CESifo Working Papers
www.cesifo.org/wp

Intention-Based Reciprocity and the Hidden Costs of Control

Ferdinand von Siemens

CESIFO WORKING PAPER NO. 3553
CATEGORY 13: BEHAVIOURAL ECONOMICS
AUGUST 2011

CESifo *Working Paper No. 3553*

Intention-Based Reciprocity and the Hidden Costs of Control

Abstract

Empirical research suggests that - rather than improving incentives - exerting control can reduce workers' performance by eroding motivation. The present paper shows that intention-based reciprocity can cause such motivational crowding-out if individuals differ in their propensity for reciprocity and preferences are private information. Not being controlled might then be considered to be kind, because not everybody reciprocates not being controlled with high effort. This argument stands in contrast to existing theoretical wisdom on motivational crowding-out that is primarily based on signaling models.

JEL-Code: A130, C700, D630, D820, L200.

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

Ferdinand von Siemens
University of Amsterdam
Faculty of Economics and Business
Roetersstraat 11
NL – 1018 WB Amsterdam
The Netherlands
f.a.vonsiemens@uva.nl

August 2011
Helpful comments from Georg Kirchsteiger, Michael Kosfeld, Matthew Rabin, Randolph Sloof, Joep Sonnemans, and Jeroen van de Ven are gratefully acknowledged.

1 Introduction

There is a wide-spread belief in human resource management and the popular business press that exerting control can damage worker performance by eroding intrinsic motivation.[1] This view is consistent with numerous empirical studies from both psychology and organizational economics.[2] In their seminal contribution Falk and Kosfeld (2006) investigate such crowding-out of motivation in an experimental work relationship. Workers can exert costly effort to increase the payoff of their bosses. Before workers choose effort, bosses decide whether to control their workers. Imposing control forces workers to exert higher minimum effort. If workers maximize their own payoff, they should exert the minimum feasible effort to save on effort costs. Falk and Kosfeld find that even though many workers indeed always choose the minimum feasible effort, a substantial fraction of workers exert less effort if controlled than if not controlled. Exerting control in fact reduces average effort contributions.[3]

This empirical finding creates an interesting theoretical challenge. Some individuals choose high effort if not controlled, and only medium effort if controlled. But choosing high effort remains feasible even when being controlled: observed behavior thus cannot be reconciled with transitive preferences defined purely over payoff outcomes. Then why does exerting control reduce voluntary cooperation? One might think that workers consider the pure act of not being controlled as kind, and reciprocate with high effort. But as Falk and Kosfeld (2006, p.1616) point out, this seems to be inconsistent with existing models of intention-based reciprocity. The reason is that if workers exert higher effort if not controlled rather than if controlled, they receive higher payoffs if controlled rather than if not controlled. But then not controlling workers has to be considered unkind, exactly because workers then exert higher effort. In short: if everybody expects workers to reciprocate not being controlled with

[1] See for example Manzoni and Barsoux (1998) and Herzberg (2003) who stress the negative consequences of exerting tight control over employees. Foss (2003) provides a careful case study on the detrimental effects of such micro-management.

[2] See for example Plant and Ryan (1985), Enzle and Anderson (1993), Barkema (1995), Ariely, Kamenica, and Prelec (2008), Dickinson and Villeval (2008), and Dominguez-Martinez, Sloof, and von Siemens (2010). These studies suggest that monitoring reduces performance if and only if the latter is perceived as imposing control. Ryan and Deci (2000) and Frey and Jegen (2001) provide some theoretical background and additional references to the extensive empirical evidence.

[3] Although there is some debate in the literature concerning the magnitude of the effects, an increasing number of studies provide experimental evidence for the existence of hidden costs of control. See in particular Charness, Cobo Reyes, Jiménez, Lacomba, and Lagos (forthcoming) and Schnedler and Vadovic (forthcoming) who study the hidden costs and benefits of delegation and control.

particularly high effort, then not exerting control is no longer a kind action, and thus cannot trigger high effort as reciprocal reaction.

Although the above argument is correct, the contribution of this paper is to demonstrate that intention-based reciprocity can explain motivational crowding-out if individuals differ in their propensity for fairness concerns and preferences are private information. The model considers a simplified version of the control game from Falk and Kosfeld. The key assumption is that individuals differ in their propensity for reciprocity: some are purely selfish in the sense that they only care for their own monetary payoffs, whereas others are reciprocal in the sense of Rabin (1993) and Dufwenberg and Kirchsteiger (2004). Preferences are private information. The analysis shows that if reciprocal workers are sufficiently reciprocal and the fraction of selfish workers is sufficiently high, there exists a pure-strategy reciprocity equilibrium in which (i) selfish workers choose the minimum feasible effort, whereas (ii) reciprocal workers choose lower effort if controlled than if not controlled. In such an equilibrium exerting no control is unkind to reciprocal workers by the above argument. But it is kind to selfish workers, since the latter are forced to choose higher effort if controlled. Bosses do not know workers' preferences, and workers know that bosses do not know workers' preferences. Define the kindness of bosses as the expected kindness towards workers. In case many workers are selfish, bosses are then - on average - kind if they do not control. In consequence, reciprocal workers reciprocate not being controlled with high effort.

2 Related Literature

The present analysis adds to the theoretical literature on crowding-out by providing a new rationale for hidden costs of control. In contrast to the present model, all prominent existing theoretical models are based on signaling motives. Most closely related are Sliwka (2007) and Ellingsen and Johannesson (2008) since they include behavioral elements in analyses and explicitly refer to the empirical evidence from Falk and Kosfeld. In Sliwka (2007) exerting no control indicates that there is a high fraction of steadfast fair-minded workers. This induces conformist workers - who want to comply with the prevalent social norm among steadfast workers - to exert high effort. Ellingsen and Johannesson (2008) argue that not controlling signals altruism. This makes it more rewarding for workers to signal their altruism. For this signaling to be credible, altruistic workers thus have to choose higher effort if not controlled than if controlled.

