.netzer@soi.uzh.ch

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=169705

Armin Schmutzler
Socioeconomic Institute
University of Zurich
Blümlisalpstr. 10
CH-8006 Zurich
SWITZERLAND

Email: arminsch@soi.unizh.ch

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=157141

Submitted 25 January 2010

We are grateful to Daniel Benjamin, Paul Heidhues, Georg Kirchsteiger, Ulrike Malmendier, Matthew Rabin, Stefano della Vigna, and seminar participants at the Universities of Frankfurt and Zurich, UC Berkeley and UC Irvine, and the Meeting of German Economists Abroad in Heidelberg for valuable suggestions. All errors are our own.

1 Introduction

One of the earliest and best-known results in the theory of social preferences is Gary Becker's Rotten Kid Theorem (Becker 1974, 1981). Becker considers a framework where an egoistic player (the "rotten kid") can take an action that increases joint ("family") income, but reduces his own income, before an altruistic player (the "parent") makes a transfer to the kid. According to the Rotten Kid Theorem, such one-sided altruism can induce efficient behavior of the kid (provided its consumption is a normal good from the parent's point of view.)[1] While the original specification based on altruism is well-suited to describe interactions within the family, the belief that social preferences can have beneficial effects by facilitating interactions is much more general. For example, Fehr, Gächter, and Kirchsteiger (1997) provide experimental evidence showing that social preferences can be exploited to motivate workers and to enforce contracts that would otherwise be infeasible. In other situations, social preferences tend to work in favor of fair (equitable) allocations, sometimes at the cost of lower economic efficiency (Fehr and Schmidt 1999; Bolton and Ockenfels 2000). In a more recent paper, Benjamin (2008) presents the most benign view according to which both efficiency *and* equity are fostered by the existence of one-sided social preferences. His model is cast as a gift-exchange game where a profit-maximizing firm pays a wage to a worker who then chooses efforts. Under plausible (and sufficiently strong) social preferences such as inequity-aversion (Fehr and Schmidt 1999), both efficiency and equity will emerge.

What the above discussed models have in common is the assumption of *outcome-based* social preferences. Experimental evidence, however, suggests that social preferences exhibit a strong *intention-based* component (Charness and Rabin 2002; Falk, Fehr and Fischbacher 2003a,b; Falk and Fischbacher 2006). In this paper, we examine the impact of intention-based reciprocity concepts (Rabin 1993; Dufwenberg and Kirchsteiger 2004) in the rotten kid framework.

This setting, where one materialistic and one reciprocal player interact, appears natural, for example, when the relation between profit-maximizing firms and their employees is modeled. Before examining such specific applications, however, we first consider a general game with finitely many stages, with observed actions and without nature. As in Rabin (1993) or Dufwenberg and Kirchsteiger (2004), the reciprocal player benefits from rewarding kind behavior of the other player with kind behavior and from punishing unkind behavior with unkind behavior, where kindness is defined relative to a reference point. A (sequential) intentions equilibrium (IE) consists of strategies and correct first- and second-order beliefs such that, in each subgame, the materialistic player maximizes his payoffs given his first-order

[1] See also Bergstrom (1989) for the role of transferability of utility.

beliefs about the other player's strategy, whereas the reciprocal player maximizes his utility given his first-order beliefs about the material player's strategy and his second-order beliefs about the material player's first-order beliefs. We prove a fundamental result for this general set-up. In every IE of the game, under very weak restrictions on the kindness concept and, in particular, on the reference point, behavior of both agents must be weakly unkind, and with mild additional restrictions, it must be strictly unkind. Positive kindness is impossible in the rotten kid setup with one-sided intention-based social preferences, because an egoistic player who maximizes own material payoffs does never have good intentions.

To explore the equilibrium properties further, we then concentrate on a two-stage "action-reaction" setting as in Becker (1974) and Benjamin (2008), where the materialistic player moves first, followed by the reciprocal player. The first player's material payoff is decreasing in own actions ("wages") and increasing in those of the second player ("efforts"). The reciprocal player's material payoff is increasing in the first player's action. This setting encompasses, for instance, standard principal-agent and gift-giving games. As it turns out, mutual unkindness generally prevents equilibrium from being materially Pareto efficient in this framework.

We then compare the material equilibrium outcomes in IE with those in the standard subgame-perfect equilibrium (SPE), where both players are materialistic. The first striking result is that the materialistic player always suffers from the reciprocity of the other player: His payoffs are lower in any IE than in any SPE. This contrasts sharply with results obtained for altruism or inequity aversion, where the egoist generally benefits from the other player's social preferences.

In a large class of games, the wages paid in IE tend to be larger than in SPE, as could also be expected with outcome-based social preferences. However, they are not paid to trigger the return of a benevolent gift, but to prevent the second player from punishing the first player's bad intentions too strongly. Hence the payoff implications are fundamentally different. As a consequence of the higher wages, the reciprocal player obtains higher material payoffs. Note the irony of this result: Even though the reciprocal player is relatively less concerned with material payoffs than without reciprocity, he ends up with higher material payoffs.

The positive effects of reciprocity on wages are very robust: Essentially, as wages increase, (i) the materialistic player becomes kinder and the reciprocal player feels less need to punish him, and (ii) punishment becomes less effective, because when the materialistic player pays higher wages, he loses less material payoffs from effort reductions. The effects of reciprocity on equilibrium efforts are more subtle. In the most general setting, there are two countervailing effects: On the one hand, effort reacts more strongly to wage increases than in the SPE, giving the materialistic player incentives to induce higher efforts. On the other hand, as the materialistic player has to pay higher wages to induce any given effort level, he will benefit less from an increase in efforts. However, in several quite familiar examples of the general action-reaction framework, we obtain a positive effect of reciprocity on efforts.

These findings have interesting implications for efficiency and distribution. In the examples that we examine, greater effort is desirable from an efficiency perspective. Thus, reciprocity is indeed efficiency-enhancing compared to the subgame-perfect equilibrium, but is not able to achieve Pareto efficiency. The distributional consequences are, however, quite surprising. The player who benefits is the reciprocator, while the egoist loses. In the moral hazard application, for example, this implies that a firm should have little interest in hiring workers with intention-based social preferences. Also, reciprocity does not lead to greater equity. First, obviously, if the second player has higher material payoffs in the SPE than the first player, then reciprocity reinforces this inequality. Second, if the reciprocal player has lower material payoffs in the SPE than the materialistic player, reciprocity only leads to more equity as long as it is not too pronounced. We are able to show that, when the agent's materialistic concerns become negligible, the agent eventually obtains almost the entire gains from trade, resulting in the strongest possible inequality.

An important technical aspect of the paper concerns the multiplicity of equilibria. Even though this is arguably a smaller problem when there is only one reciprocal player, multiple equilibria are still pervasive when reciprocity is strong enough. It is therefore remarkable that several of our results on the comparison between SPE and IE hold for all of these equilibria. The paper is organized as follows. Section 2 contains a discussion of related literature. Section 3 introduces the general framework. In Section 4, we provide our general "bad intentions" result. Section 5 deals with the action-reaction model in detail and Section 6 illustrates the ideas with concrete examples. Section 7 contains a critical discussion of our results, and Section 8 concludes. Some proofs are relegated to the Appendix.

2 Related Literature

Among the papers discussed in the previous section, Benjamin (2008) provides the most general analysis for *outcome*-based social preferences against which our results can be compared. The material payoff functions reflect a gift-exchange setup, which is a special case of our action-reaction model. In an application with Fehr-Schmidt preferences, the equilibrium wage offer is Pareto efficient if and only if the aversion against treating the principal unfairly is sufficiently strong, because the worker will exert more efforts as the wage increases. Also, inequity aversion necessarily works towards equity. This distributional effect of social preferences differs from our framework with reciprocity. In our framework, even though the player with social preferences benefits (in a material sense) from having such preferences and the player with material preferences loses, redistribution will be excessive in the sense that, for strong reciprocity, the material player has much lower payoffs than the reciprocal player.

Dur and Glazer (2008) study optimal incentives when a worker envies the employer. While the employer is risk-neutral and has purely materialistic preferences (as in our model), the agent is risk-averse and envious, with utility depending negatively on the difference between the principal's profit and wages. The agent's preferences are therefore again outcome-based. The main results are as follows: As in our model, profits decline as social preferences become more important. Also, the bonus payment is positively affected by envy. The effects on wages (base salaries) and efforts are ambiguous.[2] Hence our results for intention-based preferences are closer to those for models of envy rather than for models of altruism or inequity-aversion. Interestingly, this occurs endogenously in our framework of reciprocity that a priori allows for both positive (altruism-type) or negative (envy-type) emotions.

Few papers have attempted to model *intentions* in contractual environments. Englmaier and Leider (2008) is closely related to our contribution in that the authors consider the interaction between a materialistic principal and a reciprocal agent. Their set-up differs from ours in several respects. To understand the most important difference, in the central result of our analysis (that the equilibrium is characterized by mutually unkind behavior) we essentially allow for arbitrary reference points except those that are at the extremes of the Pareto frontier. Contrary to our approach, Englmaier and Leider (2008) specify the reference point as one of the extremes. They consider a payoff for the agent as equitable if he obtains his outside option. This leads to very different implications. Most importantly, the principal benefits from reciprocity.[3] By construction, he can never be unkind, and as soon as he offers more to the agent than the outside option, the agent responds by behaving kindly. A similar assumption underlies the results in Dufwenberg and Kirchsteiger (2000). Here, a profit-maximizing firm generates positive reciprocity effects from its employee by not hiring an outside worker at a lower wage. This is only possible, however, because there are just two wage levels, with the lower wage being too unkind as to make it a profitable offer. The conclusions from this strand of literature are hence largely in line with Fehr, Gächter, and Kirchsteiger (1997): Incentive schemes should make use of reciprocal motives

[2] The authors apply their model to argue that workers should be given stock options in spite of risk aversion, that stock options for the CEO have the additional cost that they increase worker envy, and to explain why the public sector (and non-profit organizations more generally) pay lower wages and use incentive pay less than the private sector.

[3] The differences between the frameworks imply that different issues are pursued otherwise. Englmaier and Leider (2008) say nothing about the optimal effort level. However, they can address the question of whether efforts should optimally be induced with relatively flat incentives (appealing to reciprocity) or with steep, strongly outcome-dependent incentives.

to improve the outcome, because firms can benefit from gift-exchange effects. Our analysis provides further support for the efficiency-enhancing role of reciprocity, but questions the potential benefits for employers.

Finally, Falk and Fischbacher (2006) develop a theory that *combines* intention- and outcome-based components. Their modeling of intentions differs, however, from the approach by Rabin (1993) and Dufwenberg and Kirchsteiger (2004), which we adopt here. We will compare their result for the gift-exchange game to our findings in Section 6.

3 General Notation and Definitions

We first introduce a general set-up that follows closely Dufwenberg and Kirchsteiger (2004), with one major difference to be emphasized below.

Game and Strategies. As Dufwenberg and Kirchsteiger (2004), we consider a multi-stage game with finitely many stages, observed actions and without nature. We restrict attention to two player games. H denotes the set of histories, where each history is a list of previous actions profiles. Each history corresponds to an information set for each player, and is also the root of a subgame. The symbol $\varnothing \in H$ represents the root of the complete game. We let A_i be the set of pure strategies for player i, with elements $a_i \in A_i$ that are mappings from H to available actions at the corresponding information set. Given $a_i \in A_i$ and $h \in H$, we denote by a_i^h the updated strategy that coincides with a_i except that we use player i's (pure) actions that would potentially lead to h instead of actions that are inconsistent with h. In particular, $a_i^{\varnothing} = a_i$.[4]

Beliefs. We denote by $b_{ij} \in A_j$ player i's beliefs about player j's strategy, and by $c_{iji} \in A_i$ player i's beliefs about player j's beliefs (about player i's strategy). Again, given a belief b_{ij} and a history h, let b_{ij}^h be the new belief updated from the belief b_{ij} by presuming that player j chooses his actions from h. Analogously, c_{iji}^h is the second order belief updated from c_{iji} using player i's actions that lead to h. We again have $b_{ij}^{\varnothing} = b_{ij}$ and $c_{iji}^{\varnothing} = c_{iji}$.

Material Payoffs. Let $A = A_1 \times A_2$. We define player i's material payoffs π_i directly on the set of strategy profiles A, so that $\pi_i(a_i, a_j)$ is player i's payoff if the strategy profile is (a_i, a_j). Further, for any history $h \in H$, let $\pi_i(a_i, a_j|h) = \pi_i(a_i^h, a_j^h)$ be the payoff of player i if the updated profile (a_1^h, a_2^h) is played instead of (a_1, a_2). In Section 7, we will suggest a

[4] This should not be confused with $a_i(h')$ or $a_i^h(h')$, which is the action that the strategies a_i and a_i^h, respectively, prescribe in information set h'.

straightforward generalization of the model in which π_i is interpreted as player i's *outcome*-based payoff, which could already contain non-pecuniary components.

Kindness. Based on any collection $(a_i, b_{ij}, c_{iji})_{i,j=1,2,i\neq j}$, we can now assign measures of kindness and beliefs about them to every information set $h \in H$. We denote by $k_{ij}(a_i, b_{ij}|h)$ the kindness of i to j in information set h.[5] We let $\lambda_{iji}(b_{ij}, c_{iji}|h)$ denote player i's belief about how kind j is to him in information set h. The definition of k_{ij} and λ_{iji} will make use of the concept of a player i's *equitable payoff*, which is the payoff that player j considers as the norm for judging the payoff that he gives to i by choice of the own strategy. Specifically, for any $h \in H$, the equitable payoff for this information set and player i is denoted by $\pi_i^e(a_i|h)$. Then, the kindness terms are defined to be

$$k_{ij}(a_i, b_{ij}|h) = \pi_j(b_{ij}, a_i|h) - \pi_j^e(b_{ij}|h)$$

and

$$\lambda_{iji}(b_{ij}, c_{iji}|h) = \pi_i(c_{iji}, b_{ij}|h) - \pi_i^e(c_{iji}|h),$$

so that positive (negative) kindness arises if a player is given a larger (smaller) material payoff than equitable.

It remains to be specified exactly how the equitable payoff is calculated. Here, we will deviate from Dufwenberg and Kirchsteiger (2004) and adopt the approach by Rabin (1993) instead. Let $\Pi_i(a_i|h) = \{(\pi_i(a_i^h, a_j), \pi_j(a_j, a_i^h))|a_j \in A_j\}$ be the set of payoff pairs that can be achieved if player i plays strategy a_i^h while player j plays an arbitrary strategy. Let $\Pi_i^E(a_i|h)$ be the Pareto efficient payoff pairs in $\Pi_i(a_i|h)$, i.e. it contains those payoff pairs from $\Pi_i(a_i|h)$ for which there is no other payoff pair in $\Pi_i(a_i|h)$ with a strictly larger payoff for one and a payoff at least as large for the other player. We will then require the equitable payoff π_i^e to correspond to an interior element of $\Pi_i^E(a_i|h)$. Dufwenberg and Kirchsteiger (2004) proceed analogously but invoke a different definition of efficiency. We follow Rabin (1993) in defining efficiency conditional on the strategy a_i chosen by player i. This assumption is important, and we will discuss it in greater detail in Section 7.

