Intention-Based Reciprocity
and the Hidden Costs of Control*

Ferdinand von Siemens
University of Amsterdam & CESifo

Abstract

Empirical research shows that - rather than improving incentives - exerting more control can reduce performance by crowding out intrinsic motivation. The present paper shows that such motivational crowding-out can be explained in a model with intention-based reciprocity. It demonstrates that if individuals differ in their propensity for reciprocity concerns while preferences are private information, then the very act of being controlled can be considered to be unfriendly. Imposing control can then induce low effort from individuals who - absent control - voluntary exert high effort. This new intention-based argument stands in stark contrast to existing theoretical wisdom on motivational crowd-out that is exclusively based on signaling motives.

*Helpful comments and support from Michael Kosfeld, Randolph Sloof, and Jeroen van de Ven are gratefully acknowledged.

Keywords: extrinsic and intrinsic motivation, crowd-out, intention-based reciprocity, incomplete information, hidden costs of control.
1 Introduction

Many human resource managers argue that - quite contrary to standard economic incentive theory - exerting control can damage worker performance by reducing intrinsic motivation. Over the past decades numerous empirical studies from organizational economics, cognitive psychology, and behavioral economics have firmly validated this claim. In a prominent recent article Falk and Kosfeld (2006) investigate the detrimental effect of control on intrinsic motivation in an experimental work relationship. Workers can exert costly effort to increase the payoff of their bosses. Before workers choose their effort, bosses decide how strongly whether to control their worker or not. Imposing control forces the worker to exert a higher minimum effort level. If workers rationally maximize their own payoff, they should always exert the minimum feasible effort to save on effort costs. Falk and Kosfeld find that even though many workers indeed exhibit this behavior, a substantial fraction of workers exert significantly less effort if controlled rather than if not controlled. In summary, exerting control severely damages average effort contribution.

This empirical finding creates an interesting theoretical challenge. At first impulse one might think that workers consider the pure act of not being controlled as kind or fair, and thus reciprocate with high effort. But as already pointed out by Falk and Kosfeld on p.1616 this intuition is inconsistent with existing intention-based reciprocity models like Rabin (1993) and Dufwenberg and Kirchsteiger (2004). The reason is the following. If in a fairness equilibrium worker reciprocate not being controlled with high effort, they receive higher payoffs if controlled rather than if not controlled. But in this case not controlling workers is actually unkind. This makes high effort choices as response to not being controlled unsupportable in equilibrium. In other words: if bosses expect workers to reciprocate not being controlled with high effort choices, then not controlling workers is no longer a kind action, and thus cannot trigger high effort as reciprocal reaction.

Although the above argument is absolutely correct, the present paper shows that intention-based reciprocity might well explain motivational crowd-out if it is a priori unclear whether forcing an individual to act in a certain way is really in line with that individual’s preferences. Consider the following general example. Suppose that a worker can choose between two alternative actions A and B. Suppose that a priori - for whatever reason - some workers prefer A over B, but other workers prefer B over A. Preferences are private information. Now suppose some boss forces his worker to choose A. Unknowingly this boss might be kind to his
particular worker if that worker prefers A over B, but he might also be unkind if his worker had other preferences. If it is now commonly believed that workers typically prefer B over A, forcing a worker with unknown preferences to choose A might be considered as on average unkind. This perception is also shared by workers who prefer A over B. In response, a worker who actually prefers the action that he is forced to do, might consider being forced to act in this way as rather unkind. Consequently, he adversely adjust his behavior if driven by reciprocity concerns.

To clarify the above general example, the present model considers a simplified version of the control game in Falk and Kosfeld. The crucial assumption is that a large fraction of individuals is purely selfish in the sense that they only care for their own monetary payoff, but that some individuals are fair-minded in the sense of Rabin (1993) and Dufwenberg and Kirchsteiger (2004). Preferences are private information. If the fair-minded workers are sufficiently fair-minded and the fraction of selfish workers is sufficiently high, there then exists an equilibrium in which (i) selfish workers always choose the minimum feasible effort, (ii) fair-minded workers choose low effort if controlled, but high effort if not controlled, (iii) selfish bosses do not control workers, while (iv) fair-minded bosses control workers. In this equilibrium exerting no control is unkind to fair-minded workers by the above argument. But it is kind to selfish workers, since the latter are forced to choose high effort if controlled. If bosses do not know their workers’ preferences while most workers are selfish, they are on average kind if they choose not to control their worker. Sufficiently fair-minded workers reciprocate this average kindness with high effort. Further, selfish bosses might find it optimal not to control if their fair-minded workers’ high effort response sufficiently increases their payoff. Fair-minded workers, however, might decide to control because they suffer a utility loss when not controlling since their on average their kind decision is on average not reciprocated.