These studies build on an extensive signaling literature on motivational crowding-out. The following list of articles summarizes some of the most popular arguments. In Spier (1992) bosses might leave contracts incomplete - wages do not condition on project outcomes - to signal that the job environment is more likely to yield favorable outcomes. This facilitates worker participation. Bénabou and Tirole (2003) focus on effort choices. Monetary incentives or control signals that workers effort is less productive. Although monetary incentives increase motivation in the short run, in the long run the effect is detrimental. In Suvorov and van de Ven (2009) bosses transmit information on performance by paying discretionary bonuses. This conveys information on workers' ability, which in turn affects the latters' future effort decisions. Herold (2010) argues that leaving contracts incomplete reveals that bosses believe their workers to be trustworthy. Bosses' beliefs affect their own future effort contributions, which in turn influences workers' effort choices.

In the present paper private information on individual preferences is crucial. But in contrast to all prominent existing arguments, the resulting explanation for motivational crowding-out is completely independent from any signaling incentives. The reason is that workers perceive the act of not being controlled as kind, and reciprocate with high effort. They do not care about the types of their bosses at all, and any revealed information on bosses' preferences has absolutely no impact on behavior.

Apart from complementing the existing theoretical arguments, the present model can also be distinguished empirically from signaling explanations of motivational crowing-out. Signaling only works if bosses possess information that is relevant for workers. This assumption is not always equally plausible. In stable environments experienced workers could have gathered a lot of information on prevalent social norms and the fraction of fair-minded steadfast workers. Experienced workers might also know themselves and the characteristics of their job. The altruism of a boss could have been revealed by his previous behavior towards workers and stakeholders. It is also not clear - especially in anonymous laboratory situations - whether being controlled can reveal information on what the boss believes concerning that particular worker. Getting additional information in these situations - via the control choice of the boss - might then have no big impact on beliefs. Motivational crowding-out should then be limited according to signaling models. The present explanation based on intention-based reciprocity predicts an undiminished deterioration of motivation, since not being controlled is experienced as kind or friendly as such.

Finally, the present analysis complements economic theory on intention-based reciprocity by elaborating on the impact of incomplete information. Geaneakoplos, Pearce, and Stacchetti (1989, pp.67-68) already indicate that psychological game theory might encompass incomplete information. Battigalli and Dufwenberg (2009) further study psychological game theory in dynamic situations that include incomplete information. Concerning reciprocity, Rabin (1993, p.1296) argues that "Extending the model to incomplete-information games is essential for applied research, but doing so will lead to important issues." The present paper illustrates and discusses some of these issues, and it proposes one model specification that is consequently shown to be fruitfully applicable. Rabin furthermore argues that incomplete information might strongly affect the consequences of reciprocity. The present results fully corroborate this view, because intention-based reciprocity can explain motivational crowding-out if and only if reciprocity preferences are private information. The findings suggest that a further investigation into intention-based reciprocity and incomplete information might constitute an interesting topic for future research.

3 The Model

Consider one boss interacting with one worker. The strategic situation is as follows. First, the boss decides whether to control or not control the worker, $a_b \in A_b = \{c, nc\}$. Controlling causes no costs. The worker observes the decision of the boss and then decides on his effort a_w. Effort can be high, medium, or low. The set of possible effort choices $A_w(a_b)$ depends on the control decision a_b of the boss, where $A_w(c) = \{h, m\}$ and $A_w(nc) = \{h, m, \ell\}$. By controlling the worker, the boss can thus prevent the worker from exerting only low effort. The worker's effort choice determines both his payoff and the payoff of the boss as determined by the payoff functions $\pi_b : A_w \to \mathbb{R}$ and $\pi_w : A_w \to \mathbb{R}$. More effort strictly increases the payoff of the boss and strictly decreases the payoff of the worker, thus $\pi_b(h) > \pi_b(m) > \pi_b(\ell)$ and $\pi_w(\ell) > \pi_w(m) > \pi_w(h)$. The strategic situation - without accounting for incomplete information - is summarized in Figure 1.

The crucial assumptions in this paper are that individuals can differ in their propensity for fairness concerns while individual preferences are private information. An individual is either selfish or reciprocal, so that the type space is $\Theta = \{s, r\}$. The individual type $\theta \in \Theta$ is private information, but it is common knowledge that each individual is reciprocal with prior probability $\lambda \in]0, 1[$. Selfish individuals are exclusively interested in their own payoff. Reciprocal individuals have intention-based fairness concerns in the spirit of Rabin (1993) and

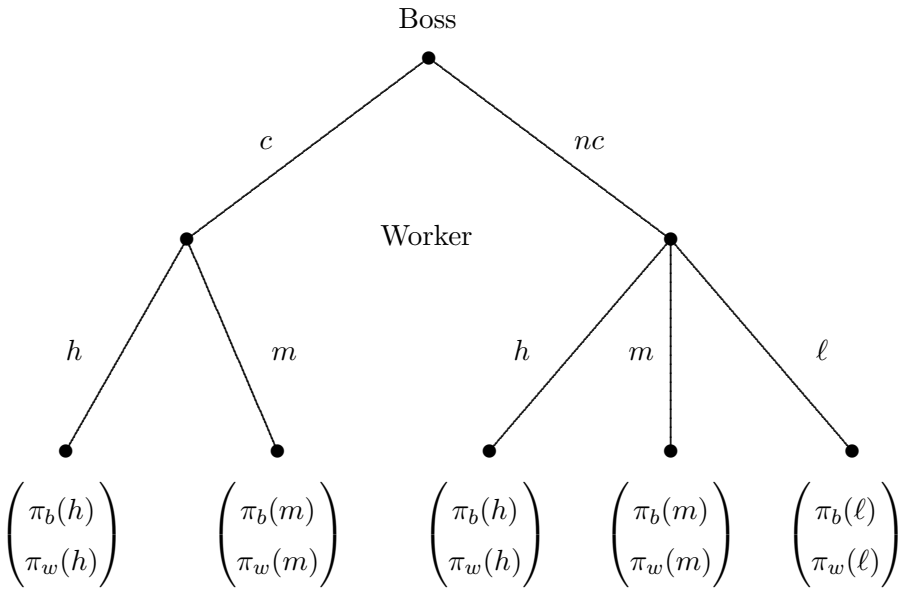

Figure 1: Control Game

Dufwenberg and Kirchsteiger (2004). These worker's utility is therefore not fully determined by the above payoffs, but also depends on strategies and beliefs concerning strategies. The paper only considers pure strategies.[4] A fully specified pure strategy for the boss is a function $\alpha_b : \Theta \to A_b$ that specifies type-dependent control choices. A fully specified strategy for the worker is a function $\alpha_w : \Theta \times A_b \to A_w(a_b)$ that specifies type-dependent effort choice conditional on the observed control choice of the boss.