Earlier papers have used (variants of) the average between player i's largest and smallest payoff within the efficient set $\Pi_i^E(a_i|h)$ as player i's equitable payoff. We will indeed use this specific form in our later examples, but some results can be proven based on much weaker requirements. We only impose the following minimal requirements for every player, strategy, and history:

[5] Actually, k_{ij} should be called player i's belief about how kind he is to player j, because it is based on the belief b_{ij}.

(A1) (i) Whenever $\Pi_i^E(a_i|h)$ has more than one element, there exist $(\pi_i', \pi_j') \in \Pi_i^E(a_i|h)$ and $(\pi_i'', \pi_j'') \in \Pi_i^E(a_i|h)$ with $\pi_i' < \pi_i^e(a_i|h) < \pi_i''$.
(ii) If $\Pi_i^E(a_i|h) = \{(\pi_i', \pi_j')\}$, then $\pi_i^e(a_i|h) = \pi_i'$.

According to (i), the equitable payoff $\pi_i^e(a_i|h)$ does not correspond to an extreme point within $\Pi_i^E(a_i|h)$. This encompasses, for example, specifications where $\pi_i^e(a_i|h)$ is the average between the largest and the smallest payoff of player i in $\Pi_i^E(a_i|h)$, as in earlier contributions.[6] Our formulation is more general, however, and would even allow the equitable payoff to depend on how costly it is for the opponent to give player i a larger payoff within $\Pi_i^E(a_i|h)$. In addition, we also require the following:

(A2) (i) $\pi_i^e(a_i|h) = \pi_i^e(a_i^h|h)$ for both players and all strategies and histories.
(ii) $\pi_i^e(a_i|h) = \pi_i^e(a_i|h')$ if $a_i^h = a_i^{h'}$ for a player i, strategy a_i, and histories $h \neq h'$.

Intuitively, properties (A2)(i) and (A2)(ii) ensure that the equitable payoff only depends on the set of efficient payoffs that are achievable at a history h when player i plays a_i, but not on the *deviations* of player i from a_i that were made to reach h. Hence (A2)(i) and (A2)(ii) are trivially satisfied when the equitable payoff depends on the achievable payoffs only, e.g. when it is defined as the average between the largest and the smallest payoff for player i in $\Pi_i^E(a_i|h)$. (A2) also guarantees that along the equilibrium path, the equitable payoff remains unchanged.

Utility. To specify the complete utility functions that players strive to maximize, we start with a very general approach. Let $F : \mathbb{R}^2 \to \mathbb{R}$ be a function that assigns a psychological utility score $F(k_{ij}, \lambda_{iji})$ to each combination of kindness k_{ij} and belief about reciprocated kindness λ_{iji}. Throughout we assume that $F(k_{ij}, 0)$ is independent of k_{ij}, i.e. player i's kindness has no impact on own psychological utility if i expects to be treated neutrally. We also assume that $F(k_{ij}, \lambda_{iji})$ is strictly increasing in k_{ij} whenever $\lambda_{iji} > 0$ and strictly decreasing if $\lambda_{iji} < 0$. We do not impose any assumptions about curvature, continuity, or even differentiability of F.[7] In Dufwenberg and Kirchsteiger (2004), $F(k_{ij}, \lambda_{iji}) = k_{ij}\lambda_{iji}$. Rabin (1993) imposes $F(k_{ij}, \lambda_{iji}) = \lambda_{iji}(1 + k_{ij})$. In contrast, our more general formulation would allow for decreasing marginal psychological utility, i.e. the F is not necessarily linear. Finally, we let $y_i \geq 0$ denote the relative weight that player i assigns to psychological payoffs.

[6] Observe that (A1)(i) does not require that $\pi_i^e(a_i|h)$ is itself the payoff of player i in an element of $\Pi_i^E(a_i|h)$.

[7] Some combination of such assumptions will, of course, become necessary to guarantee equilibrium existence, but several general results can be proven without them.

Then, for every information set $h \in H$ and player i, let

$$U_i(a_i, b_{ij}, c_{iji}|h) = \pi_i(a_i, b_{ij}|h) + y_i F\left(k_{ij}(a_i, b_{ij}|h), \lambda_{iji}(b_{ij}, c_{iji}|h)\right) \tag{1}$$

be player i's utility in information set h, which is based on material payoffs from updated strategies, and contains the (updated) reciprocity term added with a weight of y_i.

Equilibrium. Following Dufwenberg and Kirchsteiger (2004) or Battigalli and Dufwenberg (2009), we require sequential rationality, that is, each player must maximize U_i in each information set $h \in H$. To distinguish our approach from the sequential reciprocity equilibrium by Dufwenberg and Kirchsteiger (2004), due to the differences discussed above, we refer to equilibria as "intentions equilibria".

Definition 1. *A strategy profile $(\hat{a}_1, \hat{a}_2) \in A$ is an intentions equilibrium (IE) if for $i = 1, 2$, $j \neq i$, and all $h \in H$,*

(i) $\hat{a}_i \in \arg\max_{a_i \in A_i} U_i(a_i, b_{ij}, c_{iji}|h)$,

(ii) $b_{ij} = \hat{a}_j$, and

(iii) $c_{iji} = \hat{a}_i$.

4 The Intentions of Rotten Kids

From now on, and for the rest of the paper, we will assume that player 1 is materialistic ($y_1 = 0$) while player 2 is motivated by reciprocal concerns ($y_2 > 0$). In this setting, we provide the general result that behavior is never kind in any intentions equilibrium.

Proposition 1. *Suppose (A1) and (A2) hold and $y_2 > y_1 = 0$. Then, in any IE $(\hat{a}_1, \hat{a}_2)$, it holds that $k_{ij}(\hat{a}_i, \hat{a}_j|h) \leq 0$ for $i = 1, 2$, $i \neq j$ and any $h \in H$ that is reached on the equilibrium path. The inequalities for k_{ij} are strict if $|\Pi_j^E(\hat{a}_j|\varnothing)| \geq 2$.*

Proof. See Appendix A.1. □

The Proposition states that, with reciprocity concerns and one materialistic player, any equilibrium must necessarily involve negative emotions, in every subgame that is reached, including, of course, the complete game starting from history $h = \varnothing$. It is remarkable at what level of generality this result holds. Specifically, it holds for any (two-player, finite stage, observed actions, no nature) game under only minimal assumptions (A1) and (A2) on equitable payoffs and no additional assumptions on F.

To understand the intuition, first consider the perspective of player 1. Since $y_1 = 0$, the optimality condition for player 1 from Definition 1 simplifies to $\hat{a}_1 \in \arg\max_{a_1 \in A_1} \pi_1(a_1, \hat{a}_2|h)$, where $b_{12} = \hat{a}_2$ (condition (ii) in Definition 1) has been substituted. Given any strategy of player 2, player 1 will choose the strategy that maximizes his material payoffs. Thus, whenever there is a conflict of interest between the two players concerning the (materially Pareto efficient) allocations that player 1 can induce, he will avoid leaving more material payoff on the table than necessary, which is unkind behavior.[8] Player 2 in turn does not expect to be treated kindly and does not benefit from being kind to player 1 either.

This result is in stark contrast to the original Rotten Kid intuition. With altruism on the side of the parent, a rotten kid anticipates monetary rewards for increasing family income. Maximizing joint income then becomes the kid's self-interest. But acts motivated by self-interest are not kind, such that an analogous argument fails with intention-based preferences. As we will discuss in greater detail in Section 5, the resulting IE share quite a few properties with equilibria under envy preferences (Dur and Glazer 2008), rather than altruism. This is interesting as our model allows for both positive and negative emotions *a priori*, and envy-type behavior arises endogenously.

From this different perspective, Proposition 1 could also be interpreted as an equilibrium selection result. As already emphasized by Rabin (1993), intention-based preferences generally imply a multiplicity of equilibria, some of them with kind, others with unkind behavior. While Rabin (1993) has proven that an equilibrium with negative kindness always exists in his setup, we show that making one of the players materialistic eliminates any possibility for positive equilibrium kindness, so that *only* unkindness survives.

We could even strengthen our result as follows. From the above arguments it is clear that player 1 chooses a strategy that gives player 2 minimal material payoffs among the efficient payoff combinations for given a_2. This is stronger than just being unkind: it means player 1 is as unkind as possible, without violating Pareto efficiency. The same then holds for player 2, who obtains increasing psychological utility from decreasing player 1's material payoff. Formally, in any IE $(\hat{a}_1, \hat{a}_2)$, if $a_1' \in A_1$ and $(\pi_1(a_1', \hat{a}_2), \pi_2(\hat{a}_2, a_1')) \in \Pi_2^E(\hat{a}_2|\varnothing)$, then $\pi_2(\hat{a}_2, a_1') \geq \pi_2(\hat{a}_2, \hat{a}_1)$, and analogously for the other player.

5 Action-Reaction Games

5.1 A Class of Two-Stage Games

To explore the implications of Proposition 1, we now consider a structure analogous to the original Rotten Kid model, where the materialistic player 1 moves first, followed by the

[8] Observe that, off the equilibrium path, kind behavior is compatible with equilibrium, as will be illustrated in Section 5.

reciprocal player 2 ("action-reaction games"). Specifically, in the root ($h = \varnothing$) player 1 chooses a "wage" from some set W. This wage becomes observable and player 2 reacts by choosing an "effort" from a set E. We can thus simplify notation and write $a_1 \in A_1 = W$ and $a_2 : W \rightarrow E$, with $A_2 = E^W$ being the set of all such functions. The set of histories can be written as $H = \{\varnothing\} \cup W$.[9]

We write $\tilde{\pi}_1(w, e)$ and $\tilde{\pi}_2(e, w)$ to denote the players' payoffs defined on action profiles, such that the above introduced payoffs π_i on strategy profiles are $\pi_1(a_1, a_2) = \tilde{\pi}_1(a_1, a_2(a_1))$ and $\pi_2(a_2, a_1) = \tilde{\pi}_2(a_2(a_1), a_1)$. Also, since we always assume that player 1 is materialistic ($y_1 = 0$), we denote player 2's reciprocity intensity by y, skipping the player index.

According to Definition 1, player 1's strategy has to be optimal (only) in $h = \varnothing$, i.e. the equilibrium wage must maximize $\tilde{\pi}_1(w, \hat{a}_2(w))$, because $y_1 = 0$ and $b_{12} = \hat{a}_2$ in equilibrium. Player 2 must best-respond to every history $w \in W$, which implies that $U_2(a_2, b_{21}, c_{212}|w)$ must be maximized for every $w \in W$. Under (A2), we can substitute the updated belief $b_{21}^w = w$ for b_{21} in the equitable payoff $\pi_1^e(b_{21}|w)$ and simplify notation to $\pi_1^e(w) \equiv \pi_1^e(w|w)$. Observing that $a_2^w = a_2$, $\forall w \in W$, we can analogously simplify to $k_{21}(e, w) \equiv \tilde{\pi}_1(w, e) - \pi_1^e(w)$. Since $c_{212}^w = c_{212}$, $\forall w \in W$, the equitable payoff $\pi_2^e(c_{212}|w)$ must be independent of w under (A2) and can be simplified to $\pi_2^e(c_{212})$. This, finally, makes it possible to simplify $\lambda_{212}(w, c_{212}) \equiv \tilde{\pi}_2(c_{212}(w), w) - \pi_2^e(c_{212})$, and we can summarize:

Lemma 1. *Suppose (A2) holds. A strategy profile $(\hat{a}_1, \hat{a}_2) \in A$ is an IE of the action-reaction game if and only if*
(i) $\hat{a}_1 \in \arg\max_{w \in W} \tilde{\pi}_1(w, \hat{a}_2(w))$, and
(ii) $\hat{a}_2(w) \in \arg\max_{e \in E} \tilde{\pi}_2(e, w) + yF\left(k_{21}(e, w), \lambda_{212}(w, \hat{a}_2)\right)$ for all $w \in W$.

Observe that, in condition (*ii*), maximization is over effort levels $e \in E$, so that $\hat{a}_2$ on the RHS is treated as fixed.

We now specify further assumptions. First, we assume that both W and E are compact subsets of $\mathbb{R}$, denoted by $[\underline{w}, \overline{w}]$ and $[\underline{e}, \overline{e}]$ whenever convex. The payoff functions $\tilde{\pi}_i$ are assumed to be continuously differentiable on every open subset of their domain. The following assumptions specify the economic substance of the game.

(A3) (i) $\tilde{\pi}_1(w, e)$ is strictly decreasing in w and strictly increasing in e.
(ii) $\tilde{\pi}_2(w, e)$ is strictly increasing in w.
(iii) For each $w \in W$, there is a unique effort level that maximizes $\tilde{\pi}_2(e, w)$ on E.

[9] To be precise, these simplifications fit into the general framework introduced in section 3 by assuming that both players move in both periods, i.e. after all histories, but their action sets are singletons whenever they do actually not play. The set of histories is simplified correspondingly, by recording only actions that arise from actual choice.

Parts (i) and (ii) capture a conflict of interest with respect to the wage w. Also, player 1 always prefers a larger effort level of player 2. Part (iii) is less strict: it allows for the possibility that player 2 himself finds some effort materially desirable, as will be the case in the following moral hazard application for positive incentive wages. Finally, for the rest of the paper we will assume that assumptions (A1) and (A2) hold, without further mention. We conclude this section by introducing two main examples for our setup.

Example 1: Gift Exchange.

In a gift exchange game, $W = [0, \overline{w}]$, $E = [0, \overline{e}]$, $\tilde{\pi}_1(w, e) = e - w$ and $\tilde{\pi}_2(e, w) = v(w) - e$ for some continuously differentiable, strictly increasing and strictly concave function v.

Example 2: Moral Hazard.

(a) Assume player 2 (the agent) chooses an unobservable effort level from $E = [0, 1]$, which is interpreted as the probability of completing a project successfully. The success of the project is observable, with project payoffs $V > 0$ in case of success and zero otherwise. Player 1 (the principal) offers a wage (or bonus) from $W = [0, V]$ to be paid in case of success. Hence expected payoffs are $\tilde{\pi}_1(w, e) = e(V - w)$ and $\tilde{\pi}_2(e, w) = ew - d(e)$, where d is a strictly increasing and strictly convex effort cost function.[10]

(b) Several insights can already be gained in a binary version of the model, where $E = \{0, p\}$ for some $p \in (0, 1)$.[11]

In the following section, we will present general results for the class of action-reaction games. We return to the above examples in Section 6.