The present paper adds to the existing theoretical literature on motivational crowding out by providing an entirely new rational for the empirically observed behavior. All commonly known existing theoretical models on motivation crowding out are essentially based on signalling motives: by exerting control bosses reveal some new information, and workers respond with their effort choice to this new information. Two prominent recent example are Sliwka (2006) and Ellingsen and Johannesson (2008). In Sliwka there are three types of workers: selfish workers who only care for their own payoff, fair-minded workers who also care for the payoffs of their bosses, and conformist workers whose preferences depend on their belief concerning the relative frequency of selfish or fair-minded workers. A boss who controls hi worker
thereby signals that he believes that there are many selfish workers. This makes conformist workers adapt the preferences of selfish workers, which results in low average effort choices. A boss who does not control thereby signals that there are many fair-minded workers. This makes conformist workers adapt the preferences of fair-minded workers, which results in high average effort choices. Exerting control can thus reduce average effort.

In Ellingsen and Johannesson there are only two types of individuals: selfish individuals and altruistic individuals. Their crucial assumption is that individuals care for their reputation - they want to appear to be altruistic - where in addition the opinion of peers is the more important the more likely these peers are altruistic themselves. Workers might signal their altruism to their boss by exerting effort, where altruistic workers have lower signalling costs than selfish workers. The reason is that altruistic workers also receive some direct benefit from increasing the payoff of their boss by exerting more effort. In a separating equilibrium altruistic workers convincingly signal their altruism by exerting more effort than selfish workers. Now a boss who does not control his worker signals that he is altruistic. This makes it more important for the worker to convince the boss that he is also altruistic. To ensure separation an altruistic worker must thus choose a very high effort. A boss who does control his worker signals that he is selfish. Signalling altruism to that boss is then no longer that rewarding for the worker. Consequently, an altruistic worker does not have to exert very high effort to ensure separation. Exerting control might therefore diminish effort. Both the above arguments hinge crucially on signalling motives. In contrast, the present paper provides an explanation that is completely independent from any signalling incentives. The reason is that workers perceive the pure act of not being controlled as kind, and thus reciprocate with high effort. Workers actually do not care about the type of their boss at all.

2 The Model

Strategic Situation, Strategies, and Types

Consider one boss who interacts with one worker. The strategic interaction between boss and worker can be characterized as follows. First, the boss decides whether he wants to control or trust the worker, \( a_b \in A_b = \{ c, nc \} \). Controlling causes no costs. The worker observes the decision of the boss and then decides how much effort \( a_w \) to choose. Effort can be high, medium, or low. The set of his possible effort choices \( A_w(a_b) \) depend on the control decision \( a_b \) of the boss, where \( A_w(c) = \{ h, m \} \) and \( A_w(nc) = \{ h, m, l \} \). By controlling the worker the boss can thus prevent the worker from fully shirking and exerting only low effort. The
worker’s effort choice determines both his payoff and the payoff of the boss. These are determined by the payoff functions $\pi_b : A_w \rightarrow \mathbb{R}$ and $\pi_w : A_w \rightarrow \mathbb{R}$. More effort strictly increases the payoff of the boss and decreases the payoff of the worker, thus $\pi_b(h) > \pi_b(m) > \pi_b(l)$ and $\pi_w(l) > \pi_w(m) > \pi_w(h)$.

In the present model preferences over outcome are not always fully determined by the above payoffs, but can also depend on strategies, intentions, and beliefs concerning intentions. Before defining preferences, it is thus necessary to define strategies. The paper only considers pure strategies. A fully specified pure strategy for a boss is a function $\alpha_b : \Theta \rightarrow A_b$ that specifies for each type his control choice. A fully specified strategy for the worker is a function $\alpha_w : \Theta \times A_b \rightarrow A_w(a_b)$ that specifies for each type his effort choice conditional on the control choice of the boss.

The crucial assumptions in this paper are that boss and worker can differ in their propensity for fairness concerns, where individual preferences are private information. An individual is either selfish or fair-minded, so that the type space is $\Theta = \{s, f\}$. The individual type $\theta \in \Theta$ is private information, but it is common knowledge that each individual is fair-minded with probability $\lambda \in [0, 1]$. Selfish and fair-minded individuals differ in their preferences as follows. Selfish individuals are exclusively interested in their own payoff. Fair-minded individuals have intention-based fairness concerns that follow the specification in Rabin (1993).