Perceived Kindness of Control Effort Choices

The present paper applies the definitions of reciprocity by Rabin (1993) and Dufwenberg and Kirchsteiger (2004) while incorporating incomplete information.[5] The kindness of an action

[4]In psychological games it is not trivial to evaluate the kindness of an observed action resulting from a mixed strategy, as it depends on whether mixed strategies are viewed as deliberate mixing by an individual. Further, the intentions of any observed actions will be interpreted as if behavior is always fully deliberate. This seems to be more appropriate in the context of pure strategies. Like the present paper, the existing literature on intention-based reciprocity focuses on pure strategies. For further discussion see Rabin (1993, p.1286), Dufwenberg and Kirchsteiger (2004, p.275 and pp.279-280), and Segal and Sobel (2007, pp.209-210).

[5]Dufwenberg and Kirchsteiger (2004, pp.287-290) explain the subtle differences between their and Rabin's specification. The changes are primarily adopted to facilitate equilibrium existence in sequential games. Since equilibrium existence is not the main focus of the analysis, the present model largely follows Rabin's specification. Section 5 further discusses the assumptions and related literature on reciprocity; special focus is put on the robustness of results to the exact specification of reciprocity.

is thus assessed by putting the payoff consequences of that action in relation to the set of payoff consequences that could have been achieved by choosing alternative actions.[6]

Consider workers whose bosses haven chosen control action a_b. Let $\pi_b^{max}(a_b)$ and $\pi_b^{min}(a_b)$ be the maximum and minimum payoffs that theses worker can then give their bosses. These payoffs depend on the control choice a_b. Since there is a finite number of possible actions, all maximum and minimum payoffs are well defined. Define $\pi_b^e(a_b) = \big(\pi_b^{max}(a_b) + \pi_b^{min}(a_b)\big)/2$ as the resulting equitable payoff. Then

$$k_{wb}(a_b, a_w) = \frac{\pi_b(a_w) - \pi_b^e(a_b)}{\pi_b^{max}(a_b) - \pi_b^{min}(a_b)} \tag{1}$$

describes the kindness of workers with effort choice a_w towards bosses with control choice a_b as perceived by workers when they make their effort choice. Since workers can always choose between at least two different effort levels that yield different payoffs, the denominator of the above fraction always differs from zero so that the above expression is well defined. Note that the kindness of an effort choice depends on the set of feasible payoff combinations and thus on the control choice.

Consider the kindness of certain control choices made by bosses. It depends on what payoff bosses expect to give to workers with their control choice. Beliefs now matter since they determine expected payoff consequences. Furthermore, reciprocal and selfish workers might respond differently to control choices. Certain control choices can therefore be rather kind to selfish workers, and at the same time rather unkind to reciprocal workers. The crucial assumption is that the kindness of an action depends on the information that the individual holds when taking that action. Bosses cannot know their workers' types, and this is common knowledge among bosses and workers. The kindness of bosses' control choices is therefore the expected kindness towards workers. More specifically, bosses first assess the kindness of an action towards workers with a particular type, that is, towards workers with particular effort responses. The expected kindness is the sum of the type-dependent kindnesses towards selfish and reciprocal workers, weighted with the respective probabilities with which bosses believe to be facing these types of workers, or with which workers believe bosses to believe to be facing these types of workers.

[6]Following Rabin (1993) only those alternative payoff combinations are taken into account that are Pareto-efficient conditional on equilibrium play. If in equilibrium an action increases the payoffs of both worker and boss, it must thus be considered kindness neutral. Strict monotonicity implies that all payoff combinations are Pareto-efficient when considering pure strategies; but see the discussion in Section 5.

Let β_w be the belief of bosses concerning the strategy of workers. Let $\pi_w^{max}(\beta_w, \theta)$ be the maximum payoff bosses believe to be able to give to workers with type θ if bosses hold belief β_w concerning workers' strategy. Define $\pi_w^{min}(\beta_w, \theta)$ and $\pi_w^e(\beta_w, \theta)$ as the respective minimum and equitable payoff. Then

$$k_{bw}(a_b, \beta_w, \theta) = \frac{\pi_w(\beta_w(\theta, a_b)) - \pi_w^e(\beta_w, \theta)}{\pi_w^{max}(\beta_w, \theta) - \pi_w^{min}(\beta_w, \theta)} \tag{2}$$

is the kindness of bosses towards workers as perceived by bosses taking control action a_b if workers have type θ while bosses hold belief β_w concerning workers' strategy. If the above denominator is zero, bosses expect workers to always get the same payoff no matter what they do. There is no room for kindness, and kindness is normalized to zero. Given belief β_w concerning workers' strategy, bosses compute their kindness $k_{bw}(a_b, \beta_w, \theta)$ to workers with type θ conditional on their control choice a_b. The expected kindness of bosses imposing control choice a_b is then

$$E_\theta\, k_{bw}(a_b, \beta_w, \theta) = \mu\, k_{bw}(a_b, \beta_w, r) + (1 - \mu)\, k_{bw}(a_b, \beta_w, s) \tag{3}$$

where expectations are formed with probability $\mu \in [0, 1]$ with which bosses believe workers to be reciprocal.[7]