5.2 Equilibrium Inefficiency

First we want to examine whether intention-based social preferences foster the emergence of materially Pareto efficient outcomes as in the models discussed in Section 2. Benjamin (2008) distinguishes between *material* and *utility* efficiency, where the latter also takes into account utility from social comparisons. With intentions-based preferences, the concept of overall utility efficiency, which includes the psychological utility component, is problematic. Knowing the outcome of an interaction is not sufficient to derive psychological utility, because

[10] Strictly speaking, these payoff functions do not satisfy assumption (A3), because for $e = 0$, $\tilde{\pi}_1$ is only weakly decreasing and $\tilde{\pi}_2$ is only weakly increasing in w. For $w = V$, $\tilde{\pi}_1$ is only weakly increasing in e. None of this, however, will constitute a problem in the following.

[11] Again, assumption (A3)(iii) is not exactly satisfied in the binary model, because at wage $w = (d(p) - d(0))/p$ player 2 will be indifferent between the effort levels.

it depends on the way the outcome was achieved. Hence we cannot derive a utility Pareto frontier a priori, and then compare equilibrium outcomes to this frontier. For that reason, we refrain from analyzing utility Pareto efficiency and focus on material efficiency only.

Under assumption (A3)(iii), player 2 has a unique material best-response to any wage w. Then, let $\tilde{a}_2$ denote the best-response function defined by $\tilde{a}_2(w) = \arg\max_{e\in E} \tilde{\pi}_2(e, w)$, $\forall w \in W$. As an initial step, we are going to show under which conditions player 2 does indeed deviate from his material best reply to punish player 1 for being unkind in equilibrium.[12]

Lemma 2. *In any IE $(\hat{a}_1, \hat{a}_2)$ it holds that $\hat{a}_2(\hat{a}_1) \leq \tilde{a}_2(\hat{a}_1)$, with strict inequality if $|\Pi_2^E(\hat{a}_2)| \geq 2$ and $\tilde{a}_2(\hat{a}_1) \in int\, E$.*

Proof. Fix any IE $(\hat{a}_1, \hat{a}_2)$. By Lemma 1,

$$\hat{a}_2(\hat{a}_1) \in \arg\max_{e\in E} \tilde{\pi}_2(e, \hat{a}_1) + yF(\tilde{\pi}_1(\hat{a}_1, e) - \pi_1^e(\hat{a}_1), \lambda), \tag{2}$$

where $\lambda = \lambda_{212}(\hat{a}_1, \hat{a}_2)$ is independent of e and satisfies $\lambda \leq 0$ according to Proposition 1, with strict inequality if $|\Pi_2^E(\hat{a}_2)| \geq 2$. Also, $\tilde{a}_2(\hat{a}_1)$ by definition maximizes $\tilde{\pi}_2(e, \hat{a}_1)$, the first term in (2). Then, if $\lambda = 0$, we must have $\hat{a}_2(\hat{a}_1) = \tilde{a}_2(\hat{a}_1)$, because $F(k_{21}, 0)$ is independent of k_{21}, so that the reciprocity term can be omitted in (2). If $\lambda < 0$, the reciprocity term $yF(\tilde{\pi}_1(\hat{a}_1, e) - \pi_1^e(\hat{a}_1), \lambda)$ is strictly decreasing in e, which immediately implies $\hat{a}_2(\hat{a}_1) \leq \tilde{a}_2(\hat{a}_1)$. If $\tilde{a}_2(\hat{a}_1) \in \text{int}\, E$, then $\tilde{a}_2(\hat{a}_1)$ satisfies the necessary first order condition $\partial\tilde{\pi}_2(\tilde{a}_2(\hat{a}_1), \hat{a}_1)/\partial e = 0$, so that the objective (2) is strictly decreasing in e at $e = \tilde{a}_2(\hat{a}_1)$, implying $\hat{a}_2(\hat{a}_1) < \tilde{a}_2(\hat{a}_1)$. □

The lemma states that player 2 responds to the IE wage $\hat{a}_1$ with weakly less effort than would be optimal from a purely materialistic perspective. Whenever $|\Pi_2^E(\hat{a}_2)| \geq 2$, so that player 1 is strictly unkind in equilibrium (see Proposition 1), and the materially optimal effort level is not a corner solution, then the equilibrium effort is strictly lower.[13] Since player 1 suffers from reduced effort, this is equivalent to saying that player 2 punishes player 1 at an own material cost. From this argument we obtain the following immediate corollary.

Corollary 1. *Any IE $(\hat{a}_1, \hat{a}_2)$ with $|\Pi_2^E(\hat{a}_2)| \geq 2$ and $\tilde{a}_2(\hat{a}_1) \in int\, E$ is materially Pareto inefficient.*

So whenever there are indeed conflicts of interest ($|\Pi_2^E(\hat{a}_2)| \geq 2$) and punishment is viable ($\tilde{a}_2(\hat{a}_1) \in \text{int}\, E$), there is no hope to obtain an efficiency result in the spirit of the Rotten Kid Theorem for the case of intention-based social preferences.

[12] The expression $\Pi_2^E(\hat{a}_2)$ stands short for $\Pi_2^E(\hat{a}_2|\varnothing)$ in the following lemma.

[13] The qualification that $\tilde{a}_2(\hat{a}_1)$ must be an interior solution is necessary (i) to insure that reducing effort below $\tilde{a}_2(\hat{a}_1)$ is actually possible, and (ii) because $\hat{a}_2(\hat{a}_1)$ might remain an upper corner solution if $\tilde{a}_2(\hat{a}_1)$ is one.

5.3 The Materialistic Player Suffers

From now on, we want to compare the outcome of IE (when player 2 is reciprocal) to that of subgame-perfect equilibria (SPE) when both players are materialistic. For $y_1 = y_2 = 0$ our definition of IE becomes the standard SPE definition, as is obvious from Lemma 1. We will still use the different terms IE and SPE to avoid confusion. In any SPE, player 2 must clearly play the strategy $\tilde{a}_2$. Player 1 could still have more than one best reply $\tilde{a}_1$, making multiple SPE possible. We can now use the insights from Lemma 2 to compare the materialistic player's payoff in SPE and IE.

Proposition 2. *For any SPE $(\tilde{a}_1, \tilde{a}_2)$ and IE $(\hat{a}_1, \hat{a}_2)$ it holds that $\pi_1(\hat{a}_1, \hat{a}_2) \leq \pi_1(\tilde{a}_1, \tilde{a}_2)$, with strict inequality if $\hat{a}_2(\hat{a}_1) < \tilde{a}_2(\hat{a}_1)$.*

Proof. We have that $\pi_1(\hat{a}_1, \hat{a}_2) = \tilde{\pi}_1(\hat{a}_1, \hat{a}_2(\hat{a}_1)) \leq \tilde{\pi}_1(\hat{a}_1, \tilde{a}_2(\hat{a}_1)) = \pi_1(\hat{a}_1, \tilde{a}_2)$, because $\hat{a}_2(\hat{a}_1) \leq \tilde{a}_2(\hat{a}_1)$ according to Lemma 2 and because $\tilde{\pi}_1$ is increasing in e. The inequality is strict whenever $\hat{a}_2(\hat{a}_1) < \tilde{a}_2(\hat{a}_1)$. To obtain a contradiction, first assume $\pi_1(\tilde{a}_1, \tilde{a}_2) < \pi_1(\hat{a}_1, \hat{a}_2)$. Together with the above inequality this implies $\pi_1(\tilde{a}_1, \tilde{a}_2) < \pi_1(\hat{a}_1, \tilde{a}_2)$, which contradicts $\tilde{a}_1 \in \arg\max_{w \in W} \pi_1(w, \tilde{a}_2)$ and hence that $(\tilde{a}_1, \tilde{a}_2)$ is an SPE. If $\hat{a}_2(\hat{a}_1) < \tilde{a}_2(\hat{a}_1)$, we obtain an analogous contradiction under the assumption that $\pi_1(\tilde{a}_1, \tilde{a}_2) \leq \pi_1(\hat{a}_1, \hat{a}_2)$. □

Proposition 2 shows that the materialistic player does not profit from facing an opponent who is reciprocal. His equilibrium payoff in *any* IE must necessarily be weakly smaller than in *any* SPE, and strictly so whenever punishment actually takes place in the IE, which, according to Lemma 2, will be the case except if there are common interests. This result stands in stark contrast to results obtained with outcome-based preferences. For completeness, the following proposition confirms this claim within an altruism model that is similar in generality to our intention-based model. We assume that player 1 still maximizes material payoffs $\pi_1(a_1, a_2)$. Player 2 is altruistic, maximizing $\pi_2(a_2, a_1) + G(\pi_1(a_1, a_2))$, where G is an arbitrary but strictly increasing function of player 1's payoff. We are interested in subgame-perfect equilibria $(\bar{a}_1, \bar{a}_2)$ of the action-reaction game with altruism, which we refer to as *altruism equilibria* (AE).

Proposition 3. *For any SPE $(\tilde{a}_1, \tilde{a}_2)$ and AE $(\bar{a}_1, \bar{a}_2)$ it holds that $\pi_1(\tilde{a}_1, \tilde{a}_2) \leq \pi_1(\bar{a}_1, \bar{a}_2)$, with strict inequality if $\tilde{a}_2(\tilde{a}_1) \in int\, E$.*

Proof. Arguing as in the proof of Lemma 2, it immediately follows that $\tilde{a}_2(\tilde{a}_1) \leq \bar{a}_2(\tilde{a}_1)$, with strict inequality if $\tilde{a}_2(\tilde{a}_1) \in \operatorname{int} E$. But then $\pi_1(\tilde{a}_1, \tilde{a}_2) = \tilde{\pi}_1(\tilde{a}_1, \tilde{a}_2(\tilde{a}_1)) \leq \tilde{\pi}_1(\tilde{a}_1, \bar{a}_2(\tilde{a}_1)) = \pi_1(\tilde{a}_1, \bar{a}_2)$, with strict inequality if $\tilde{a}_2(\tilde{a}_1) < \bar{a}_2(\tilde{a}_1)$. Since $\bar{a}_1 \in \arg\max_{a_1 \in A_1} \pi_1(a_1, \bar{a}_2)$ by definition of AE, we obtain $\pi_1(\tilde{a}_1, \bar{a}_2) \leq \pi_1(\bar{a}_1, \bar{a}_2)$, which completes the proof. □

Propositions 2 and 3 together show that reciprocity and altruism have completely opposite effects concerning player 1's payoff. Depending on the specific application, this can have quite important implications. For example, we could conclude that a profit-maximizing principal should try to hire altruistic agents but stay away from reciprocators, because he can exploit the social preferences of the first but not those of the latter.

5.4 Equilibrium Wages and Efforts

We now examine how reciprocity affects the equilibrium actions. To do so, we invoke some additional conditions. The upshot of the analysis will be that reciprocity typically has a positive effect on wages and a more ambiguous effect on efforts.

Our analysis in the following will be based on the standard psychological utility score $F\left(k_{ij}, \lambda_{iji}\right) = k_{ij} \cdot \lambda_{iji}$. Player 2's optimality condition from Lemma 1 thus becomes $\hat{a}_2(w) \in \arg\max_{e \in E} \tilde{\pi}_2(e, w) + y \cdot (\tilde{\pi}_1(w, e) - \pi_1^e(w)) \cdot \lambda_{212}(w, \hat{a}_2)$ for all $w \in W$. We will also assume convexity of both W and E, i.e. $W = [\underline{w}, \overline{w}]$ and $E = [\underline{e}, \overline{e}]$. Second, some additional but standard assumptions on payoff functions will be used, which are satisfied in both Example 1 (gift-exchange) and Example 2 (moral hazard).

(A4) (i) $\tilde{\pi}_1(w, e)$ is submodular on $W \times E$.
(ii) $\tilde{\pi}_2(e, w)$ is supermodular on $E \times W$.
(iii) $\tilde{\pi}_1(w, e)$ is (weakly) concave in e.

The following proposition again compares IE to SPE. For sake of clarity, it applies only to the simplified case when there is a unique SPE $(\tilde{a}_1, \tilde{a}_2)$, i.e. a unique value $\tilde{a}_1 \in W$ that maximizes $\tilde{\pi}_1(w, \tilde{a}_2(w))$. The result is readily generalizable to allow for multiple SPE, with its conclusion becoming a comparison between largest and/or smallest wages across equilibria. The proposition does, however, not require IE to be unique.

Proposition 4. *Suppose that (A3) and (A4) hold. Then $\tilde{a}_1 \leq \hat{a}_1$ holds for any IE $(\hat{a}_1, \hat{a}_2)$ in which $\Delta(w) \equiv \tilde{a}_2(w) - \hat{a}_2(w)$ is weakly decreasing in w on $[\underline{w}, \tilde{a}_1]$ or $[\hat{a}_1, \overline{w}]$ (or both).*

Proof. See Appendix A.2. □

The proposition applies to all those IE in which the punishment $\Delta(w)$ is decreasing over a suitable range of wages. In a sense, these are equilibria that preserve some properties of the SPE, namely that increasing the wage increases player 2's payoffs, hence reduces unkindness, and leads player 2 to punish less. We will show in Section 6 that, in several standard applications, any equilibrium exhibits this property.

To grasp the intuition for the result, focus on the condition that $\Delta(w)$ is decreasing on $[\underline{w}, \tilde{a}_1]$ and observe that there are three effects of reciprocity on equilibrium wages, under the assumptions of the proposition. First, we already know that player 2 responds with lower effort (than materially optimal) to the equilibrium wage and, in fact, to any unkind wage offer. Because player 1 is expecting lower effort, he wants to give higher wages by submodularity (A4)(i). The second effect is more subtle: because $\tilde{\pi}_1$ is concave in efforts by (A4)(iii), a lower value of effort makes it more attractive for player 1 to induce more effort by further increasing his wage offer. Finally, consider the third effect: the assumption that $\Delta(w)$ is decreasing in the range $[\underline{w}, \tilde{a}_1]$ implies that for all wages smaller than the SPE wage $\tilde{a}_1$, the wage-sensitivity of effort is larger in the IE than in the SPE. This again strengthens player 1's incentive to increase wages.

The proposition thus tells us that we can indeed expect reciprocity to have a positive impact on the level of transfers from player 1 to player 2, as in the standard rotten kid model or other previously discussed models with outcome-based social preferences. In our model, the reason for such increased transfers is of course player 1's attempt to reduce punishment by player 2, rather than the hope to trigger a benevolent gift in return. Thus it needs to be pointed out that, because both outcome- and intention-based models have some predictions in common, the empirical occurrence of such phenomena does not yet lend support to purely outcome-based models.