**Control, Effort, and Kindess**

The preferences of fair-minded individuals depend on the perceived kindness of the other individual, and the observed or perceived own kindness. It is thus important to define kindness. The definition essentially follows Rabin (1993). Consider first the kindness of a worker towards a boss with a particular control choice. This kindness depends on a worker’s effort choice given the feasible effort choices conditional on the control choice of the boss. Since the boss has chosen his action, the kindness of a worker does not depend on beliefs. Let $\pi_b^{max}(a_b)$ be the maximum payoff that the worker can give to a boss with control choice $a_b$, and let $\pi_b^{min}(a_w)$ be the corresponding minimum payoff that a worker can give to a boss with that particular control choice. Since there is a finite number of possible actions, all maximum and minimum payoffs are well defined. Define $\pi_b^e(a_b) = (\pi_b^{max}(a_b) + \pi_b^{min}(a_b))/2$ as the equitable payoff a worker can give to a boss with control choice $a_b$. This equitable payoff is exactly in
between the above extreme payoffs. Then

\[ k_{wb}(a_w, a_b) = \frac{(\pi_b(a_w) - \pi_b^c(a_b))}{(\pi_b^{max}(a_b) - \pi_b^{min}(a_b))} \] (1)

describes the kindness of a worker with effort \( a_w \) towards a boss with control choice \( a_b \). Note that given the strategic situation a worker can always choose between at least two different effort levels, thus the denominator of the above fraction always differs from zero.

Consider next the kindness of a boss towards a worker as perceived by the boss. This kindness depends on what payoff the boss expects to give with his control choice to a worker, which depends on belief the boss holds concerning the worker’s effort choice conditional on his own control choice. Let \( \beta_w \) be the belief of boss concerning the strategy - and thus the conditional effort choices - of the worker. Selfish and fair-minded workers might react differently to the control choice of the boss. The perceived kindness of the boss is thus contingent on the worker’s type \( \theta \). Define \( \pi^{max}_w(\beta_w, \theta) \) as the maximum payoff the boss believes to be able to give to a worker with type \( \theta \) if the boss holds belief \( \beta_w \) concerning the worker’s strategy. Define \( \pi^{min}_w(\beta_w, \theta) \) and \( \pi^e_w(\beta_w, \theta) \) as the respective minimum and equitable payoff. Then

\[ k_{bw}(a_b, \beta_w, \theta) = \frac{(\pi_w(\beta_w(\theta, a_b)) - \pi^e_w(\beta_w, \theta))}{(\pi^{max}_w(\beta_w, \theta) - \pi^{min}_w(\beta_w, \theta))} \] (2)

is the kindness of a boss with belief \( \beta_w \) concerning the worker’s strategy to a worker with type \( \theta \) as perceived by the boss. This holds unless the denominator equals zero. Otherwise, the boss cannot affect the worker’s payoff, and there is no room for kindness. The kindness of the boss is then normalized to zero. Again, note that since the boss might believe that selfish and fair-minded workers differ in their effort choice after observing his control choice, his expected kindness might depend on the worker’s type. This in combination with incomplete information on the worker’s type will essentially drive all results.

Finally, consider the kindness of a boss towards a worker with type \( \theta \) as perceived by the worker. This kindness depends on the worker’s belief concerning the belief of the boss concerning the worker’s strategy. Let \( \gamma_w \) denote this belief of the worker concerning his own strategy. Then the kindness of a boss towards a worker with type \( \theta \) if the worker believes the boss to hold belief \( \gamma_w \) concerning his own strategy equals

\[ k_{bw}(a_b, \gamma_w, \theta) = \frac{(\pi_w(\gamma_w(\theta, a_b)) - \pi^e_w(\gamma_w, \theta))}{(\pi^{max}_w(\gamma_w, \theta) - \pi^{min}_w(\gamma_w, \theta))} \] (3)

\(^1\)In Rabin the equitable payoff only takes into account payoffs that arise from Pareto-efficient payoff combinations. In this application strict monotonicity implies that all payoff combinations are Pareto-efficient.
where maximum, minimum, and equitable payoffs are defined analogously to the above definition. Kindness is set to zero if the denominator equals zero.

Preferences, Fairness Equilibrium, and Simplifying Assumptions

It is now possible to define the preferences of a fair-minded boss. Given his belief \( \beta_w \) concerning the worker’s strategy, a boss has a belief how the kindness \( k_{bw}(a_b, \beta_w, \theta) \) a particular control choice \( a_b \) is to a worker of type \( \theta \). But he does not know the worker’s type. He forms his expected kindness towards the worker, which is then

\[
E_\theta k_{bw}(a_b, \beta_w, \theta) = \lambda k_{bw}(a_b, \beta_w, f) + (1 - \lambda) k_{bw}(a_b, \beta_w, s) \tag{4}
\]

given his belief \( \lambda \) that the worker is fair-minded. Equally, the boss does not know how kind will be the worker’s effort response since he does not know the worker’s type or strategy. He forms the expected kindness of the worker, which equals

\[
E_\theta k_{wb}(a_b, \beta_w, \theta, a_b) = \lambda k_{wb}(a_b, \beta_w(f, a_b)) + (1 - \lambda) k_{wb}(a_b, \beta_w(s, a_b)) \tag{5}
\]

given belief \( \lambda \) concerning the worker’s type and his belief \( \beta_w \) concerning his effort choice. Given action \( a_b \) and the involved beliefs \( \lambda \) and \( \beta_w \) define

\[
U_b(a_b, \beta_w) = E_\theta \pi_b(\beta_w(\theta, a_b)) + \eta E_\theta k_{wb}(a_b, \beta_w(\theta, a_b)) E_\theta k_{bw}(a_b, \beta_w, \theta) \tag{6}
\]
as the respective expected utility of the boss. Parameter \( \eta \in \mathbb{R}^+ \) characterized the relative importance of fairness concerns. A fair-minded boss therefore cares for his expected payoff, but also cares for fairness. However, he does not know the worker’s type. He thus computes the expected kindness of the worker, and multiplies it with his expected kindness towards the worker.