The kindness of bosses towards workers as perceived by workers follows the above definitions. As in Dufwenberg and Kirchsteiger (2004) all observed control choices are considered to be fully deliberate and intentional; this important assumption and the resulting updating of the kindness of bosses is further discussed below. This perceived kindness depends on the belief of workers γ_w concerning the beliefs of bosses concerning the workers' strategy. Then

$$k_{bw}(a_b, \gamma_w, \theta) = \frac{\pi_w(\gamma_w(\theta, a_b)) - \pi_w^e(\gamma_w, \theta)}{\pi_w^{max}(\gamma_w, \theta) - \pi_w^{min}(\gamma_w, \theta)} \tag{4}$$

is the kindness of bosses towards workers with type θ as perceived by workers if workers believe bosses to hold belief γ_w concerning their strategy. The maximum, minimum, and equitable payoffs are defined analogously to the above. The above kindness is set to zero if the denominator equals zero. Since workers know that bosses do not know workers' types, the expected kindness of bosses taking action a_b is

$$E_\theta\, k_{bw}(a_b, \gamma_w, \theta) = \nu\, k_{bw}(a_b, \gamma_w, r) + (1 - \nu)\, k_{bw}(a_b, \gamma_w, s) \tag{5}$$

[7] An alternative approach would be to consider type-dependent effort choices as mixed strategy, and to compute the kindness towards workers as if all worker were using that particular mixed strategy. Results are largely robust to this alternative specification; see the discussion in Section 5.

where expectation are formed with probability $\nu \in [0,1]$ with which workers believe bosses to believe that workers are reciprocal.

Finally, workers know the control choices of their bosses when making their effort choices. Bosses take this into account and assess the expected kindness of workers conditional on their own control choice.[8] This implies that if bosses choose to control their workers, the expected kindness of their workers does not depend on workers' effort choices if bosses had chosen not to control their worker. The expected kindness of workers towards bosses with control choice a_b then equals

$$E_\theta\, k_{wb}\big(a_b, \beta_w(\theta, a_b)\big) = \mu\, k_{wb}\big(a_b, \beta_w(r, a_b)\big) + (1-\mu)\, k_{wb}\big(a_b, \beta_w(s, a_b)\big). \tag{6}$$

The kindness of workers' effort choices follows from (1) and does not depend on any beliefs. It is thus not necessary for bosses to form beliefs about the probability with which workers believe to be facing reciprocal or selfish bosses. Expectations are formed with probability $\mu \in [0,1]$ with which bosses believe workers to be reciprocal.

Selfish and Reciprocal Preferences

It is now possible to define selfish and reciprocal preferences. Given action a_b and the involved beliefs μ and β_w define

$$U_b(a_b, \beta_w, \mu) = E_\theta\, \pi_b\big(\beta_w(\theta, a_b)\big) + \eta\, E_\theta\, k_{wb}\big(a_b, \beta_w(\theta, a_b)\big)\, E_\theta\, k_{bw}(a_b, \beta_w, \theta) \tag{7}$$

as the expected utility of reciprocal bosses. Parameter $\eta \in \mathbb{R}^+$ characterizes the relative importance of reciprocity concerns as compared to monetary payoffs. Reciprocal bosses thus care for their expected monetary payoff. But they also care for fairness. However, they do not know workers' responses to their control choices. They thus compute the expected kindness of workers, and multiply it with their expected kindness towards workers, all conditional on the considered control choice.

Equally, the expected utility of workers given effort choice a_w, control choice a_b by their bosses, and beliefs γ_w and ν concerning the beliefs of bosses is

$$U_w(a_w, a_b, \gamma_w, \nu) = \pi_w(a_w) + \eta\, E_\theta\, k_{bw}(a_b, \gamma_w, \theta)\, k_{wb}(a_b, a_w)\,. \tag{8}$$

Reciprocal workers thus care for their own payoff, but they also care for fairness. All reciprocal individuals put equal relative weight η on their reciprocity concerns. Selfish bosses and

[8] In this respect the present paper differs from Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006). Results are not very sensitive to the exact specification of reciprocity; see the discussion in Section 5.

workers only care for their own expected payoffs. The respective utility functions of selfish bosses and workers are characterized by (7) and (8) with η set equal to zero. All individuals are fully rational and maximize expected utility.

Reciprocity Equilibrium

The equilibrium notions of Rabin (1993) and Dufwenberg and Kirchsteiger (2004) are adapted to the present context with incomplete information as follows. First, equilibrium strategies α_b^* and α_w^* maximize expected utility given the others' equilibrium strategies, given bosses' equilibrium first-order beliefs μ^* and β_w^*, and given workers' equilibrium second-order beliefs ν^* and γ_w^*. Second, first-order and second-order beliefs are consistent with prior beliefs and equilibrium strategies. Concerning equilibrium strategies this implies $\alpha_w^* = \beta_w^* = \gamma_w^*$. It further yields $\mu^* = \lambda$ as bosses beliefs must be consistent with the common prior at the beginning of the strategic interaction. This yields $\nu^* = \mu^* = \lambda$ since in addition workers' beliefs must coincide with the beliefs of bosses. Thirdly, the above holds at all decision nodes so that equilibrium cannot be based on unreasonable behavior off the equilibrium path. Fourthly, kindness is updated as the game progresses under the assumption that observed actions have been chosen completely deliberately.

To avoid tedious and uninformative case distinctions, the following cutoff rule is implemented: in case of indifference workers choose the lowest effort, and bosses control their workers. This actually makes it harder to achieve the goal of the study - to find an equilibrium in which some bosses do not control while reciprocal workers reciprocate not being controlled with exerting high effort.

Irrelevance of Signaling

Since selfish and reciprocal bosses might behave differently in equilibrium, workers might update beliefs after they observe their bosses' control choices. The following arguments show that equilibrium behavior is completely independent from the probability with which workers believe their bosses to be reciprocal or selfish. Reciprocal workers care about the kindness of the observed control choice. The latter depends on their beliefs γ_w and ν concerning the beliefs of bosses concerning workers' strategy and types. In equilibrium these beliefs must be consistent with workers' actual strategy α_w and the prior probability λ with which workers are reciprocal. The beliefs γ_w and ν - and thus also the perceived kindness of a control choice - thus do not not depend on the type of bosses. Consequently, any reciprocal effort reaction by workers does not depend on the probability with which workers belief their bosses to be selfish or reciprocal.