The results so far suggest an ambiguous relation between reciprocity and effort. On the one hand, we already know by Lemma 2 that IE efforts are lower than the material best response of player 2 to the IE equilibrium wage, due to punishment taking place. On the other hand, under the assumptions of Proposition 4, wages in IE will be higher than in SPE, which, with upward-sloping reaction functions, suggests higher efforts. This ambiguity is confirmed by a closer look at the economic intuition. Recall the following properties:

(a) (A4) requires $\tilde{\pi}_1(w, e)$ to be submodular.

(b) By Lemma 2, efforts at IE wages are lower with IE reaction functions than with the SPE reaction function. The same can be shown (and actually has been shown in the proof of Proposition 4) to be true at SPE wages.

(c) We have argued (and will show in our examples) that, in regions with negative kindness, the wage-sensitivity of effort should be higher in the IE than in the SPE.

The first two properties together suggest that, near the SPE effort, the benefits of inducing higher effort are lower with reciprocal players than with materialistic players: By (b), near the SPE and IE efforts, reciprocity concerns increase the wage level required to reach a desired effort level. By (a), at these higher wage levels, inducing higher efforts has a smaller effect on player 1's payoffs, because a greater part of the benefits accrues to player 2. Property (c), however, suggests that the costs of inducing higher efforts are also lower, because player 2 reacts more strongly to wage increases. The net effect of these countervailing forces is unclear at this level of generality. However, in all the examples in Section 6 where there is a change in effort at all, the cost-reduction effect dominates.

6 Examples

6.1 Gift Exchange

Gift exchange games have served as example both in theoretical and in experimental work (see Falk and Fischbacher (2006) and the discussion therein). Therefore, we briefly examine the implications of one-sided reciprocity in our gift exchange game as a starting point. Assuming an interior solution, the efficient gift $w^* \in (0, \bar{w})$ is characterized by the condition $v'(w^*) = 1$. We also immediately obtain $\tilde{a}_2(w) = 0$ for all w, which implies that there is a unique SPE $(\tilde{a}_1, \tilde{a}_2) = (0, \tilde{a}_2)$ where no gift is given and no transfer is paid. An analogous statement holds for IE.

Proposition 5. *In the gift exchange game, any IE $(\hat{a}_1, \hat{a}_2)$ satisfies $\hat{a}_1 = 0$ and $\hat{a}_2(\hat{a}_1) = 0$.*

Proof. Lemma 2 implies $\hat{a}_2(\hat{a}_1) \leq \tilde{a}_2(\hat{a}_1) = 0$, so that $\hat{a}_2(\hat{a}_1) = 0$ because $\hat{a}_2(\hat{a}_1) \in [0, \bar{e}]$. Then, if $\hat{a}_1 > 0$, we have $\tilde{\pi}_1(\hat{a}_1, \hat{a}_2) = 0 - \hat{a}_1 < 0 \leq \hat{a}_2(0) - 0 = \tilde{\pi}_1(0, \hat{a}_2)$, a contradiction. □

Proposition 5 tells us that intention-based preferences, in contrast to outcome-based altruism, do not help to solve the inefficiency problem in the simple gift exchange game. Falk and Fischbacher (2006, p. 305) present a result according to which gift and transfer are strictly positive even if only player 2 has social preferences.[14] In their model, outcome- and intention-based components are intertwined in the definition of social preferences. Our above result indicates that any deviation from the SPE must be due to the outcome-based component.

6.2 Moral Hazard

6.2.1 The Binary Model

We now turn to the richer moral hazard game, and first examine its binary version in greater detail. Despite its simplicity, there is a variety of interesting insights to be learned from it.

[14] Falk and Fischbacher (2006) use the slightly different material payoff functions $\tilde{\pi}_1(w, e) = ve - w$ and $\tilde{\pi}_2(e, w) = w - \alpha e^2$.

Without loss of generality, we normalize $d(0) = 0$ and write $d(p) = d > 0$. Furthermore, we assume that $pV > d$, which implies that the large effort $e = p$ is the efficient action. For the reciprocity part of preferences we use the standard functional form $F(k_{ij}, \lambda_{iji}) = k_{ij}\lambda_{iji}$. We calculate the equitable payoff $\pi_i^e(a_i)$ as the mean between the largest and smallest payoff of player i within the efficient set $\Pi_i^E(a_i)$, in line with previous models.[15] Formally, let $\pi_i^l(a_i) = \inf \{\pi_i | (\pi_i, \pi_j) \in \Pi_i^E(a_i)\}$ and $\pi_i^h(a_i) = \sup \{\pi_i | (\pi_i, \pi_j) \in \Pi_i^E(a_i)\}$, and define $\pi_i^e(a_i) = (\pi_i^l(a_i) + \pi_i^h(a_i))/2$.

The binary game has a unique SPE $(\tilde{a}_1, \tilde{a}_2)$ in which $\tilde{a}_2$ is of a cut-off form: for any $w < d/p$ we must have $\tilde{a}_2(w) = 0$, while $\tilde{a}_2(w) = p$ for all $d/p \leq w$.[16] Player 1 then pays a wage of $\tilde{a}_1 = d/p$ and player 2 supplies effort. We can compare this outcome with the outcomes of IE, which are characterized in the following proposition.

Proposition 6. *Consider the binary moral hazard game with $pV > d$. There exist two values w^l and w^h with $d/p < w^l < w^h < V$, such that*
(i) any IE $(\hat{a}_1, \hat{a}_2)$ satisfies $\hat{a}_1 \in [w^l, w^h]$ and $\hat{a}_2(\hat{a}_1) = p$, and
(ii) for any $\hat{w} \in [w^l, w^h]$ there exists an IE $(\hat{a}_1, \hat{a}_2)$ with $\hat{a}_1 = \hat{w}$ and $\hat{a}_2(\hat{a}_1) = p$.

Proof. See Appendix, Section A.3. □

According to the proposition, the wage paid in any IE must be from the interval $[w^l, w^h]$, and it actually induces the high effort. Conversely, there is such an IE for any wage $\hat{w} \in [w^l, w^h]$. The equilibria we construct to prove the second claim are of the same cut-off form as the SPE: player 2 supplies effort only for wages $w \geq \hat{w}$. It is worth pointing out, however, that not all IE are necessarily cut-off equilibria. Player 2 could supply effort at wage $\hat{w}$ but not at some higher wage w'. If player 1, who anticipates this, actually offers wage w', he must be considered especially unkind because he intends to induce zero effort and give player 2 a payoff of zero. This can make it optimal from player 2's perspective to punish player 1 by indeed supplying no effort.[17] Such equilibria are rather counterintuitive, and we want to emphasize that no such construction is used to prove statement (*ii*) in Proposition 6.

The proposition illustrates some of our general results from Sections 4 and 5. Consider any IE $(\hat{a}_1, \hat{a}_2)$, where $\hat{a}_1 \in [w^l, w^h]$ must hold according to statement (*i*). Then, as shown in the proof of the proposition, player 2's (correct) belief about player 1's equilibrium kindness is given by $\lambda_{212}(\hat{a}_1, \hat{a}_2) = -(p/2)(V - \hat{a}_1)$, which is strictly negative even in the IE with largest possible equilibrium wage $w^h < V$. That is what Proposition 1 predicts. On the

[15] As argued before, we can omit conditioning on the history in our two-stage action-reaction game.

[16] At a wage of d/p, player 2 is indifferent between both effort levels. An equilibrium with $\tilde{a}_2(d/p) = 0$ cannot exist, however, because a best-response of player 2 would not exist for such a strategy. This, essentially, is the reason why the violations of Assumption A(3) that we discussed earlier are innocuous.

[17] This artefact has already been observed by Falk, Fehr, and Fischbacher (2003b, p. 294f).

other hand, even the lowest possible equilibrium wage w^l is still larger than d/p, i.e. the wage paid in the unique SPE. These two facts together are rather surprising: player 1 pays a higher wage in any IE than in the SPE, but this does not imply he is considered kind: after all, player 1 is still a rotten kid with bad intentions.

Player 2 is faced with the following trade-off. For wages $d/p \leq w$, supplying the high effort $e = p$ is the unique action that induces a Pareto efficient payoff pair. This implies that supplying effort is not a kind action from point of view of player 2, but it is simply neutral ($k_{21}(p, w) = 0$). Strictly positive kindness would require the sacrifice of own payoffs in favor of player 1. Then, since $F(k_{ij}, \lambda_{iji}) = k_{ij}\lambda_{iji}$, the reciprocity term is irrelevant for player 2's evaluation of high effort, so that only material payoffs matter for him. On the other hand, if he decides to supply no effort ($e = 0$), his material payoffs are zero and only psychological utility matters. Hence, to induce effort, player 1 needs to make effort sufficiently attractive from a purely material perspective, relative to the psychologically appealing option of punishment. Higher wages help on both sides: they make effort materially more rewarding, and they reduce player 1's potential payoff and therefore decrease player 2's scope for punishment.[18]

The fact that the high effort $e = p$ is supplied both in the SPE and in *any* IE is also of interest. First, it makes it possible to compare material equilibrium payoffs simply by comparing equilibrium wages. Proposition 6 therefore confirms Proposition 2's prediction that player 1 is worse off in any IE then in the SPE. Analogously, player 2 is strictly better off. On the other hand, both the SPE and all IE are materially Pareto efficient due to high effort. The reason that this is true for IE is, of course, the (artificial) restriction to binary actions, which implies that $\tilde{a}_2(\hat{a}_1) \notin \operatorname{int} E$ and makes Corollary 1 inapplicable. We will return to this issue in the next subsection when discussing a richer moral hazard model.

Finally, we can examine the boundary values w^l and w^h more closely. As shown in the proof of Proposition 6, they are implicitly defined by

$$y\left[(p/2)(V - w^l)\right] = \frac{pw^l - d}{p(V - w^l)}, \text{ and} \tag{3}$$

$$y\left[(p/2)(V + w^h) - d\right] = \frac{pw^h - d}{p(V - w^h)}. \tag{4}$$

We can explicitly write them as functions of the parameters y, V and d and derive comparative static effects.

[18] We would also want to point out that $\tilde{a}_1 \leq \hat{a}_1$ for any IE also follows from our general Proposition 4. The punishment function $\Delta(w) = \tilde{a}_2(w) - \hat{a}_2(w)$ is constant and equal to zero for all wages $w \in [0, d/p]$, which makes the proposition applicable.

Proposition 7. *The values $w^l(y, V, d)$ and $w^h(y, V, d)$ are strictly increasing in y, V and d, with $\lim_{y\to\infty} w^l(y, V, d) = \lim_{y\to\infty} w^h(y, V, d) = V$.*

Proof. See Appendix, Section A.4. □

The boundary values are strictly increasing in all three parameters. For the effort cost d, this effect is standard: If supplying effort becomes more expensive, player 2 demands a higher wage, i.e. both the largest and the smallest equilibrium wage increases. The same comparative statics effect does hold for the SPE wage $\tilde{a}_1$. The SPE wage is independent of V, however: If player 2 cares only for own material payoffs, the surplus left to player 1 is irrelevant in his effort decision. This is no longer the case with intention-based preferences. For a given wage, increasing the project payoff V makes the option of punishment more attractive, because player 2 can deprive player 1 of higher payoffs by not supplying effort. In this sense, a larger project value V implies a larger potential for sabotage, and the reciprocal player wants to be compensated for not using this option.

This finding could again have interesting implications for job design. On the one hand we should expect jobs with more responsibility to be paid better, even if they require exactly the same skills. On the other hand, employers could benefit from systematically structuring jobs so as to minimize what we have called potential for sabotage, even if destructive behavior is never observed in equilibrium.

The positive effect of the degree of reciprocity y on w^l and w^h is now also obvious. As psychological payoffs become more important, larger material payoffs are required to still supply effort to the unkind player 1. More interesting is the statement that, as $y \to \infty$, both w^l and w^h converge to V. This implies that player 2's material payoffs become larger and he eventually reaps the complete gains from trade as he cares less and less for material payoffs.

6.2.2 A Continuous Model

We now modify the principal-agent example to make the following two points. First, we show the robustness of the key insights from the binary model. Second, in the new example the effects of reciprocity are not merely distributionary; instead, efforts and hence efficiency will be affected positively compared to the SPE, while material Pareto efficiency is at least not generally attainable.

We modify the previous example by assuming that $E = [0, 1]$ and $d(e) = e^2$, which results in monetary payoffs $\tilde{\pi}_1(w, e) = e(V - w)$ and $\tilde{\pi}_2(e, w) = ew - e^2$. Analogous to the assumption $pV > d$ in the previous section, we will assume $V > 2$ which implies that full effort $e = 1$ is efficient. Then it is immediate to show that $\tilde{a}_2(w) = \min\{w/2, 1\}$ is player 2's material best response, and player 1 will offer the wage $\tilde{a}_1 = \min\{V/2, 2\}$ in the (unique)

SPE. We are now first interested whether there are still IE $(\hat{a}_1, \hat{a}_2)$ of the cut-off form, i.e. equilibria where

$$\hat{a}_2(w) = \begin{cases} 1 & \text{if} \quad \hat{w} \leq w, \\ 0 & \text{if} \quad w < \hat{w}, \end{cases}$$

for some $\hat{w} \in [0, V]$, and $\hat{a}_1 = \hat{w}$. Such equilibria are interesting for at least reasons. First, they are materially Pareto efficient. Second, the possibility that a discontinuity emerges in an otherwise completely continuous model is of its own interest.

Proposition 8. *Consider the continuous moral hazard game with $V > 2$ and $d(e) = e^2$.*
(i) IE of the cut-off form exist if and only if

$$y \geq \delta(V) \equiv \frac{2}{(V-1)^{3/2} - (V-1)}. \tag{5}$$

(ii) If (5) is satisfied, then there exist two values w^l and w^h with $2 < w^l \leq w^h < V$ such that a cut-off profile is an IE if and only if $\hat{w} \in [w^l, w^h]$.

Proof. See Appendix, Section A.5. □

Condition (5) can be used to describe the parameter region in (V, y)-space for which cut-off equilibria exist (see Figure 1). High values of V and y (above the downward sloping line) are conducive to existence. As the size of the surplus grows (V becomes sufficiently large), small levels of reciprocity suffice to guarantee existence. Intuitively, cut-off equilibria need a sufficiently strong reciprocity effect for existence, to render credible the threat of supplying zero effort for small wages. As argued before, this is the case if either y or V are large enough.