Equally, define the expected utility of a worker given his own action \( a_w \), the control choice \( a_b \) of the boss, and his belief \( \gamma_w \) concerning the belief of the boss concerning the strategy of the worker as

\[
U_w(a_w, a_b, \gamma_w) = \pi_w(a_w) + \eta E_\theta k_{bw}(a_b, \gamma_w, \theta) k_{wb}(a_b, a_w) \tag{7}
\]

A fair-minded worker thus also cares for his own payoff, but also for fairness. It is implicitly assumed that all fair-minded individuals bosses and workers, put equal relative weight \( \eta \) on their fairness concerns. Equally, workers believe selfish and fair-minded bosses to hold equal belief \( \gamma_w \) concerning workers’ conditional effort strategy.
Note that fair-minded bosses and workers only care for the expected kindness of the action chosen by the other individual boss. Fair-minded bosses only care for the type of their workers in as far as the latter determine the workers’ effort choice. They do not care for their workers’ types as such. Fair-minded workers actually know the control choice of their bosses, and thus do not care about the type of the boss at all. Their utility does not depend on the ex-ante probability with which the boss is fair-minded. Selfish bosses and workers only care for their own expected payoffs. They respectively maximize and with set equal to zero.

In the present context with incomplete information, a fairness equilibrium is then defined as follows. First, equilibrium strategies and maximize type-dependent expected utility given the other individual’s equilibrium strategies and, given the bosses’ equilibrium first-order beliefs concerning the workers’ equilibrium strategy, and given workers’ equilibrium second-order beliefs concerning the equilibrium belief of bosses concerning workers’ equilibrium strategies. Second, first-order and second-order beliefs are consistent with the equilibrium strategies so that . Thirdly, the above holds in all subgames, so that equilibrium strategies and beliefs form a fairness equilibrium also in all subgames following the control choices of workers.

The following assumption simplifies the analysis without driving any of the results. It concerns control and effort choices in case of indifference.

Assumption 1 (Indifference) In case of indifference

(i) workers choose the lower possible effort, and
(ii) bosses control the worker.

This assumption avoids tedious case distinctions. It actually makes it harder to achieve the goal of the study - to find a fairness equilibrium in which some bosses do not control while fair-minded workers reciprocate not being controlled with exerting high effort.

3 Main Results

Since the kindness of a worker towards a boss does not depend on any beliefs or types, implies the following in any fairness equilibrium.
Lemma 1 (Kindness of Worker) In any fairness equilibrium

\[ k_{wb}(c, m) = -\frac{1}{2}, \quad k_{wb}(c, h) = +\frac{1}{2}, \]  
\[ k_{wb}(nc, l) = -\frac{1}{2}, \quad k_{wb}(nc, h) = +\frac{1}{2}, \]  
\[ k_{wb}(c, m) = \frac{2\pi b(m) - \pi_b(h) - \pi_b(l)}{2(\pi_b(h) - \pi_b(l))}. \]  

characterize the kindness of a worker towards a boss.

A selfish worker always maximizes his payoff, and therefore chooses the minimum possible effort. This directly implies the following.

Lemma 2 (Behavior Selfish Worker) In any fairness equilibrium

\[ \alpha^*_w(s, nc) = l \quad \text{and} \quad \alpha^*_w(s, c) = m \]  

characterize the equilibrium behavior of selfish workers.

Since the behavior of selfish workers is identical in all fairness equilibria, Lemma 2 yields directly \( \pi^\text{max}_w(\alpha^*_w, s) = \pi_w(l) \), \( \pi^\text{min}_w(\alpha^*_w, s) = \pi_w(m) \), and \( \pi^e_w(\alpha^*_w, s) = (\pi_w(l) + \pi_w(m))/2 \). This implies that controlling a selfish worker is unkind, whereas not controlling a selfish worker is kind. This yields the following.

Lemma 3 (Kindness Boss towards Selfish Worker) In any fairness equilibrium

\[ k_{bw}(nc, \alpha^*_w, s) = \frac{1}{2} \quad \text{and} \quad k_{bw}(c, \alpha^*_w, s) = -\frac{1}{2} \]  

characterize the kindness of a boss towards a selfish worker.