Equally, reciprocal bosses care about the expected kindness of workers. That kindness only depends on workers' actions, which by the above argument do not depend on beliefs concerning bosses' types. Selfish bosses only care for the expected effort choice of workers, and selfish workers only care for their own monetary payoff, which depends on their own effort choice. In all cases, it is thus irrelevant with what probability workers believe their bosses to be reciprocal or selfish. For that reason none of the ensuing results is caused by any signaling concerns. Exactly this distinguishes the present analysis from existing theoretical studies on motivational crowding-out.

4 Results

Since the kindness of workers towards bosses does not depend on any beliefs or types, (1) implies the following in any equilibrium.

Lemma 1 (Kindness of Worker) *In any reciprocity equilibrium*

$$k_{wb}(c,m) = -\tfrac{1}{2},\ k_{wb}(c,h) = +\tfrac{1}{2}, \tag{9}$$

$$k_{wb}(nc,\ell) = -\tfrac{1}{2},\ k_{wb}(nc,h) = +\tfrac{1}{2},\ \textit{and} \tag{10}$$

$$k_{wb}(c,m) = \frac{2\,\pi_b(m) - \pi_b(h) - \pi_b(\ell)}{2\big(\pi_b(h) - \pi_b(\ell)\big)} \tag{11}$$

characterize the kindness of workers towards bosses.

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 reciprocity requilibrium*

$$\alpha_w^*(s, nc) = \ell \quad \textit{and} \quad \alpha_w^*(s, c) = m \tag{12}$$

characterize the equilibrium behavior of selfish workers.

Since the behavior of selfish workers is identical in all equilibria, Lemma 2 yields directly $\pi_w^{max}(\alpha_w^*, s) = \pi_w(\ell)$, $\pi_w^{min}(\alpha_w^*, s) = \pi_w(m)$, and $\pi_w^e(\alpha_w^*, s) = \big(\pi_w(\ell) + \pi_w(m)\big)/2$. This implies that controlling a selfish worker is unkind, whereas not controlling a selfish worker is kind. This yields the following.

Lemma 3 (Kindness of Bosses towards Selfish Worker) *In any reciprocity equilibrium*

$$k_{bw}(nc, \alpha_w^*, s) = +\frac{1}{2} \quad \textit{and} \quad k_{bw}(c, \alpha_w^*, s) = -\frac{1}{2} \tag{13}$$

characterize the kindness of bosses towards selfish workers.

Before presenting the main result define

$$\eta_1 = \frac{\pi_w(\ell) - \pi_w(h)}{1 - 2\lambda} \tag{14}$$

and

$$\eta_2 = \frac{2\big(\pi_w(m) - \pi_w(h)\big)\big(\pi_b(h) - \pi_b(\ell)\big)}{(1 - 2\lambda)\big(\pi_b(h) - \pi_b(m)\big)} \tag{15}$$

Further, let λ_1 be the unique solution in $]0, 1[$ that solves

$$\pi_b(m) - \pi_b(\ell) = \lambda_1 \big(\pi_b(h) - \pi_b(\ell)\big) - \eta \frac{1}{2}(1 - \lambda_1)(1 - 2\lambda_1) \tag{16}$$

and define

$$\lambda_2 = \frac{\pi_b(m) - \pi_b(\ell)}{\pi_b(h) - \pi_b(\ell)}. \tag{17}$$

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 Reciprocation) *Consider a reciprocity equilibrium in which reciprocal workers reciprocate not being controlled so that* $\alpha_w^*(r, nc) = h$ *and* $\alpha_w^*(r, c) = m$. *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 $\lambda \leq \lambda_1$. *Reciprocal 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 reciprocal bosses choose to control.*

This result is based on the following intuition. If workers are selfish workers with sufficiently high probability - the ex-ante probability λ is larger than $1/2$ - then controlling workers with unknown type is on average unkind, whereas not controlling workers with unknown type is on average kind. If reciprocal workers sufficiently care for reciprocity - the level parameter η exceeds a certain threshold - then they reciprocate not being controlled by exerting high effort. This equilibrium exhibits the most important qualitative features of motivational crowding-out as observed in the empirical studies.[9]

In contrast to the prominent existing theoretical explanations for motivational crowding-out, signaling plays no role whatsoever in the above characterized reciprocity equilibrium. In fact, in this equilibrium both selfish and reciprocal bosses might make the same control choices. Workers then do not learn anything about their bosses' preferences.

But heterogeneous control behavior can arise in the above reciprocity equilibrium although everybody holds the same equilibrium beliefs concerning the fraction of reciprocal individuals in the population and their respective equilibrium behavior. This contrasts Sliwka (2007) and Ellingsen and Johannesson (2008) who must assume heterogeneous beliefs to rationalize heterogeneous control choices in equilibrium. The argument runs as follows. In the above reciprocity equilibrium bosses' control choices primarily depend on the fraction λ of reciprocal individuals in the population. They also depend on the payoff increase $\pi_b(h) - \pi_b(m)$ that bosses receive if workers exert high rather than low effort. If this payoff increase is high, it pays not to control workers even if only few workers reciprocate with high effort. The

[9]The exact cutoff values for λ and η depend on the specification of reciprocity preferences. In particular, with an alternative specification there might exist similar equilibria even if the majority of workers are reciprocal and exert high effort if not controlled. See the discussion in Section 5.

control choice of reciprocal bosses is also influenced by their reciprocity concerns and the expected kindness of workers. If bosses exert control, workers are unkind and shirk. Since controlling is equally unkind, it generates reciprocal bosses some utility kick from spite. Not controlling workers is kind, but typically not reciprocated. The unkind negative reaction to a kind control choice hurts reciprocal bosses. In consequence, reciprocal bosses have stronger incentives to control workers than selfish bosses. Selfish and reciprocal bosses might thus make different control choices in equilibrium.

All Other Pure-Strategy Reciprocity Equilibria

It turns out that in the present context there exist only rather intuitive equilibria in pure strategies that depend in an obvious way on the level of reciprocity concerns. This section describes the intuition for these equilibria; the technical details are available upon request.