The boundaries w^l and w^h described in statement (ii) are characterized by conditions similar to those in the binary moral hazard model (see equations (16) and (17) in Appendix A.5), and they exhibit the same properties as those described in Proposition 7. Specifically, as $y \to \infty$, they converge to V. Since $\tilde{a}_1 \leq 2$ but $2 < w^l \leq \hat{w} = \hat{a}_1$ holds, we again have that the wage in any cut-off IE is larger than the SPE wage.[19] An example for the case where $V = 3$ is given in Figure 2. According to (5), cut-off equilibria then exist whenever $y \geq 1/(\sqrt{2} - 1) \approx 2.41$. The first equilibrium that emerges (for $y = 1/(\sqrt{2} - 1)$) has the cutoff $\hat{w} = 1/(\sqrt{2} - 1)$. Several equilibria exist for larger values of y, but both the largest and the smallest equilibrium are strictly increasing, their cut-offs are larger than $\tilde{a}_1 = 1.5$, and they converge towards a situation where the agent obtains the entire project payoff.

[19] To apply Proposition 4, observe that $\Delta(w)$ is constant on $[\hat{a}_1, \overline{w}]$ in any cut-off IE, because $\tilde{a}_1(w) = \hat{a}_1(w) = 1$ in this region.

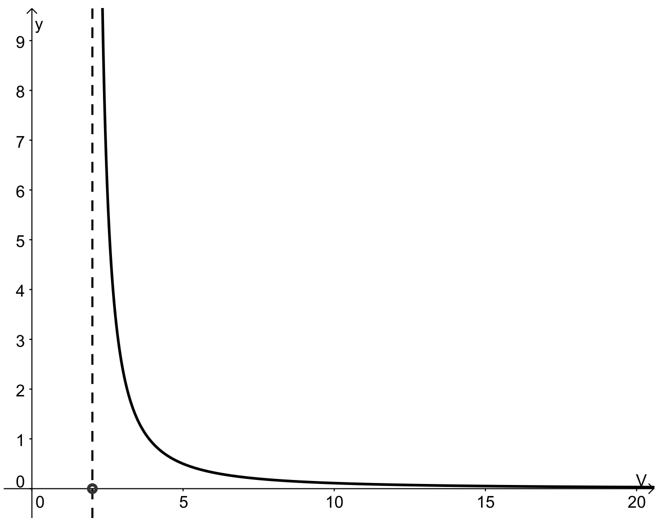

Figure 1: Existence of Cut-off Equilibria

Now whenever $V < 4$ so that $\tilde{a}_1 < 2$ and $\tilde{a}_2(\tilde{a}_1) < 1$, the effort in any cut-off IE is strictly larger than in the SPE. In fact, the efficient effort is chosen in any such IE. Note however that this requires a relatively strong concern for reciprocity. At $V = 4$ the existence condition (5) becomes $y \geq 1/(\sqrt{27} - 3) \approx 0.91$. Hence whenever the IE has the potential to outperform the SPE ($V < 4$), this still requires player 2 to put about at least equal weights on material and on psychological payoffs.

The quadratic model also has additional equilibria that are not of the cut-off form. The goal for the remainder of the section will be to illustrate such equilibria and discuss interesting properties. We should point out, however, that we do not attempt a complete equilibrium analysis, but construct equilibria as follows.[20] First, given some second-order belief c_{212}, player 2's optimization yields a unique best effort response to any offered wage. The best response of player 1 to this best-response function can then also be determined. Under the (potentially restrictive) assumption that player 2's material payoff is increasing in the wage in equilibrium, his equitable payoff can then be calculated as the average between his equilibrium payoff and the payoff obtained for wage $w = V$. Then, the fixed point condition $\hat{a}_2 = c_{212}$ can be invoked to determine IE. It can now be shown that, as y increases from 0, we pass through the following equilibrium regions (see Table 1 for the numerical results when $V = 3$).

(i) For very low values of y_2 (e.g., $y_2 = 0.1$), the reaction function of player 2 is zero up to some threshold after which it becomes positive and is strictly increasing. Eventu-

[20] The details can be found in Appendix A.6.

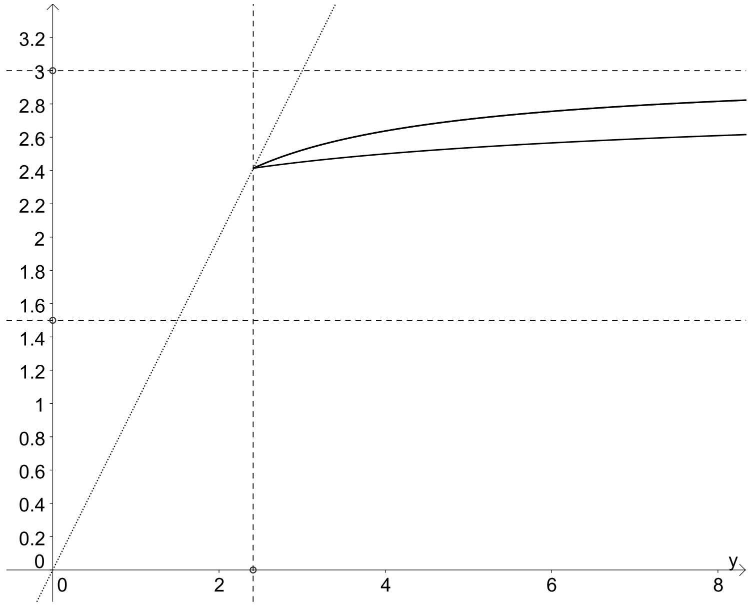

Figure 2: Cut-off Boundaries

y	0	1	1.5	2	5	7	9
Type	O	O	O	O	O+C	O+C	C
Lowest wage	1.5	2.22	2.33	2.38	2.54	2.59	2.63
Highest wage	1.5	2.22	2.33	2.38	2.71	2.79	2.84
Effort	0.75	0.95	1	1	1	1	1
π_1	1.13	0.75	0.67	0.62	0.46	0.41	0.37
π_2	0.56	1.21	1.33	1.38	1.54	1.59	1.63

Table 1: Numerical Results. C = cut-off equilibrium, O = other equilibrium.

ally, a point where the agent supplies full effort is reached. The reaction function is continuous, and the offered equilibrium wage is too low to induce full effort.

(ii) For larger values of y the equilibrium has similar properties except that, at the threshold below which the reaction function is zero, there is an upward jump to an effort level between 0 and 1. As long as y is not too large (e.g., $y_2 = 1$), player 1 will not find it optimal to induce full effort. For higher values ($y_2 \geq 1.5$), player 1 pays an equilibrium wage that induces full effort.

Equilibria of the described type cease to exist if y is larger than some critical value (between 8 and 9 if $V = 3$). Figure 3 illustrates the equilibrium strategies of player 2 for different values of y. Hence, there is a parameter region where the last type of equilibrium coexists with cut-off equilibria. Several other observations are worth mentioning:

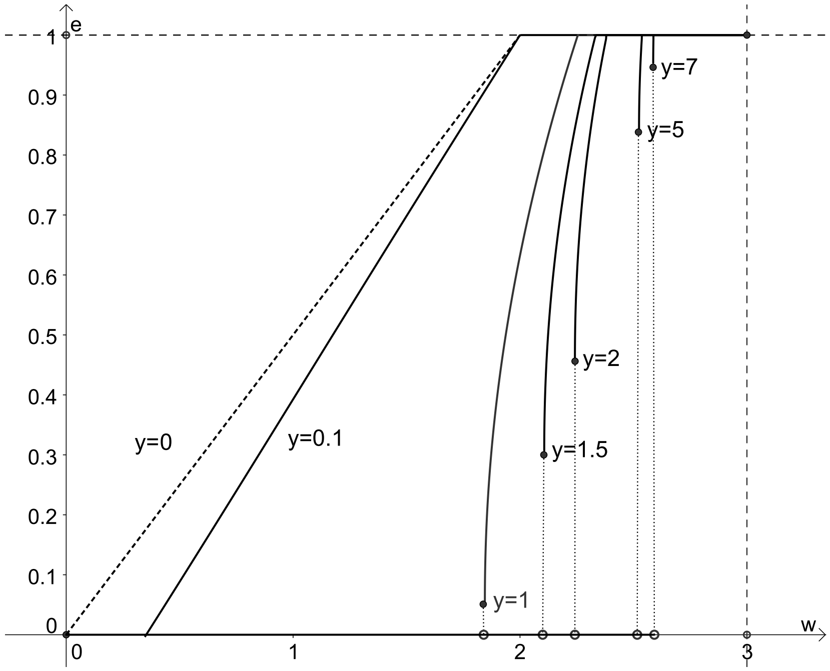

Figure 3: Other Equilibria

(i) Where cut-off equilibria co-exist with the above derived equilibrium, the wage in the latter corresponds to the lowest wage in a cut-off equilibrium.

(ii) The relation between y_2 and both the lowest and the highest equilibrium wage is monotone: As the reciprocity parameter increases, so does the equilibrium wage and the equilibrium effort.

(iii) The material payoff of player 1 decreases with reciprocity, while the material payoff of player 2 increases with reciprocity.

(iv) As y_2 approaches infinity, the equilibrium wage approaches V, so that the agent receives the entire surplus.

6.2.3 A General Limit Result

We conclude the moral hazard example by providing a general limiting result for the case when $y \to \infty$. It applies to the continuous moral hazard model for any differentiable (and strictly increasing and convex) effort cost function $d(e)$, provided that $V > d'(0)$, i.e. that some effort provision is efficient. Let $\mathscr{A}(y)$ denote the set of IE for reciprocity parameter y, and assume $\mathscr{A}(y) \neq \varnothing$ for all $y \in \mathbb{R}_+$. Then, let $a^* : \mathbb{R}_+ \to A_1 \times A_2$ be an arbitrary equilibrium selection function, i.e. $a^*(y) = (a_1^*(y), a_2^*(y)) \in \mathscr{A}(y)$ for all $y \in \mathbb{R}_+$. We denote the value of $a_2^*(y)$ at a wage w by $a_2^*(y)(w)$.

Proposition 9. *Consider the continuous moral hazard game where $d(e)$ is differentiable and $V > d'(0)$. Then any equilibrium selection function a^* satisfies* $\lim_{y\to\infty} a_1^*(y) = V$.

Proof. See Appendix, Section A.7. □

Proposition 9 establishes in general what the previous examples have illustrated: as player 2's concern for material payoffs becomes negligible relative to psychological payoffs, he eventually reaps the entire gains from trade, i.e. the equilibrium wage converges to the full project payoff V. This result is perfectly robust, in the sense that it holds for every conceivable equilibrium selection.

7 Discussion

Our model with one-sided reciprocity is not supposed to deliver a comprehensive picture of bilateral interactions. First, many interactions (e.g. among experiment participants) will naturally involve social preferences on both sides. As Rabin (1993) has already shown, positive kindness is then compatible with equilibrium. Other interactions, such as between a profit-maximizing firm and an employee, are more in line with our setup. Second, and more importantly, substantial evidence for the simultaneous presence of both intention- and outcome-based social preferences has been accumulated by now (Falk, Fehr, and Fischbacher 2003b). Still, we believe that our approach is valuable for at least two reasons.

1. We have chosen to focus on intention-based preferences for conceptual clarity. As a result, we have been able to show that existing results for purely outcome-based preferences are changed considerably. Our approach makes it possible to isolate the effects of intention-based preferences, which would be hard if not impossible in a model that combines different types of social preferences. Similarly, the role of the reference point against which kindness is measured has been clarified. Our results rest on the assumption of interior reference points. From a different perspective, our findings therefore imply that positive equilibrium kindness necessarily requires the use of a reference point that is on an extreme point of the efficiency frontier.

2. While we described π_i as player i's *material* payoff throughout the paper, a more appropriate name would actually be *outcome-based* payoff. Assuming that π_i includes nonpecuniary utility components such us altruism or envy is compatible with our model as long as they can be derived from outcomes only. Hence many of the effects that we have derived will still exist in addition to possible outcome-based effects. In particular,

this holds for the bad intentions result, which requires essentially no assumptions on the payoff functions π_i.

Utikal and Fischbacher (2009) report on an experiment where individuals were asked to evaluate the intentions behind actions of a profit-maximizing firm: one treatment involved positive and one negative externalities on a third party. They find that, when the firm is in a dominant position and positive externalities are small, positive externalities are perceived as unintentional while negative externalities are perceived as fully intentional. The effect is no longer present when the firm has small economic status. While the overall approach is not fully comparable to our model, for example because the person to judge and punish is not the one experiencing the externality, these results still confirm earlier findings according to which negative reciprocity, as in our model, seems to be the more widespread phenomenon (see Utikal and Fischbacher (2009) for a detailed discussion).

Finally, we have already emphasized in Section 3 that our definition of equitable payoff rests on an efficiency concept adopted from Rabin (1993) rather than Dufwenberg and Kirchsteiger (2004). This is indeed crucial for our results. The concept of sequential reciprocity equilibrium due to Dufwenberg and Kirchsteiger (2004) would require to define $\Pi_i^E(a_i|h)$ as the payoff pairs achievable when player $j \neq i$ can play any *efficient strategy* a_j. A strategy a_j is efficient except if there exists another strategy a'_j that always yields the same and sometimes higher payoffs (to both players), where "always" refers to all histories and all strategies of player i. With this concept, the set $\Pi_i^E(a_i|h)$ would become substantially larger and include payoff pairs that are in fact Pareto dominated when h and a_i are fixed. As a result, equilibrium kindness would become possible even with one-sided reciprocity. In the binary moral hazard game, for instance, paying a wage below the agent's threshold $\hat{w}$ would be efficient. Given the agent's *actual* strategy, the resulting outcome is Pareto inefficient. There are, however, non-equilibrium strategies of the agent for which the low wage would induce a Pareto efficient outcome, in which the principal would obtain very large payoffs. Hence the principal could be considered kind when offering $\hat{w}$, even though he knows the agent's actual equilibrium strategy and in fact does not sacrifice own payoffs in favor of the agent.

8 Conclusions

This paper has shown that, when materialistic and reciprocal players interact, both will typically display unkind behavior in equilibrium. This result, which requires only weak assumptions, stands in stark contrast to familiar findings that materialistic players ("rotten kids") benefit from outcome-based social preferences of others, such as altruism and inequity aversion. We explore the implications of our general result in standard principal-agent games. Compared to the subgame-perfect equilibrium, the intentions equilibrium usually displays higher wages and higher efforts. While the reciprocal player obtains higher material payoffs thanks to the higher wages even though he chooses higher efforts, the materialistic player loses, because the higher effort does not compensate for higher wage costs. When reciprocity concerns are sufficiently pronounced, the gains from trade may accrue fully to the reciprocal player. Thus, contrary to inequity aversion, reciprocity works towards a more unequal distribution.