Before presenting the main result define

\[ \eta_1 = \frac{(\pi_w(l) - \pi_w(h))}{1 - 2\lambda} \]  
\[ \eta_2 = \frac{2(\pi_w(m) - \pi_w(h))(\pi_b(h) - \pi_b(l))}{(1 - 2\lambda)(\pi_b(h) - \pi_b(l))} \]  

Further, let \( \lambda_1 \) be the unique solution in \([0, 1]\) that solves

\[ \pi_b(m) - \pi_b(l) = \lambda_1 (\pi_b(h) - \pi_b(l)) - \eta(1 - \lambda_1)(1 - 2\lambda_1)/2 \]  

and define

\[ \lambda_2 = \frac{(\pi_b(m) - \pi_b(l))}{(\pi_b(h) - \pi_b(l))}. \]

The present paper’s main result is made formally precise in the following proposition. All formal proofs can be found in the appendix.
Proposition 1 (High Effort Reciprocity) Consider a fairness equilibrium in which fair-minded workers reciprocate not being controlled in the sense that $\alpha_{w}^{*}(f, nc) = h$ and $\alpha_{w}^{*}(f, c) = m$. Selfish workers behave as characterized in Lemma 2. Then the following holds.

(i) Such an equilibrium exists if and only if $\lambda < 1/2$ and $\eta > \max\{\eta_1, \eta_2\}$.

(ii) Selfish bosses choose control if and only if $\lambda \leq \lambda_1$. Fair-minded bosses choose control if and only if $\lambda \leq \lambda_2$.

(iii) If $\lambda_2 < 1/2$ then $0 < \lambda_1 < \lambda_2 < 1/2$. For $\lambda \in ]\lambda_1, \lambda_2[$ selfish bosses then choose not to control, whereas fair-minded bosses choose to control.

Proposition 1 is based on the following intuition. Lemma 3 implies that if a boss is facing selfish workers with sufficiently high probability - the ex-ante probability $\lambda$ is larger than $1/2$ - then controlling a worker with unknown type is on average unkind, whereas not controlling a worker with unknown type is on average kind. If fair-minded workers sufficiently care for fairness - their fairness parameter $\eta$ exceeds a certain threshold - then they reciprocate note being controlled by voluntarily exerting high effort. This generates a fairness equilibrium that captures the most important qualitative feature of motivational crowd-out as observed in the empirical studies.

It is important to note that signalling motives play no role whatsoever in the above characterized fairness equilibrium with motivational crowd-out. Given the fraction $\lambda$ of fair-minded individuals the control choice of the boss might convey information on his reciprocity type. But workers do not care for the type of the boss, but for the latter’s intentions. If a boss is typically screwed when not controlling the worker, then not exerting control is considered a kind action as such. Fair-minded workers reciprocate this kind action with high effort.

Whether bosses want to control their workers primarily depends on the fraction $\lambda$ of fair-minded individuals in the population and the benefit for the boss of a worker exerting high rather than low effort. If this difference $\pi_b(h) - \pi_b(m)$ is high then it pays not to control the worker even if only very few workers then reciprocate with high effort. For fair-minded bosses their control choice is also influenced by their own fairness concerns and the expected kindness of the worker. If bosses control, then in equilibrium all worker shirk and are thus unkind. Since controlling is also unkind, controlling workers generates a spiteful utility kick to fair-minded bosses. Further, not controlling workers is in equilibrium typically not reciprocated by workers. On average workers who are not controlled are therefore unkind. Since not
controlling workers is on average kind, the unkind negative average reaction by workers really hurts fair-minded workers. In consequence, fair-minded bosses are more likely to control workers than selfish bosses. Heterogeneous control behavior can thus arise in a fairness equilibrium in which all individuals hold the same equilibrium beliefs concerning the fraction of fair-minded individuals in the population and their respective equilibrium behavior.

4 Further Results

This section completes the equilibrium analysis by characterizing all remaining fairness equilibria. It starts with fairness equilibria in which fair-minded workers reciprocate not being controlled only by exerting medium effort. Define

\[ \eta_3 = \frac{2(\pi_w(l) - \pi_w(m))(\pi_b(h) - \pi_w(l))}{(1 - \lambda)(\pi_b(m) - \pi_b(l))} \]  
\[ \eta_4 = \frac{2(\pi_w(m) - \pi_w(h))(\pi_b(h) - \pi_w(m))}{(1 - \lambda)(\pi_b(h) - \pi_b(l))}. \]

There then is the following result.

**Proposition 2 (Medium Effort Reciprocity)** Consider a fairness equilibrium in which fair-minded workers reciprocate not being controlled in the sense that \( \alpha_{w}(f, nc) = m \) and \( \alpha_{w}(f, c) = m \). Selfish workers behave as characterized in Lemma 2. Then the following holds.

(i) Such an equilibrium exists if and only if \( \eta_4 \geq \eta_3 \) and \( \eta \in [\eta_3, \eta_4] \).

(ii) Both selfish and fair-minded bosses always control.