Proposition 1 describes the reciprocity equilibrium for high levels of reciprocity. For medium levels of reciprocity there can exist an equilibrium in which reciprocal workers always exert medium effort. Not exerting controlling is then kind to selfish workers, and kindness neutral to reciprocal workers. Reciprocity levels have to be strong enough so that uncontrolled reciprocal workers do not exert low effort, but they must not be too strong so that uncontrolled reciprocal workers do not exert high effort. Selfish bosses always control since not controlling workers never increases effort, not even for reciprocal workers. Reciprocal bosses also always control. The reason is that exerting no control is a kind action that is not well reciprocated. This reduces the utility of reciprocal bosses. Controlling is an unkind action, that triggers an unkind effort reaction. Exerting control thus generates a positive utility kick from spite. As monetary and reciprocity incentives are aligned, reciprocal bosses control.

For low levels of reciprocity reciprocal and selfish workers behave in the same way: they choose the minimum feasible effort. Not controlling is kind to both reciprocal and selfish workers. However, reciprocity concerns are so weak so that even being unambiguously kind cannot trigger effort choices that exceed low effort. All bosses exert control: selfish bosses have monetary incentives to control, and reciprocal bosses have additional reciprocity incentives to control workers in order to be spiteful.

There exist no other reciprocity equilibria in pure strategies. The reason is that reciprocal workers must then exert high effort once controlled. This can only be optimal if controlling workers is considered to be kind. The latter is never the case for selfish workers. It also cannot be kind to reciprocal workers for the following reason. Reciprocal workers cannot exert more than high effort if they are controlled. Thus, exerting control is at most kindness neutral to reciprocal workers. But then reciprocal workers maximize their payoffs and do not exert more than the minimum effort once controlled. This demonstrates that the theory possesses some predictive power as certain kind of behavior is ruled out in equilibrium: reciprocal workers never exert more than the minimum effort once controlled.

5 Discussion

This section discusses some aspects of the current model specification. It shows that the main result - in equilibrium some workers might respond to control by lowering their effort - is quite robust with respect to the exact specification of reciprocity preferences.

Viewing Heterogeneous Behavior as Mixed Strategy

One assumption of the present specification is that the kindness of bosses is assessed by (i) deriving their kindness towards workers with a particular type, and (ii) forming the average kindness over all types. This appears natural given that workers differ in their types while different types might use different strategies. Alternatively, one could simply treat all workers the same while assuming that workers employ some mixed strategy. The kindness of a control choice then depends on the expected payoffs to workers of unknown type, ignoring that different types of workers differ in their behavior and thus receive different payoffs. The following arguments show that the resulting specification of reciprocity does not change results fundamentally. Suppose workers use a mixed strategy analogous to the type-dependent behavior from Proposition 1. Bosses then expect to give workers $\lambda\pi_w(h)+(1-\lambda)\,\pi_w(\ell)$ if they choose no control, and $\pi_w(m)$ if they choose control. The logic of intention-based reciprocity implies that exerting no control should be considered kind if and only if λ is smaller than some cutoff value. Note that there might then exist an equilibrium in which the majority of workers reciprocates not being controlled with high effort.

Although results are similar under both specification of reciprocity, there is one substantial difference. The present specification only considers Pareto-efficient payoff combinations to compute the maximum and minimum payoffs that bosses believe to be able to give to workers.

If not controlling is kind, workers receive a higher expected payoff if they are not controlled than if they are controlled. But bosses might also receive a higher expected payoff if they do not control than if they control. Exerting no control then Pareto-dominates exerting control, and must be assessed as kindness neutral. There then exists no reciprocity equilibrium in which reciprocal workers always exert high effort if they are not controlled.

Kindness of Workers Not Conditional On Control Choices

Another aspect of the current specification is that bosses assess workers' kindness conditional on their own control choice. Workers' kindness as perceived by bosses then does not depend on workers' behavior given hypothetical control choices. This complies with the idea that the kindness of an action should depend on the information available to the individual choosing the action. Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006) propose a different specification: bosses compute the kindness of workers at the moment of their control choice by taking into account the induced equilibrium payoffs. This kindness is invariant and does not change when contemplating hypothetical control choices. Further, when different types of bosses make different control choices, workers' kindness is affected by their effort responses to the control choices of other types of bosses.

Using this alternative specification yields very similar results. The reason is that the behavior of workers does not change as here specifications coincide. Concerning bosses, consider the equilibrium in which reciprocal workers exert high effort if they are not controlled. Workers are initially considered to be unkind: they are unkind if bosses choose control, and since only a minority reciprocates, they are also on average unkind if bosses choose not to control. Reciprocal bosses then have additional incentives to be unkind, and are more inclined to control than selfish bosses. Qualitative equilibrium characteristics are unchanged.

Falk and Fischbacher (2006)

The present model also differs from Falk and Fischbacher (2006) because they assume that reciprocity preferences are influenced by payoff inequality. Workers then consider a control choice as kind if and only if they consequently receive higher payoffs than their bosses. The following argument demonstrates that using this definition of reciprocity does not change results fundamentally. Suppose reciprocal workers reciprocate not being controlled with high effort. Reciprocal workers consider not being controlled as kind if the resulting payoff difference $\pi_w(h) - \pi_b(h)$ is positive. But since they then exert only medium effort, being controlled is even kinder to reciprocal workers. The reason is that the payoff difference

$\pi_w(m) - \pi_b(m)$ exceeds $\pi_w(h) - \pi_b(h)$. Further, since $\pi_w(h)$ is the minimum payoff bosses can grant workers in equilibrium, the kindness of choosing no control is further reduced to capture intentions. Exerting no control is therefore less kind towards reciprocal workers than exerting control. Selfish workers choose low effort if not controlled and medium effort if controlled. Towards selfish workers exerting no control is then kinder than exerting control. If the kindness of bosses as perceived by workers is the weighted sum of the kindness of bosses towards selfish and reciprocal workers, then exerting no control can be considered as kind if and only if the fraction of reciprocal workers λ is below some cutoff. The present equilibrium arguments then apply. However, the cutoff for λ depends on payoffs and can thus differ from 1/2. There might then exist an equilibrium in which a majority of workers exerts high effort when not being controlled.