Our analysis has several interesting applications. First, most directly, firms may not want to employ reciprocal workers. Even compared to the alternative of purely materialistic workers, the employers are typically worse off with reciprocal agents. Second, materialistic workers can benefit from being perceived as reciprocators, in the sense that firms who are prepared to employ them will pay higher wages than to materialistic agents. However, in view of the first conclusion they must be concerned that firms may shy away from employing them in the first place. Third, we obtain a rationale for paying higher wages to employees with higher responsibility, as measured by the difference in payoffs that workers' actions can make. Suppose the worker has a strong potential for sabotage by choosing low effort, because high effort may lead to a substantial increase in social surplus. Contrary to the corresponding model with purely materialistic agents, our model predicts a higher wage.

While these implications follow more or less directly from our model, we are also confident that suitable extensions can be used to obtain further interesting results on organizational design. For instance, experimental observations suggests that, if a principal gives the control rights for unpopular decisions to third parties, he may benefit because he is perceived as less unkind than when he takes such decisions himself (Bartling and Fischbacher 2008). It would be interesting to see whether such behavior is consistent with our framework. Finally, an extension of our paper to a model with multiple agents would seem suitable to shed new light on the longstanding debate on the boundaries of the firm: Changes in the numbers of employees working on related projects may affect their potential for sabotage and thus the potential adverse consequences of reciprocal behavior.

References

Bartling, B., and U. Fischbacher (2008): "Shifting the Blame: On Delegation and Responsibility," SSRN Discussion Paper No. 1166544.

Battigalli, P., and M. Dufwenberg (2009): "Dynamic psychological games," *Journal of Economic Theory*, 144, 1–35.

Becker, G. (1974): "A Theory of Social Interactions," *Journal of Political Economy*, 82, 1063–1093.

——— (1981): *A Treatise on the Family.* Harvard University Press, Cambridge, Massachusetts.

Benjamin, D. (2008): "Social Preferences and the Efficiency of Bilateral Exchange," mimeo, Cornell University.

Bergstrom, T. (1989): "A Fresh Look at the Rotten Kid Theorem–and Other Household Mysteries," *Journal of Political Economy*, 97, 1138–1159.

Bolton, G., and A. Ockenfels (2000): "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, 90, 166–193.

Charness, A., and M. Rabin (2002): "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics*, 117, 817–869.

Dufwenberg, M., and G. Kirchsteiger (2000): "Reciprocity and Wage Undercutting," *European Economic Review*, 44, 1069–1078.

——— (2004): "A Theory of Sequential Reciprocity," *Games and Economic Behavior*, 47, 268–298.

Dur, R., and A. Glazer (2008): "Optimal Contracts When a Worker Envies His Boss," *Journal of Law, Economics, and Organization*, 24, 120–137.

Englmaier, F., and S. Leider (2008): "Contractual and Organizational Structure with Reciprocal Agents," Cesifo working paper 2415.

Falk, A., E. Fehr, and U. Fischbacher (2003a): "On the Nature of Fair Behavior," *Economic Inquiry*, 41, 20–26.

——— (2003b): "Testing Theories of Fairness - Intentions Matter," *Games and Economic Behavior*, 62, 287–303.

Falk, A., and U. Fischbacher (2006): "A Theory of Reciprocity," *Games and Economic Behavior*, 54, 293–315.

Fehr, E., S. Gächter, and G. Kirchsteiger (1997): "Reciprocity as a Contract Enforcement Device: Experimental Evidence," *Econometrica*, 65, 833–860.

Fehr, E., and K. Schmidt (1999): "A Theory Of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114, 817–868.

Rabin, M. (1993): "Incorporating Fairness Into Game Theory and Economics," *American Economic Review*, 83, 1281–1302.

Utikal, V., and U. Fischbacher (2009): "On the Attribution of Externalities," *TWI Research Paper*, 46.

A Appendix

A.1 Proof of Proposition 1

Step 1. Consider any history h that is reached on the equilibrium path, i.e. that satisfies $\hat{a}_i^h = \hat{a}_i$ for $i = 1, 2$. We then have that $\pi_i(\hat{a}_i, \hat{a}_j|h) = \pi_i(\hat{a}_i, \hat{a}_j|\varnothing) = \pi_i(\hat{a}_i, \hat{a}_j)$, $\Pi_i^E(\hat{a}_i|h) = \Pi_i^E(\hat{a}_i|\varnothing)$ and $\pi_i^e(\hat{a}_i|h) = \pi_i^e(\hat{a}_i|\varnothing)$ for $i = 1, 2$, $i \neq j$, the latter by (A2). Therefore, the following arguments for $h = \varnothing$ apply unaltered to any history on the equilibrium path.

Step 2. We now show that $k_{12}(\hat{a}_1, \hat{a}_2|\varnothing) \leq 0$, with strict inequality if $|\Pi_2^E(\hat{a}_2|\varnothing)| \geq 2$. To obtain a contradiction, assume first that $|\Pi_2^E(\hat{a}_2|\varnothing)| \geq 2$ but $k_{12}(\hat{a}_1, \hat{a}_2|\varnothing) \geq 0$. There are now two subcases. First, $(\pi_1(\hat{a}_1, \hat{a}_2), \pi_2(\hat{a}_2, \hat{a}_1)) \in \Pi_2^E(\hat{a}_2|\varnothing)$ may hold. Then, under (A1), there exists a strategy $a_1' \in A_1$ such that $(\pi_1(a_1', \hat{a}_2), \pi_2(\hat{a}_2, a_1')) \in \Pi_2^E(\hat{a}_2|\varnothing)$ and $\pi_2(\hat{a}_2, a_1') < \pi_2^e(\hat{a}_2|\varnothing) \leq \pi_2(\hat{a}_2, \hat{a}_1)$, where the second inequality follows from the assumption that $k_{12}(\hat{a}_1, \hat{a}_2|\varnothing) \geq 0$. Pareto efficiency of the elements in $\Pi_2^E(\hat{a}_2|\varnothing)$ then implies that $\pi_1(a_1', \hat{a}_2) > \pi_1(\hat{a}_1, \hat{a}_2)$, which contradicts that $\hat{a}_1 \in \arg\max_{a_1 \in A_1} \pi_1(a_1, \hat{a}_2)$. Second, $(\pi_1(\hat{a}_1, \hat{a}_2), \pi_2(\hat{a}_2, \hat{a}_1)) \notin \Pi_2^E(\hat{a}_2|\varnothing)$ may hold. Then, all $(\pi_1', \pi_2') \in \Pi_2^E(\hat{a}_2|\varnothing)$ must satisfy $\pi_2' > \pi_2(\hat{a}_2, \hat{a}_1)$. Assume to the contrary that $\pi_2' \leq \pi_2(\hat{a}_2, \hat{a}_1)$ for some $(\pi_1', \pi_2') \in \Pi_2^E(\hat{a}_2|\varnothing)$. Since $\hat{a}_1 \in \arg\max_{a_1 \in A_1} \pi_1(a_1, \hat{a}_2)$, $\pi_1' \leq \pi_1(\hat{a}_1, \hat{a}_2)$ must also hold. But then, if $(\pi_1', \pi_2') \in \Pi_2^E(\hat{a}_2|\varnothing)$, $(\pi_1(\hat{a}_1, \hat{a}_2), \pi_2(\hat{a}_2, \hat{a}_1)) \in \Pi_2^E(\hat{a}_2|\varnothing)$ must also hold, a contradiction. Now we immediately obtain $\pi_2(\hat{a}_2, \hat{a}_1) < \pi_2^e(\hat{a}_2|\varnothing)$ under (A1), and hence $k_{12}(\hat{a}_1, \hat{a}_2|\varnothing) < 0$.

Assume then that $|\Pi_2^E(\hat{a}_2|\varnothing)| = 1$ and denote $\Pi_2^E(\hat{a}_2|\varnothing) = \{(\pi_1', \pi_2')\}$, so that $\pi_2^e(\hat{a}_2|\varnothing) = \pi_2'$ under (A1). Then, unique Pareto efficiency of (π_1', π_2') implies that $\pi_2(\hat{a}_2, \hat{a}_1) \leq \pi_2'$, which in turn implies $k_{12}(\hat{a}_1, \hat{a}_2|\varnothing) \leq 0$. This completes step 2.

Step 3. Now consider player 2. In any IE, we have that $b_{21} = \hat{a}_1$ and $c_{212} = \hat{a}_2$ according to Definition 1, so that $\lambda_{212}(b_{21}, c_{212}|h) = k_{12}(\hat{a}_1, \hat{a}_2|h) \leq 0$ at any history $h \in H$ on the equilibrium path, including $h = \varnothing$. Hence $U_2(a_2, \hat{a}_1, c_{212}|\varnothing)$ is (weakly) decreasing in $\pi_1(\hat{a}_1, a_2|\varnothing)$. We can now go through the same cases as for player 1, repeating analogous arguments. First, if

$|\Pi_1^E(\hat{a}_1|\varnothing)| \geq 2$ and $(\pi_1(\hat{a}_1,\hat{a}_2),\pi_2(\hat{a}_2,\hat{a}_1)) \in \Pi_1^E(\hat{a}_1|\varnothing)$, we obtain a contradiction to weakly positive kindness because $\exists a_2' \in A_2$ such that $\pi_1(\hat{a}_1,a_2') < \pi_1(\hat{a}_1,\hat{a}_2)$ and $\pi_2(a_2',\hat{a}_1) > \pi_2(\hat{a}_2,\hat{a}_1)$. If $(\pi_1(\hat{a}_1,\hat{a}_2),\pi_2(\hat{a}_2,\hat{a}_1)) \notin \Pi_1^E(\hat{a}_1|\varnothing)$, on the other hand, all profiles $(\pi_1',\pi_2') \in \Pi_1^E(\hat{a}_1|\varnothing)$ must satisfy $\pi_1' > \pi_1(\hat{a}_1,\hat{a}_2)$. Otherwise, if $\pi_1' \leq \pi_1(\hat{a}_1,\hat{a}_2)$ held for some $(\pi_1',\pi_2') \in \Pi_1^E(\hat{a}_1|\varnothing)$, $\pi_2' > \pi_2(\hat{a}_1,\hat{a}_2)$ would have to be true, because (π_1',π_2') is Pareto efficient but $(\pi_1(\hat{a}_1,\hat{a}_2),\pi_2(\hat{a}_2,\hat{a}_1))$ is not. Then, player 2 would prefer the strategy inducing (π_1',π_2') over $\hat{a}_2$. But then again $\pi_1(\hat{a}_1,\hat{a}_2) < \pi_1^e(\hat{a}_1|\varnothing)$. Finally, if $|\Pi_1^E(\hat{a}_1|\varnothing)| = 1$, it again follows immediately that $k_{21}(\hat{a}_2,\hat{a}_1|\varnothing) \leq 0$.

A.2 Proof of Proposition 4

Step 1. First we show that $\hat{a}_2(\tilde{a}_1) \leq \tilde{a}_2(\tilde{a}_1)$ must hold. To obtain a contradiction, assume $\hat{a}_2(\tilde{a}_1) > \tilde{a}_2(\tilde{a}_1)$. Then, $\pi_1(\tilde{a}_1,\tilde{a}_2) = \tilde{\pi}_1(\tilde{a}_1,\tilde{a}_2(\tilde{a}_1)) < \tilde{\pi}_1(\tilde{a}_1,\hat{a}_2(\tilde{a}_1))$ under (A3)(i). Since $\hat{a}_1 \in \arg\max_{w\in W} \tilde{\pi}_1(w,\hat{a}_2(w))$ according to Lemma 1, $\tilde{\pi}_1(\tilde{a}_1,\hat{a}_2(\tilde{a}_1)) \leq \tilde{\pi}_1(\hat{a}_1,\hat{a}_2(\hat{a}_1)) = \pi_1(\hat{a}_1,\hat{a}_2)$ must hold, which implies $\pi_1(\tilde{a}_1,\tilde{a}_2) < \pi_1(\hat{a}_1,\hat{a}_2)$ and contradicts Proposition 2.

Step 2. To prove the proposition, we are going to show that, for any $w \in [\underline{w},\tilde{a}_1)$, it holds that $\tilde{\pi}_1(\tilde{a}_1,\tilde{a}_2(\tilde{a}_1)) - \tilde{\pi}_1(w,\tilde{a}_2(w)) \leq \tilde{\pi}_1(\tilde{a}_1,\hat{a}_2(\tilde{a}_1)) - \tilde{\pi}_1(w,\hat{a}_2(w))$. The LHS of this inequality is strictly positive by definition of $\tilde{a}_1$ as the unique maximizer of $\tilde{\pi}_1(w,\tilde{a}_2(w))$. Then, if the inequality holds, the RHS must also be strictly positive. But $\tilde{\pi}_1(\tilde{a}_1,\hat{a}_2(\tilde{a}_1)) - \tilde{\pi}_1(\hat{a}_1,\hat{a}_2(\hat{a}_1))$ cannot be strictly positive, again because $\hat{a}_1$ maximizes $\tilde{\pi}_1(w,\hat{a}_2(w))$. We then know that $\hat{a}_1 \notin [\underline{w},\tilde{a}_1)$, which is the desired conclusion.

The above inequality can be rearranged to $\tilde{\pi}_1(\tilde{a}_1,\tilde{a}_2(\tilde{a}_1)) - \tilde{\pi}_1(\tilde{a}_1,\hat{a}_2(\tilde{a}_1)) \leq \tilde{\pi}_1(w,\tilde{a}_2(w)) - \tilde{\pi}_1(w,\hat{a}_2(w))$. Now, $\tilde{\pi}_1(\tilde{a}_1,\tilde{a}_2(\tilde{a}_1)) - \tilde{\pi}_1(\tilde{a}_1,\hat{a}_2(\tilde{a}_1)) \leq \tilde{\pi}_1(w,\tilde{a}_2(\tilde{a}_1)) - \tilde{\pi}_1(w,\hat{a}_2(\tilde{a}_1))$ holds due to $w < \tilde{a}_1$, $\hat{a}_2(\tilde{a}_1) \leq \tilde{a}_2(\tilde{a}_1)$ from step 1, and submodularity of $\tilde{\pi}_1$ (assumption (A4)(i)). Supermodularity of $\tilde{\pi}_2$ (assumption (A4)(ii)) implies that $\tilde{a}_2(w) \leq \tilde{a}_2(\tilde{a}_1)$. Then, concavity of $\tilde{\pi}_1$ in e (assumption (A4)(iii)) implies that $\tilde{\pi}_1(w,\tilde{a}_2(\tilde{a}_1)) - \tilde{\pi}_1(w,\hat{a}_2(\tilde{a}_1)) \leq \tilde{\pi}_1(w,\tilde{a}_2(w)) - \tilde{\pi}_1(w,\hat{a}_2(\tilde{a}_1) - \tilde{a}_2(\tilde{a}_1) + \tilde{a}_2(w))$. Now observe that $\hat{a}_2(w) \leq \hat{a}_2(\tilde{a}_1) - \tilde{a}_2(\tilde{a}_1) + \tilde{a}_2(w)$, which follows immediately from the fact that $\Delta(w)$ is decreasing in $w \in [\underline{w},\tilde{a}_1]$. Thus $\tilde{\pi}_1(w,\tilde{a}_2(w)) - \tilde{\pi}_1(w,\hat{a}_2(\tilde{a}_1) - \tilde{a}_2(\tilde{a}_1) + \tilde{a}_2(w)) \leq \tilde{\pi}_1(w,\tilde{a}_2(w)) - \tilde{\pi}_1(w,\hat{a}_2(w))$. Combining all these inequalities yields $\tilde{\pi}_1(\tilde{a}_1,\tilde{a}_2(\tilde{a}_1)) - \tilde{\pi}_1(\tilde{a}_1,\hat{a}_2(\tilde{a}_1)) \leq \tilde{\pi}_1(w,\tilde{a}_2(w)) - \tilde{\pi}_1(w,\hat{a}_2(w))$, which is the desired result.