In such an equilibrium not controlling is kind to selfish workers while reciprocally neutral to fair-minded workers. Reciprocity levels then have to be intermediate: they must be strong enough so that uncontrolled fair-minded workers do not exert low effort, but they must not be too strong so that uncontrolled fair-minded workers do not exert high effort. Selfish bosses always control since not controlling workers never increases effort not even for fair-minded workers. Fair-minded workers also always control. The reason is that exerting no control is a kind action that is not really well reciprocated, so that fair-minded bosses then suffer a utility loss. Controlling is an unkind action, that triggers an unkind effort reaction, which in turn generates a positive spiteful utility kick. As monetary and reciprocity incentives are aligned, fair-minded workers optimally control.

Next define

\[ \eta_5 = \frac{2(\pi_w(l) - \pi_w(h))}{(\pi_b(h) - \pi_w(l))} \]  
\[ \eta_6 = \frac{2(\pi_w(l) - \pi_w(m))(\pi_b(h) - \pi_b(l))}{(\pi_b(m) - \pi_b(l))}. \]
There then is the following result.

**Proposition 3 (No Reciprocity)** Consider a fairness equilibrium in which fair-minded workers do not reciprocate not being controlled in the sense that $\alpha^*_w(f,nc) = l$ and $\alpha^*_w(f,c) = m$. Selfish workers behave as characterized in Lemma 2. Then the following holds.

(i) Such an equilibrium exists if and only if $\eta \leq \min\{\eta_5, \eta_6\}$.

(ii) Both selfish and fair-minded bosses always control.

In this equilibrium not controlling is clearly very kind as it is kind to both fair-minded and selfish workers. However, reciprocity concerns are so weak so that even then not being controlled cannot trigger effort choices that exceed low effort. Selfish bosses clearly have monetary incentives to control, and as in Proposition 2 fair-minded bosses have additional reciprocity incentives to control workers in order to enjoin their spitefulness. The simple formal proof is omitted. The following result completes the analysis.

**Proposition 4 (No Other Equilibria)** There exist no other fairness equilibria besides those characterized in Propositions 1 to 3.

In any fairness equilibrium not covered in Propositions 1 to 3 fair-minded workers exert high effort if controlled. This can only be optimal if controlling workers is considered kind. The latter is never the case for selfish workers, and given fair-minded workers’ high effort choice if controlled, it can also never be the case for fair-minded workers. Controlling workers must this be considered on average as unkind concerning a worker with unknown type. But then it cannot be optimal for fair-minded workers to exert more than the minimum possible effort if controlled.

**5 Conclusion**

The present paper has shown how intention-based reciprocity can explain the crowd out of intrinsic motivation if individuals differ in their propensity for reciprocity concerns while individual preferences are private information. The main argument is that if most individuals are selfish and only care for their own payoff, then not exerting control is kind exactly because is is typically not reciprocated with high effort. The paper adds to the existing theoretical literature on extrinsic incentives and intrinsic motivation by offering an explanation for motivation crowd out that is not based on signalling motives.
Appendix

Proof of Proposition 1

The equilibrium behavior of fair-minded workers implies \( \pi_{w}^{\max}(\alpha_{w}^{*}, f) = \pi_{w}(m) \) and \( \pi_{w}^{\min}(\alpha_{w}^{*}, f) = \pi_{w}(h) \) so that \( \pi_{w}^{e}(\alpha_{w}^{*}, f) = (\pi_{w}(m) + \pi_{w}(h))/2 \). Concerning the kindness of a boss towards a fair-minded worker this yields \( k_{bw}(nc, \alpha_{w}^{*}, f) = -1/2 \) and \( k_{bw}(c, \alpha_{w}^{*}, f) = 1/2 \). The expected kindness of the boss towards a workers with unknown type is then

\[
E_{b}k_{bw}(c, \alpha_{w}^{*}, \theta) = \lambda - 1/2 \quad \text{and} \quad E_{b}k_{bw}(nc, \alpha_{w}^{*}, \theta) = 1/2 - \lambda.
\]

(21)

while

\[
E_{b}k_{wb}(nc, \alpha_{w}^{*}(\theta)) = \lambda - 1/2 \quad \text{and} \quad E_{b}k_{wb}(c, \alpha_{w}^{*}(\theta)) = -1/2.
\]

(22)

is the expected kindness of the worker towards the boss as perceived by the boss.