Segal and Sobel (2007)

Segal and Sobel (2007, pp.206-209) take a rather different approach towards reciprocity. They argue that the kindness of a control choice depends on the resulting maximum feasible payoff for workers. This definition immediately implies that exerting no control is kinder than exerting control. The reason is that exerting no control provides workers with the option to achieve a higher payoff than the maximum they could get if controlled. But this definition of kindness - kindness depends on what workers could do, but not an what they actually do - implies that a control choice is considered to be kind even if the subsequent equilibrium result is unfavorable for workers. Intention-based reciprocity does not suffer from this drawback, since the definition of kindness is based on resulting payoffs. The main contribution of the present paper is to show that such intention-based reciprocity might play a role in the crowding-out of motivation.

Battigalli and Dufwenberg (2007)

Schnedler and Vadovic (forthcoming) sketch how motivational crowding-out might be caused by guilt aversion in the spirit of Battigalli and Dufwenberg (2007). Bosses could expect - because of social norms or the framing of the decision environment - that workers exert less effort if controlled than if not controlled. Guilt averse workers match these expectations, so that behavior and expectations are consistent in equilibrium. However, in this application of psychological game theory it can be unclear what constitutes reasonable expectations. In particular, there might also exist an equilibrium in which bosses rightly expect workers to exert higher effort if controlled than if not controlled. Rabin (1993, p.1285) argues that one advantage of intention-based reciprocity is that it derives all behavioral components -

the psychological game - from the material payoffs of the underlying strategic interaction. Section 4 shows that this approach imposes structure which rules out unreasonable equilibria. The present analysis thus demonstrates that despite restrictions on reasonable equilibrium behavior, intention-based reciprocity might well explain motivational crowding-out if there is incomplete information on individual reciprocity preferences.

6 Conclusion

The present paper shows that intention-based reciprocity can explain the crowding-out of motivation if individuals differ in their propensity for reciprocity concerns while preferences are private information. The main argument is that if many 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. This complements the existing theoretical literature on extrinsic incentives and motivation by offering an explanation for motivation crowding-out that is not based on signaling motives. The present paper thereby adds to a small but growing economic literature that applies models of reciprocity. For example, Englmaier and Leider (2008) study optimal contracts and organizational structure in the presence of moral hazard if workers are reciprocal, von Siemens (2009) shows how reciprocity can improve investment incentives in a hold-up situation, and İriş and Santos-Pinto (2010) derive conditions under which reciprocity among managers facilitates or complicates collusion in the product market. The application of intention-based reciprocity to economic problems might thus offer a fruitful avenue for future investigation.

Appendix

Proof of Proposition 1

The equilibrium behavior of reciprocal workers implies $\pi_w^{max}(\alpha_w^*, f) = \pi_w(m)$ and $\pi_w^{min}(\alpha_w^*, f) = \pi_w(h)$ so that $\pi_w^e(\alpha_w^*, f) = \big(\pi_w(m) + \pi_w(h)\big)/2$. Concerning the kindness of a boss towards a reciprocal 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

$$E_\theta k_{bw}(c, \alpha_w^*, \theta) = \lambda - 1/2 \quad \text{and} \quad E_\theta k_{bw}(nc, \alpha_w^*, \theta) = 1/2 - \lambda. \tag{18}$$

while

$$E_\theta k_{wb}(nc, \alpha_w^*(\theta)) = \lambda - 1/2 \quad \text{and} \quad E_\theta k_{wb}(c, \alpha_w^*(\theta)) = -1/2. \tag{19}$$

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 the cutoff rule the behavior of reciprocal workers is optimal if and only if

$$\pi_w(h) + \eta(1/2 - \lambda)(1/2) > \pi_w(\ell) + \eta(1/2 - \lambda)(-1/2) \tag{20}$$

$$\pi_w(h) + \eta(1/2 - \lambda)/2 > \pi_w(m) + \eta(1/2 - \lambda)(2\pi_b(m) - \pi_b(h) - \pi_b(\ell))/(\pi_b(h) - \pi_b(\ell)) \tag{21}$$

$$\pi_w(h) + \eta(1/2 - \lambda)(1/2) \geq \pi_w(h) + \eta(\lambda - 1/2)(1/2). \tag{22}$$

Constraints (20) and (21) ensure that if not controlled, the worker prefers to exert high rather than medium or low effort. (22) ensures that if controlled, the worker prefers to exert medium rather than high effort. As $\pi_w(h) < \pi_w(\ell)$ constraint (20) can hold only if $\lambda < 1/2$. This condition with $\pi_w(m) > \pi_w(h)$ implies that (22) can be ignored. Rearranging (20) yields as condition $\eta > \eta_1$ and (21) yields $\eta > \eta_2$ with the cutoffs η_1 and η_2 as defined above. Note that since $\pi_b(h) - \pi_b(\ell) > \pi_b(h) - \pi_b(m)$ it is not clear whether (20) or (21) is binding even though $\pi_w(\ell) - \pi_w(h) > \pi_w(m) - \pi_w(h)$.

Consider next the optimality of the equilibrium behavior of bosses. Given the cutoff rule selfish bosses control in equilibrium if and only if $\pi_b(m) \geq \lambda\pi_b(h) + (1 - \lambda)\pi_b(\ell)$ or

$$\pi_b(m) - \pi_b(\ell) \geq \lambda\big(\pi_b(h) - \pi_b(\ell)\big) = A(\lambda). \tag{23}$$

This yields as condition $\lambda \leq \lambda_2$ with λ_2 as defined above which directly yields $\lambda_2 \in]0, 1[$. Given the cutoff rule reciprocal bosses control in equilibrium if and only if

$$\pi_b(m) - \pi_b(\ell) \geq \lambda\big(\pi_b(h) - \pi_b(\ell)\big) + \eta(1 - \lambda)(\lambda - 1/2) = B(\lambda). \tag{24}$$