Step 3. The proof for the interval $[\hat{a}_1,\overline{w}]$ is analogous.

A.3 Proof of Proposition 6

Statement (i). Fix any IE $(\hat{a}_1,\hat{a}_2)$.

Step 1. We will first derive the kindness terms $k_{21} = \tilde{\pi}_1(w,e) - \pi_1^e(w)$ for all $w \in [0,V]$. For $w < d/p$ it holds that $\Pi_1^E(w) = \{(0,0),(p(V-w),pw-d)\}$, because $pw - d < 0 < p(V-w)$. This implies $\pi_1^e(w) = (p/2)(V-w)$ and

$$k_{21}(e,w) = \begin{cases} (p/2)(V-w) & \text{if} \quad e = p, \\ -(p/2)(V-w) & \text{if} \quad e = 0. \end{cases}$$

For $d/p \leq w$ we have $\Pi_1^E(w) = \{(p(V-w), pw-d)\}$. This implies $\pi_1^e(w) = p(V-w)$ and

$$k_{21}(e,w) = \begin{cases} 0 & \text{if} \quad e = p, \\ -p(V-w) & \text{if} \quad e = 0. \end{cases}$$

Step 2. We can now show that $\hat{a}_2(V) = p$ must hold. Indeed, from the above results we have $k_{21}(e,V) = 0$ for both $e \in \{0,p\}$. Hence $\hat{a}_2(V) = \tilde{a}_2(V) = p$ according to Lemma 1.

Step 3. Next, we show that $\exists w' < V$ with $\hat{a}_2(w') = p$. To obtain a contradiction, assume $\hat{a}_2(w) = 0$ for all $w \in [0,V)$. Then $\Pi_2^E(\hat{a}_2) = \{(pV-d, 0)\}$, $\pi_2^e(\hat{a}_2) = pV-d$ and

$$\lambda_{212}(w,\hat{a}_2) = \begin{cases} 0 & \text{if} \quad w = V, \\ -(pV-d) & \text{if} \quad w < V. \end{cases}$$

Consider any wage $w \in [d/p, V)$. According to Lemma 1, optimality of effort $\hat{a}_2(w) = 0$ requires $\tilde{\pi}_2(0,w) + yk_{21}(0,w)\lambda_{212}(w,\hat{a}_2) \geq \tilde{\pi}_2(p,w) + yk_{21}(p,w)\lambda_{212}(w,\hat{a}_2)$. Using the above derived terms this can be rearranged to

$$yp(V-w) \geq \frac{pw-d}{pV-d}.$$

But as $w \to V$, the LHS of this condition goes to zero while the RHS goes to 1, so it must be violated for sufficiently large wages $w < V$, which precludes the possibility of $(\hat{a}_1, \hat{a}_2)$ being an IE.

Step 4. Given $\hat{a}_2$, let $W_0 = \{w \in W | \hat{a}_2(w) = 0\}$ and $W_p = \{w \in W | \hat{a}_2(w) = p\}$. We know from step 3 that $|W_p| \geq 2$. Since any $w \in W_0$ induces material payoffs $\tilde{\pi}_1(w, \hat{a}_2(w)) = \tilde{\pi}_2(\hat{a}_2(w), w) = 0$, while for $w = V$ we have $\tilde{\pi}_1(V, \hat{a}_2(V)) = 0$ and $\tilde{\pi}_2(\hat{a}_2(V), V) = pV - d > 0$ according to step 2, wages $w \in W_0$ induce Pareto inefficient payoffs. For wages $w \in W_p$, $\tilde{\pi}_1(w, \hat{a}_2(w)) = p(V-w)$ is strictly decreasing and $\tilde{\pi}_2(\hat{a}_2(w), w) = pw - d$ is strictly increasing in w, which implies $\Pi_2^E(\hat{a}_2) = W_p$. From Lemma 1 we can also conclude that $\hat{a}_1 = \min W_p$, so that $\hat{a}_1 < V$ and $\hat{a}_2(\hat{a}_1) = p$. Player 2's equitable payoff can then be expressed as $\pi_2^e(\hat{a}_2) = (p/2)(V + \hat{a}_1) - d$, because $w = \hat{a}_1$ yields his smallest and $w = V$ yields his largest among Pareto efficient payoffs. As for the kindness term λ_{212}, we obtain

$$\lambda_{212}(w,\hat{a}_2) = \begin{cases} pw - (p/2)(V + \hat{a}_1) & \text{if} \quad w \in W_p, \\ d - (p/2)(V + \hat{a}_1) & \text{if} \quad w \in W_0. \end{cases}$$

Step 5. Since $\hat{a}_1 \in W_p$ we have that $\lambda_{212}(\hat{a}_1, \hat{a}_2) = (p/2)(\hat{a}_1 - V) < 0$. This implies that $d/p \leq \hat{a}_1$ must be true, since otherwise, $\hat{a}_2(\hat{a}_1) = p \neq 0 = \arg\max_{e \in \{0,p\}} \tilde{\pi}_2(e, \hat{a}_1) + yk_{21}(e, \hat{a}_1)\lambda_{212}(\hat{a}_1, \hat{a}_2)$, for both material and reciprocity motives. Since $\hat{a}_1 = \min W_p$, we obtain $d/p \leq w$ for all $w \in W_p$ as a corollary.

Step 6. For $(\hat{a}_1, \hat{a}_2)$ to be an IE it needs to hold that $\tilde{\pi}_2(p,w) + yk_{21}(p,w)\lambda_{212}(w,\hat{a}_2) \geq \tilde{\pi}_2(0,w) + yk_{21}(0,w)\lambda_{212}(w,\hat{a}_2)$ for all $w \in W_p$. Using the above derived expressions for the case $d/p \leq w$, this can be simplified to

$$y\left[(p/2)(V + \hat{a}_1) - pw\right] \leq \frac{pw - d}{p(V - w)}. \tag{6}$$

Since the LHS of (6) is strictly decreasing in w and the RHS is strictly increasing, this is satisfied $\forall w \in W_p$ iff it is satisfied at $\hat{a}_1 = \min W_p$:

$$y\left[(p/2)(V - \hat{a}_1)\right] \leq \frac{p\hat{a}_1 - d}{p(V - \hat{a}_1)}. \tag{7}$$

Since the LHS of (7) is decreasing and the RHS is increasing in $\hat{a}_1$, condition (7) implicitly defines a lower bound w^l for $\hat{a}_1$ given by

$$y\left[(p/2)(V - w^l)\right] = \frac{pw^l - d}{p(V - w^l)}, \tag{8}$$

such that $(\hat{a}_1, \hat{a}_2)$ can be an IE only if $w^l \leq \hat{a}_1$. To see that $d/p < w^l$, observe that the RHS of (8) becomes zero if $w^l = d/p$, while the LHS is still positive. As the LHS is decreasing and the RHS is increasing in w^l we must have $d/p < w^l$.

Step 7. For $(\hat{a}_1, \hat{a}_2)$ to be an IE, it also needs to hold that $\tilde{\pi}_2(0,w) + yk_{21}(0,w)\lambda_{212}(w,\hat{a}_2) \geq \tilde{\pi}_2(p,w) + yk_{21}(p,w)\lambda_{212}(w,\hat{a}_2)$ for all $w \in W_0$. For $w \in W_0$ with $d/p \leq w < \hat{a}_1$ (which exist since $d/p < w^l \leq \hat{a}_1$), this can be rearranged to

$$y\left[(p/2)(V + \hat{a}_1) - d\right] \geq \frac{pw - d}{p(V - w)}.$$

Since the RHS is continuous and increasing in w, this is satisfied for all $d/p \leq w < \hat{a}_1$ iff it is satisfied at $\hat{a}_1$, i.e.

$$y\left[(p/2)(V + \hat{a}_1) - d\right] \geq \frac{p\hat{a}_1 - d}{p(V - \hat{a}_1)}. \tag{9}$$

Both the LHS and the RHS of (9) are increasing in $\hat{a}_1$. We can calculate the derivative of the LHS w.r.t. $\hat{a}_1$ as $(p/2)y$ and the derivative of the RHS as $(V - d/p)/(V - \hat{a}_1)^2$. The condition that $(p/2)y < (V - d/p)/(V - \hat{a}_1)^2$ can be rearranged to

$$y\left[(p/2)(V - \hat{a}_1)\right] < \frac{pV - d}{p(V - \hat{a}_1)},$$

which is satisfied whenever (7) is satisfied. Thus (9) defines the upper bound w^h by

$$y\left[(p/2)(V + w^h) - d\right] = \frac{pw^h - d}{p(V - w^h)}, \tag{10}$$

and $(\hat{a}_1, \hat{a}_2)$ is an IE only if $\hat{a}_1 \leq w^h$. Arguments analogous to step 6 reveal that $d/p < w^h < V$.

To show that $w^l < w^h$, it suffices to show that

$$y\left[(p/2)(V - w^h)\right] < y\left[(p/2)(V + w^h) - d\right]. \tag{11}$$

If (11) holds and w^h solves (10), then for $w^l = w^h$ the LHS of (8) is smaller than the RHS of (8), which implies the desired result $w^l < w^h$. But inequality (11) is equivalent to $d/p < w^h$, which is true.

Statement (ii). For any $\hat{w} \in [w^l, w^h]$ we construct an IE $(\hat{a}_1, \hat{a}_2)$ with $\hat{a}_1 = \hat{w}$ and the cut-off strategy

$$\hat{a}_2(w) = \begin{cases} p & \text{if} \quad \hat{w} \le w, \\ 0 & \text{if} \quad w < \hat{w} \end{cases}$$

for player 2, so that $\hat{a}_2(\hat{a}_1) = \hat{a}_2(\hat{w}) = p$.

Step 1. Given $\hat{a}_2$, the fact that $\hat{w}$ (uniquely) maximizes $\tilde{\pi}_1(w, \hat{a}_2(w))$ as required by Lemma 1 is immediate.

Step 2. Since $w^l \le \hat{w}$, the argument in step 6 for statement (i) immediately implies that $\hat{a}_2(w) = p$ is indeed a best-response for player 2 to any $w \in W_p = [\hat{w}, V]$.

Step 3. Analogously, the argument in step 7 above implies that $\hat{a}_2(w) = 0$ is a best response for player 2 to any $w \in [d/p, \hat{w})$. Since $W_0 = [0, \hat{w})$, it only remains to be checked that $\hat{a}_2(w) = 0$ is also a best response to wages $w \in [0, d/p)$. But the corresponding payoff comparison can easily be rearranged to $(pw - d)/2 \le y$ which is always satisfied because $pw - d < 0$ in that range. Thus whenever $\hat{w} \in [w^l, w^h]$, the above defined cut-off profile $(\hat{a}_1, \hat{a}_2)$ is an IE.

A.4 Proof of Proposition 7

Step 1. Consider w^l first. Equation (3) can be rearranged to

$$w^l - (d/p) - (yp/2)(V - w^l)^2 = 0. \tag{12}$$

Implicit differentiation of (12) yields

$$\frac{\partial w^l}{\partial y} = \frac{(p/2)(V - w^l)^2}{1 + yp(V - w^l)} > 0$$

due to $w^l < V$. Next,

$$\frac{\partial w^l}{\partial V} = \frac{yp(V - w^l)}{1 + yp(V - w^l)} > 0,$$

and

$$\frac{\partial w^l}{\partial d} = \frac{1/p}{1 + yp(V - w^l)} > 0.$$

Step 2. As for w^h, (4) can be rearranged to

$$w^h - (d/p) - (yp/2)(V + w^h)(V - w^h) + yd(V - w^h) = 0. \tag{13}$$

From (13) we obtain

$$\frac{\partial w^h}{\partial y} = \frac{(V - w^h)\left[(p/2)(V + w^h) - d\right]}{1 + y(pw^h - d)} > 0$$

because $d/p < w^h < V$. Next,

$$\frac{\partial w^h}{\partial V} = \frac{y(pV - d)}{1 + y(pw^h - d)} > 0$$

and

$$\frac{\partial w^h}{\partial d} = \frac{(1/p) - y(V - w^h)}{1 + y(pw^h - d)}. \tag{14}$$

The condition that (14) is strictly positive can be rearranged to $y < (1/p(V - w^h))$ and, by multiplying both sides with $pw^h - d$, to $y(pw^h - d) < (pw^h - d)/(p(V - w^h))$. In this expression, the RHS $(pw^h - d)/(p(V - w^h))$ is equal to the RHS of (4), but the LHS $y(pw^h - d)$ is strictly smaller than the LHS of (4), because $w^h < V$. Since (4) holds as an equality, we can conclude $\partial w^h/\partial d > 0$.
Step 3. Rewrite (12) as

$$y\left(\frac{p}{2}\right) = \frac{w^l - d/p}{(V - w^l)^2}. \tag{15}$$

As $y \to \infty$, the LHS of (15) goes to infinity, and so must the RHS, immediately implying that $w^l \to V$ because $d/p < w^l < V$. Since $w^l < w^h < V$, the same limit statement must hold for w^h.