Now consider the optimality of the equilibrium behavior of bosses and workers. Consider first workers. The behavior of selfish workers is optimal given Lemma 2. Given Assumption 1 the behavior of fair-minded workers is optimal if and only if

\[
\pi_{w}(h) + \eta(1/2 - \lambda)(1/2) > \pi_{w}(l) + \eta(1/2 - \lambda)(-1/2)
\]

(23)

\[
\pi_{w}(h) + \eta(1/2 - \lambda)/2 > \pi_{w}(m) + \eta(1/2 - \lambda)(2\pi_{b}(m) - \pi_{b}(h) - \pi_{b}(l))/(\pi_{b}(h) - \pi_{b}(l))
\]

(24)

\[
\pi_{w}(h) + \eta(1/2 - \lambda)(1/2) \geq \pi_{w}(h) + \eta(\lambda - 1/2)(1/2).
\]

(25)

Constraints (23) and (24) ensure that if not controlled, the worker prefers to exert high rather than medium or low effort. As \( \pi_{w}(h) < \pi_{w}(l) \) constraint (23) can hold only if \( \lambda < 1/2 \). This condition with \( \pi_{w}(m) > \pi_{w}(h) \) implies that (25) can be ignored. Rearranging (23) yields as condition \( \eta > \eta_{1} \) and (24) yields \( \eta > \eta_{2} \) with the cutoffs \( \eta_{1} \) and \( \eta_{2} \) as defined above. Note that since \( \pi_{b}(h) - \pi_{b}(l) > \pi_{b}(h) - \pi_{b}(m) \) it is not clear whether (23) or (24) is binding even though \( \pi_{w}(l) - \pi_{w}(h) > \pi_{w}(m) - \pi_{w}(h) \).

Consider next the optimality of the equilibrium behavior of bosses. Given Assumption 1 selfish bosses control in equilibrium if and only if \( \pi_{b}(m) \geq \lambda\pi_{b}(h) + (1 - \lambda)\pi_{b}(l) \) or

\[
\pi_{b}(m) - \pi_{b}(l) \geq \lambda(\pi_{b}(h) - \pi_{b}(l)) = A(\lambda).
\]

(26)

This yields as condition \( \lambda \leq \lambda_{2} \) with \( \lambda_{2} \) as defined above which directly yields \( \lambda_{2} \in [0, 1] \).

Given Assumption 1 fair-minded bosses control in equilibrium if and only if

\[
\pi_{b}(m) - \pi_{b}(l) \geq \lambda(\pi_{b}(h) - \pi_{b}(l)) + \eta(1 - \lambda)(\lambda - 1/2) = B(\lambda).
\]

(27)
Since $B(0) = -\eta/2 < \pi_b(m) - \pi_b(l)$ and $B(1) = \pi_b(h) - \pi_b(l) > \pi_b(m) - \pi_b(l)$ continuity of $B$ on $[0,1]$ and the intermediate value theorem imply that there exists $\lambda_1 \in ]0,1[$ such that $B(\lambda_1) = \pi_b(m) - \pi_b(l)$. Further, function $B$ is strictly concave on $[0,1]$ since $B''(\lambda) = -2\eta$. It is thus quasi-concave so that the upper contour sets $P(x) = \{\lambda \in ]0,1[ : B(\lambda) \geq x\}$ are compact intervals. $B(1) > \pi_b(m) - \pi_b(l)$ implies $1 \in P(\pi_b(m) - \pi_b(l))$ which yields that $P(\pi_b(m) - \pi_b(l)) = [\lambda_1,1]$. Convexity of the sets $P$ implies $B(\lambda) < \pi_b(m) - \pi_b(l)$ for all $\lambda < \lambda_1$. Strict concavity of $B$ further implies that $B(t\lambda_1 + (1-t)) > tB(\lambda_1) + (1-t)B(1) > \pi_b(m) - \pi_b(l)$ for all $t \in ]0,1[$. Thus, there exists a unique $\lambda_1$ in $]0,1[$ that solves (15) with equality so that (??) holds for all $\lambda > \lambda_1$.

Finally, suppose $\lambda_1 < 1/2$. Then $A(\lambda) > B(\lambda)$ and therefore $B(\lambda_1) < A(\lambda_1) = \pi_b(m) - \pi_b(l)$. Further, $B(1/2) = A(1/2) > A(\lambda_1) = \pi_b(m) - \pi_b(l)$. Then the above arguments imply that there exists a unique $\lambda_2 \in ]\lambda_1,1/2[$ with the property $B(\lambda_2) = \pi_b(m) - \pi_b(l)$. For $\lambda \in ]\lambda_1, \lambda_2[$ optimality of the behavior of selfish and fair-minded workers then follows from the above conditions. \emph{Q.E.D.}

Proof of Proposition 2

The equilibrium behavior of fair-minded workers implies $\pi^w_{\text{max}}(\alpha^*_w, f) = \pi^w_{\text{min}}(\alpha^*_w, f) = \pi^w(\alpha^*_w, f) = \pi_w(m)$. Concerning the kindness of a boss towards a fair-minded worker this yields $k_{bw}(nc, \alpha^*_w, f) = k_{bw}(c, \alpha^*_w, f) = 0$. The expected kindness of the boss towards a workers with unknown type is then

$$E_{\theta} k_{bw}(c, \alpha^*_w, \theta) = (\lambda - 1)/2 \quad \text{and} \quad E_{\theta} k_{bw}(nc, \alpha^*_w, \theta) = (1 - \lambda)/2. \quad (28)$$