Since $B(0) = -\eta/2 < \pi_b(m) - \pi_b(\ell)$ and $B(1) = \pi_b(h) - \pi_b(\ell) > \pi_b(m) - \pi_b(\ell)$ continuity of B on $[0, 1]$ and the intermediate value theorem imply that there exists $\lambda_1 \in]0, 1[$ such that $B(\lambda_1) = \pi_b(m) - \pi_b(\ell)$. Further, function B is strictly concave on $[0, 1]$ since $B''(\lambda) = -2\eta$. It is thus quasi-concave and the upper contour sets $P(x) = \{\lambda \in [0, 1] : B(\lambda) \geq x\}$ are compact intervals. $B(1) > \pi_b(m) - \pi_b(\ell)$ implies $1 \in P\big(\pi_b(m) - \pi_b(\ell)\big)$ which yields that $P\big(\pi_b(m) - \pi_b(\ell)\big) = [\lambda_1, 1]$. Convexity implies $B(\lambda) < \pi_b(m) - \pi_b(\ell)$ for all $\lambda < \lambda_1$. Strict concavity of B further implies that $B(t\lambda_1 + (1 - t)) > tB(\lambda_1) + (1 - t)B(1) > \pi_b(m) - \pi_b(\ell)$ for all $t \in]0, 1[$. Thus, there exists a unique λ_1 in $]0, 1[$ that solves (16) with equality so that reciprocal bosses do not control in equilibrium if and only if $\lambda > \lambda_1$.

Finally, suppose $\lambda_1 < 1/2$. Then $A(\lambda) > B(\lambda)$ and therefore $B(\lambda_1) < A(\lambda_1) = \pi_b(m) - \pi_b(\ell)$. Further, $B(1/2) = A(1/2) > A(\lambda_1) = \pi_b(m) - \pi_b(\ell)$. Then the above arguments imply that

there exists a unique $\lambda_2 \in]\lambda_1, 1/2[$ with the property $B(\lambda_2) = \pi_b(m) - \pi_b(\ell)$. For $\lambda \in]\lambda_1, \lambda_2[$ optimality of the behavior of selfish and reciprocal workers then follows from the above conditions. *Q.E.D.*

References

Ariely, D., E. Kamenica, and D. Prelec (2008): "Man's Search for Meaning: The Case of Legos," *Journal of Economic Behavior and Organization*, 67(3-4), 671–677.

Barkema, H. G. (1995): "Do Top Managers Work Harder When They Are Monitored?," *Kyklos*, 48, 19–42.

Battigalli, P., and M. Dufwenberg (2007): "Guilt in Games," *American Economic Review (Papers and Proceedings)*, 97(2), 170–176.

——— (2009): "Dynamic Psychological Games," *Journal of Economic Theory*, 144, 1–35.

Bénabou, R., and J. Tirole (2003): "Intrinsic and Extrinsic Motivation," *Review of Economic Studies*, 70, 489–520.

Charness, G., R. Cobo Reyes, N. Jiménez, J. A. Lacomba, and F. Lagos (forthcoming): "The Hidden Costs of Control: Comment," *American Economic Review*.

Dickinson, D., and M.-C. Villeval (2008): "Does Monitoring Decrease Work Effort? The Complementarity between Agency and Crowding-Out Theories," *Games and Economic Behavior*, 63, 56–76.

Dominguez-Martinez, S., R. Sloof, and F. A. von Siemens (2010): "Monitoring Your Friends, Not Your Foes: Strategic Ignorance and the Delegation of Real Authority," CESifo Working Paper Series No. 3172.

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

Ellingsen, T., and M. Johannesson (2008): "Pride and Prejudice: The Human Side of Incentive Theory," *American Economic Review*, 98(3), 990–1008.

Englmaier, F., and S. Leider (2008): "Contractual and Organizational Structure with Reciprocal Agents," *CESifo Working Paper No. 2415*.

Enzle, M. E., and S. C. Anderson (1993): "Surveillant Intentions and Intrinsic Motivation," *Journal of Personality and Social Psychology*, 64(2), 257–266.

Falk, A., and U. Fischbacher (2006): "A Theory of Reciprocity," *Games and Economic Behavior*, 54(2), 293–315.

Falk, A., and M. Kosfeld (2006): "The Hidden Costs of Control," *American Economic Review*, 96(5), 1611–1630.

Foss, N. J. (2003): "Selective Intervention and Internal Hybrids: Interpreting and Learning from the Rise and Decline of the Oticon Spaghetti Organization," *Organization Science*, 14(3), 331–349.

Frey, B. S., and R. Jegen (2001): "Motivation Crowding Theory," *Journal of Economic Surveys*, 15(5), 589–611.

Herold, F. (2010): "Contractual Incompleteness as a Signal of Trust," *Games and Economic Behavior*, 68, 180–191.

Herzberg, F. (2003): "One More Time: How Do You Motivate Employees?," *Harvard Business Review*, 81(1), 87–96.

İrış, D., and L. Santos-Pinto (2010): "Tacit Collusion under Fairness and Reciprocity," mimeo, University of Lausanne.

Manzoni, J.-F., and J.-L. Barsoux (1998): "The Set-Up-to-Fail Syndrom: How Bosses Create Their Own Poor Performers," *Harvard Business Review*, 76(2), 101–113.

Plant, R. W., and R. M. Ryan (1985): "Intrinsic Motivatin and the Effects of Self-Consciousness, Self-Awareness, and Ego-Involvement: An Investigation of Internally Controlling States," *Journal of Personality*, 53(3), 435–449.

Rabin, M. (1993): "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, 83(5), 1281–1302.

Ryan, R. M., and E. L. Deci (2000): "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being," *American Psychologist*, 55(1), 68–78.

Schnedler, W., and R. Vadovic (forthcoming): "Legitimacy of Control," *Journal of Economics and Management Strategy*.

Segal, U., and J. Sobel (2007): "Tif for Tat: Foundations of Preferences for Reciprocity in Strategic Settings," *Journal of Economic Theory*, 136, 197–216.

Sliwka, D. (2007): "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes," *American Economic Review*, 97(3), 999–1012.

Spier, K. E. (1992): "Incomplete Contracts and Signalling," *RAND Journal of Economics*, 23(3), 432–443.

Suvorov, A., and J. van de Ven (2009): "Discretionary Rewards as Feedback Mechanism," *Games and Economic Behavior*, 67, 665–681.

von Siemens, F. A. (2009): "Bargaining under Incomplete Information, Fairness, and the Hold-Up Problem," *Journal of Economic Behavior and Organization*, 71, 486–494.