A.5 Proof of Proposition 8

Step 1. As in the proof of Proposition 6, we first derive the kindness terms $k_{21}(e, w)$. Since $\tilde{a}_2(w) = \min\{w/2, 1\}$ maximizes player 2's material payoffs $\tilde{\pi}_2(e, w) = ew - e^2$, which are strictly concave in e, we obtain $\Pi_1^E(w) = \left\{\left(e(V - w), ew - e^2\right) | e \in [\min\{w/2, 1\}, 1]\right\}$. The equitable payoff when $w \in [0, 2]$ is thus $\pi_1^e(w) = (1/2)(V - w)((w/2) + 1)$, and $\pi_1^e(w) = V - w$ whenever $w \in (2, V]$. We then obtain the kindness term

$$k_{21}(e, w) = \begin{cases} (V - w)(e - 1) & \text{if} \quad w \in (2, V], \\ (V - w)\left(e - \frac{(w/2)+1}{2}\right) & \text{if} \quad w \in [0, 2]. \end{cases}$$

Step 2. Lemma 1 now implies that, in any IE $(\hat{a}_1, \hat{a}_2)$, overall utility $ew - e^2 + yk_{21}(e, w)\lambda_{212}(w, \hat{a}_2)$ must be maximized by $e = \hat{a}_2(w)$ for every $w \in [0, V]$. It is easily verified that the objective is strictly concave in e (for any fixed w). The first-order condition is identical for the cases $w \in [0, 2]$ and $w \in (2, V]$ and characterizes the following effort level:

$$e^*(w) = \frac{w}{2} + \frac{y\lambda_{212}(w, \hat{a}_2)(V - w)}{2}.$$

Concavity implies that $\hat{a}_2(w) = e^*(w)$ whenever $e^*(w) \in [0,1]$, and $\hat{a}_2(w) = 1\,(= 0)$ whenever $e^*(w) > 1\,(< 0)$.

Step 3. Now consider a cut-off profile $(\hat{a}_1, \hat{a}_2)$ with cut-off value $\hat{w}$. Arguing as for Proposition 6, we obtain $\Pi_2^E(\hat{a}_2) = \{(w-1, V-w) | \hat{w} \leq w\}$ and $\pi_2^e(\hat{a}_2) = ((V+\hat{w})/2) - 1$. This implies

$$\lambda_{212}(w, \hat{a}_2) = \begin{cases} w - \left(\frac{V+\hat{w}}{2}\right) & \text{if} \quad \hat{w} \leq w, \\ 1 - \left(\frac{V+\hat{w}}{2}\right) & \text{if} \quad w < \hat{w}. \end{cases}$$

Optimality of $\hat{a}_2(w) = 1$ for all $\hat{w} \leq w$ now requires $e^*(w) \geq 1$ for all those wages, i.e.

$$y\left[\left(\frac{V+\hat{w}}{2}\right) - w\right] \leq \frac{w-2}{V-w}$$

after substitution of λ_{212} and some rearrangements. Arguing as for Proposition 6, this yields a lower bound w^l for $\hat{w}$, implicitly defined by

$$y\left[\left(\frac{V-w^l}{2}\right)\right] = \frac{w^l-2}{V-w^l}. \tag{16}$$

Analogously, the condition for $\hat{a}_2(w) = 0$ to be optimal for all $w \in [0, \hat{w})$ yields an upper bound given by

$$y\left[\left(\frac{V+w^h}{2}\right) - 1\right] = \frac{w^h}{V-w^h}. \tag{17}$$

Step 5. The fact that $2 < w^l < V$ and $w^h < V$ is shown as for Proposition 6. Thus is remains to be shown under which conditions $w^l \leq w^h$ holds, so that the requirements for equilibrium existence can be met simultaneously.

Fix w^h as defined in (17) and suppose we evaluate (16) at the value w^h instead of w^l. Then the LHS of (16), $(y/2)(V-w^h)$, is (weakly) smaller than the RHS, $(w^h-2)/(V-w^h)$, if and only if $w^l \leq w^h$. Dividing the LHS of (17) by $(y/2)(V-w^h)$ and the RHS by $(w^h-2)/(V-w^h)$ we then obtain the equivalent condition

$$\frac{V+w^h-2}{V-w^h} \leq \frac{w^h}{w^h-2},$$

which in turn is equivalent to $1+\sqrt{V-1} \leq w^h$. Hence $w^l \leq w^h$ if and only if $1+\sqrt{V-1} \leq w^h$.

Suppose first that $w^l \leq w^h$, so that $1+\sqrt{V-1} \leq w^h$. (17) can be rearranged to

$$y = \frac{2w^h}{(V-w^h)(V+w^h-2)}. \tag{18}$$

Since the RHS of this expression is increasing in w^h for values $1 < w^h$, we can replace w^h by $1+\sqrt{V-1}$ to obtain the inequality

$$y \geq \frac{2\left[1+\sqrt{V-1}\right]}{\left[(V-1)-\sqrt{V-1}\right]\left[(V-1)+\sqrt{V-1}\right]} = \frac{2}{(V-1)\left[\sqrt{V-1}-1\right]},$$

which is condition (5). Conversely, if (5) is satisfied, then $1+\sqrt{V-1} \leq w^h$ must hold, because the RHS of (18) is too small at $w^h = 1+\sqrt{V-1}$, but increasing. This implies $w^l \leq w^h$.

A.6 Numerical Example

We search for equilibria that exhibit the properties that (i) the equilibrium effort is interior, $\hat{a}_2(\hat{a}_1) \in (0,1)$, and (ii) the monetary payoff of player 2 is increasing in wages.

Consider player 1. If he induces an interior effort level, he chooses the wage according to

$$\hat{a}_2'(w)(V-w) - \hat{a}_2(w) = 0, \tag{19}$$

provided $\hat{a}_2(w)$ is differentiable in the relevant region. Assuming that a solution of this equation exists, denote it as w_m and write $e_m = \hat{a}_2(w_m)$. If the monetary payoff of player 2 is increasing in the wage, w_m leads to the minimum payoff for player 2 among the Pareto-efficient choices given $\hat{a}_2(w)$: Any lower wage will make both players worse off, and any higher wage will make player 1 worse off. Also, $\hat{a}_2(V) = 1$ must always hold, because $k_{21}(e,V)$ as derived in the proof of Proposition 8 is independent of e, so that $\hat{a}_2(V) = \tilde{a}_2(V)$. We thus obtain

$$\pi_2^e(\hat{a}_2) = \frac{e_m w_m - (e_m)^2 + V - 1}{2}. \tag{20}$$

Using the kindness terms derived in the proof of Proposition 8, we obtain the payoff of player 2 as

$$U_2(e,w,\hat{a}_2) = \begin{cases} ew - e^2 + y(V-w)(e-1)\left(\hat{a}_2(w)w - (\hat{a}_2(w))^2 - \pi_2^e(\hat{a}_2)\right) & \text{if} \quad w \in (2,V] \\ ew - e^2 + y(V-w)\left(e - \frac{(w/2)+1}{2}\right)\left(\hat{a}_2(w)w - (\hat{a}_2(w))^2 - \pi_2^e(\hat{a}_2)\right) & \text{if} \quad w \in [0,2] \end{cases}. \tag{21}$$

The function $U_2(e,w,\hat{a}_2)$ is strictly concave in e and yields the first order condition

$$\frac{\partial U_2}{\partial e} = w - 2e + y(V-w)\left(\hat{a}_2(w)w - (\hat{a}_2(w))^2 - \pi_2^e(\hat{a}_2)\right) = 0.$$

Using (20) and the fixed point condition $e = \hat{a}_2(w)$, we obtain

$$w - 2\hat{a}_2(w) + y(V-w)\left(\hat{a}_2(w)w - (\hat{a}_2(w))^2 - \left(\frac{e_m w_m - (e_m)^2 + V - 1}{2}\right)\right) = 0. \tag{22}$$

Then, by implicit differentiation we can determined the slope of $\hat{a}_2(w)$ as

$$\hat{a}_2'(w) = -\frac{y(\hat{a}_2(w))^2 - \frac{1}{2}y - \frac{1}{2}y(e_m)^2 + \frac{1}{2}Vy + Vy\hat{a}_2(w) - 2yw\hat{a}_2(w) + \frac{1}{2}yw_m e_m + 1}{Vyw - yw^2 - 2Vy\hat{a}_2(w) + 2yw\hat{a}_2(w) - 2}. \tag{23}$$

If equilibrium wages and efforts are w_m and e_m, then the following requirements need to be fulfilled. First, (19) has to hold with $\hat{a}_2'(w)$ replaced by the right-hand side of (23) evaluated at

$w = w_m$ and $\hat{a}_2(w^m) = e_m$. Second, (22) has to hold for w_m and e_m. These two requirements give

$$V - w_m - 2e_m - yw_m - 6y(e_m)^2 + Vy + \frac{3}{2}yw_m(e_m)^2 + \frac{1}{2}yw_m e_m + 3Vye_m + \frac{1}{2}Vy(e_m)^2 - \frac{3}{2}Vyw_m e_m = 0. \tag{24}$$

and

$$w_m - 2e_m + y(V - w_m)\left(e_m w_m - (e_m)^2 - \left(\frac{e_m w_m - (e_m)^2 + 2}{2}\right)\right) = 0.$$

The last two equations can be used to calculate candidate equilibrium actions for given parameter values, provided an interior solution exists. It remains to be checked that, along the reaction curve, the equilibrium payoff of player 2 increases; but this turns out to be true in the examples.

To derive $\hat{a}_2$ at wage levels other than w_m, one has to start from (22). Except for small values of y, the curve described by this equation has no solution for low values of w. Thus a boundary solution emerges. Because $\partial U_2/\partial e < 0$ and $\partial^2 U_2/\partial e^2 < 0$, the boundary solution is $\hat{a}_2(w) = 0$. To the immediate right of the critical wage level where a solution of (22) emerges, there is a region where it is solved by two effort levels, only one of which is an original best response. At the critical level, the reaction curve jumps upwards to this effort level. For higher wage levels, (22) yields a solution larger than 1. At such wage levels, $\partial U_2/\partial e > 0$ and $\partial^2 U_2/\partial e^2 < 0$ imply $\hat{a}_2(w) = 1$.

For $y > 1.5$ it turns out that the solution derived in the above fashion is no longer in the strategy space, because it would involve $e > 1$. Then one has to modify the above procedure by taking the candidate effort level $e = 1$. Choosing w as the solution of (22) for $e = e_m = 1$, one arrives at a candidate equilibrium. Going through similar procedures as above one obtains a reaction function with similar properties up to parameter values including $y = 8$. For higher parameter values, e.g. $y = 9$, the upward-sloping part of (22) no longer lies in the strategy space, so that solutions of the type described here no longer exist.

A.7 Proof of Proposition 9

Step 1. To obtain a contradiction, assume the limit statement is not true, i.e. $\exists \epsilon > 0$ such that $\forall y \in \mathbb{R}_+$, $\exists y' > y$ with $a_1^*(y') \leq V - \epsilon$. This includes the possibility that a_1^* converges to a value other than V and that a_1^* does not converge. We keep ϵ with the above property fixed for the rest of the proof.

Step 2. We then claim that $\exists \bar{y}$ such that, for all $y > \bar{y}$, $a_2^*(y)(a_1^*(y)) = 0$ whenever $a_1^*(y) \leq V - \epsilon$. To prove this claim, observe that $e = a_2^*(y)(w)$ maximizes $U_2(e, w) = ew - d(e) + yk_{21}(e, w)\lambda_{212}(w, a_2^*(y))$ according to Lemma 1, for any $w \in [0, V]$. Since $k_{21}(e, V) = \tilde{\pi}_1(V, e) - \pi_1^e(V) = -\pi_1^e(V)$ is independent of e, we must have $a_2^*(y)(V) = \tilde{a}_2(V)$. Then, since $\tilde{\pi}_2(e, w)$ is strictly increasing in w, we obtain $\pi_2^h(a_2^*(y)) = \tilde{\pi}_2(\tilde{a}_2(V), V)$. We also must have $\pi_2^l(a_2^*(y)) \geq$

$\tilde{\pi}_2(a_2^*(y)(a_1^*(y)), a_1^*(y))$, which implies

$$\pi_2^e(a_2^*(y)) \geq \frac{\tilde{\pi}_2(\tilde{a}_2(V), V) + \tilde{\pi}_2(a_2^*(y)(a_1^*(y)), a_1^*(y))}{2}.$$

Then, since $\lambda_{212}(w, a_2^*(y)) = \tilde{\pi}_2(a_2^*(y)(w), w) - \pi_2^e(a_2^*(y))$, we obtain

$$\lambda_{212}(a_1^*(y), a_2^*(y)) \leq \frac{1}{2}\left[\tilde{\pi}_2(a_2^*(y)(a_1^*(y)), a_1^*(y)) - \tilde{\pi}_2(\tilde{a}_2(V), V)\right].$$

Furthermore, it holds that

$$\frac{1}{2}\left[\tilde{\pi}_2(a_2^*(y)(a_1^*(y)), a_1^*(y)) - \tilde{\pi}_2(\tilde{a}_2(V), V)\right] \leq \frac{1}{2}\left[\tilde{\pi}_2(\tilde{a}_2(V-\epsilon), V-\epsilon) - \tilde{\pi}_2(\tilde{a}_2(V), V)\right] \equiv \eta$$

if $a_1^*(y) \leq V - \epsilon$, so that $\lambda_{212}(a_1^*(y), a_2^*(y)) \leq \eta < 0$. Now consider the derivative of player 2's utility function w.r.t. e for given $w = a_1^*(y)$:

$$\frac{\partial U_2(e, a_1^*(y))}{\partial e} = a_1^*(y) - d'(e) + y\lambda_{212}(a_1^*(y), a_2^*(y))(V - a_1^*(y)).$$

Using the above results, it then holds that

$$\frac{\partial U_2(e, a_1^*(y))}{\partial e} \leq (V - \epsilon) + y\eta\epsilon.$$

Hence whenever $y > \bar{y} \equiv -(V-\epsilon)/(\epsilon\eta)$, we have that $U_2(e, a_1^*(y))$ is strictly decreasing in $e \in [0, 1]$, which implies $a_2^*(y)(a_1^*(y)) = 0$.

Step 3. Steps 1 and 2 together imply that $\exists y' > \bar{y}$ such that $a_1^*(y') \leq V - \epsilon$ and $a_2^*(y')(a_1^*(y')) = 0$. This implies $\tilde{\pi}_1(a_1^*(y'), a_2^*(y')(a_1^*(y'))) = 0$. We now claim that $\exists w' < V$ such that $a_2^*(y')(w') > 0$. Consider again the derivative $\partial U_2(e, w)/\partial e = w - d'(e) + y'\lambda_{212}(w, a_2^*(y'))(V - w)$. As $w \to V$, the reciprocity term goes to zero (because $\lambda_{212}(w, a_2^*(y'))$ is bounded), so $\partial U_2(0, w)/\partial e \to V - d'(0) > 0$, implying $a_2^*(y')(w') > 0$ for some sufficiently large $w' < V$. But then $\tilde{\pi}_1(w', a_2^*(y')(w')) > 0 = \tilde{\pi}_1(a_1^*(y'), a_2^*(y')(a_1^*(y')))$, which contradicts that $a^*(y') = (a_1^*(y'), a_2^*(y')) \in \mathscr{A}(y')$.