Controlling is thus always unkind for all $\lambda \in ]0,1[$ as it is unkind to selfish workers while it is neutral to fair-minded workers. Further,

$$E_{\theta} k_{wb}(nc, \alpha^*_w(\theta)) = \lambda(2\pi_b(m) - \pi_b(h) - \pi_b(l))/(2(\pi_b(h) - \pi_b(l))) + (1 - \lambda)(-1/2) \quad (29)$$

$$E_{\theta} k_{wb}(c, \alpha^*_w(\theta)) = -1/2. \quad (30)$$

is the expected kindness of the worker towards the boss as perceived by the boss. Note that $E_{\theta} k_{wb}(nc, \alpha^*_w(\theta)) < 0$ since $k_{wb}(nc, \alpha^*_w(s)) = -1/2$ while $k_{wb}(nc, \alpha^*_w(f)), 1/2$.

Now consider the optimality of the equilibrium behavior of bosses and workers. Consider first workers. The behavior of selfish workers is optimal given Lemma 2 and Given Assumption 1
simplifying some expressions the behavior of fair-minded workers is optimal if and only if
\[ \pi_w(m) + \eta(1 - \lambda)(2\pi_b(m) - \pi_b(h) - \pi_b(l))/(4(\pi_b(h) - \pi_b(l))) > \pi_w(l) - \eta(1 - \lambda)/4 \] (31)
\[ \pi_w(m) + \eta(1 - \lambda)(2\pi_b(m) - \pi_b(h) - \pi_b(l))/(4(\pi_b(h) - \pi_b(l))) \geq \pi_w(h) + \eta(1 - \lambda)/4 \] (32)
\[ \pi_w(m) - \eta(\lambda - 1)/4 \geq \pi_w(h) + \eta(\lambda - 1)/4. \] (33)

Constraint (33) ensured that if controlled, a fair-minded worker exerts medium rather than high effort. Since \( \lambda < 1 \) implies that controlling the worker is unkind, this constraint can be ignored because \( \pi_w(m) > \pi_w(h) \). Constraint (31) ensures that if not controlled, a fair-minded worker exerts medium rather than low effort. Solving for \( \eta \) yields as condition \( \eta > \eta_3 \).

Constraint (31) ensures that if not controlled, a fair-minded worker exerts medium rather than high effort. Solving for \( \eta \) yields as condition \( \eta \leq \eta_3 \). Note the weak inequality due to Assumption 1. Consequently, the above constraints can be satisfied if and only if \( \eta_4 \geq \eta_3 \) and \( \eta \in [\eta_3, \eta_4] \). Note that there are no assumptions concerning the relative magnitude of \( \pi_w(l) - \pi_w(m) \) and \( \pi_w(m) - \pi_w(h) \). Thus, the existence of the interval \( [\eta_3, \eta_4] \) in unclear and depends on the particular parameters of the model.

Consider next the optimality of the equilibrium behavior of the bosses. In this equilibrium selfish bosses always control since the latter never decreases, but sometimes increases workers’ effort choices. Fair-minded bosses control if and only if
\[ \pi_b(m) + \eta(1 - \lambda)/4 \geq \lambda\pi_b(m) + (1 - \lambda)\pi_b(l) \]
\[ + \eta(\lambda(2\pi_b(m) - \pi_b(h) - \pi_b(l))/(2(\pi_b(h) - \pi_b(l))) + (1 - \lambda)(-1/2))(1 - \lambda)/2. \] (34)

This condition is automatically satisfied since workers are on average unkind if controlled and there are no monetary incentives not to control.

Q.E.D.

Proof of Proposition 4

In any equilibrium not characterized in Propositions 1 to 3 fair-minded workers choose high effort if controlled. Consider a potential equilibrium with \( \alpha_w^*(f, c) = \alpha_w^*(f, nc) = h \). Then the respective kindness of bosses towards workers is as in the equilibrium in which fair-minded workers always choose medium effort, as bosses cannot affect the payoffs of fair-minded workers. This implies
\[ E_\theta k_{bw}(c, \alpha_w^*, \theta) = (\lambda - 1)/2 \quad \text{and} \quad E_\theta k_{bw}(nc, \alpha_w^*, \theta) = (1 - \lambda)/2. \] (35)

For fair-minded workers it is then optimal to exert high effort if controlled if and only if
\[ \pi_w(h) + \eta(1/2)(\lambda - 1)/2 > \pi_w(m) + \eta(-1/2)(\lambda - 1)/2 \] (36)
which can never hold since $\lambda < 1$ and $\pi_w(h) < \pi_w(m)$. The proof is analogous for all other equilibrium effort choices $\alpha^*_w(f, nc)$. Q.E.D.

References


