Legitimacy of Control
VERY PRELIMINARY, PLEASE DO NOT QUOTE OR CIRCULATE

Wendelin Schnedler*
Department of Economics
University of Heidelberg

Radovan Vadovic
Department of Economics
ITAM Mexico City

June 19, 2007

1 Introduction

It is relatively simple for employers to monitor the email content and internet usage of their employees. Controlling employees in this way has a direct benefit: it prevents unproductive activities such as misuse and slack. Control is hardly complete; employees still have many ways to influence productivity even when their email content and internet usage is monitored. Still, it seems sensible for employers to avert misuse and slack where they can. In a recent experiment, however, Falk and Kosfeld (2006) demonstrate that apart from the direct benefit, control entails *hidden costs*: a principal who ensures that an agent exerts a minimum level of effort obtains on average lower effort than a principal who leaves the choice free. While the findings of Falk and Kosfeld (2006) document the existence of hidden costs, we know little about the circumstances under which hidden costs occur or the determinants of these costs. Here, we enquire into the extent of hidden costs and identify a main factor that influences them.

When companies introduce measures to control the activities of their employees, they often justify this step. As an example take the following quote:[1]

> "Sometimes it is necessary to monitor employee personal communications or computer usage or to search employee workspaces for the protection of employees, company assets and other legitimate business reasons. [The company] retains the right to monitor personal communications or computer usage or to search any and all computer property at any time, including, but not limited to, offices, desks, lockers, bags, vehicles, e-mail, voice mail, pagers, ... [company] telephone usage records and computer files."

This justification encompasses two key elements. First, the mere existence of a general code of conduct emphasizes that control is not targeted at a specific employee. Second, it highlights that control prevents loss or damage of company assets. Both issues may play a role in assuring the cooperation of the employee (agent) and reducing hidden costs of control. To summarize the two issues, we

*University of Heidelberg, Department of Economics, Grabengasse 14, 69117 Heidelberg (wendelin.schnedler@awi.uni-heidelberg.de).

[1] The quote is taken from the Verizon Code of Conduct (2001).

call control *legitimate* (i) if it is not exclusively aimed at the agent or (ii) if it protects the endowment of the principal rather than forces the agent to give up some of his endowment. Our central hypothesis is that control leads to lower hidden costs of control if it is legitimate and to higher hidden costs when it is not.

In the setting examined by Falk and Kosfeld (2006), control is not legitimate: it is exclusively aimed at the agent and forces him to exert effort. In order to study hidden costs of control when control is legitimate, we consider two variations of their setting. In the first, the principal (e.g. the employer) does not know whether she faces the agent or someone who exerts minimal effort. In practice, this someone could be another worker. In order to clarify our point and maintain experimental control, we represent this someone by a pre-programmed robot. In the second variation, control does not touch the agent's endowment but protects the endowment of the principal. Both variation as well as the central treatments of Falk and Kosfeld belong to a larger class of games, which we label *control games*.

If we want to formalize the idea of legitimacy of control and its effect on hidden costs in control games, we need an explanation why hidden costs of control arise in the first place. Hidden costs do not occur when the agent is selfish, inequity averse (in the sense of Ernst Fehr and Klaus Schmidt 1999), cares about payoff-intentions (see e.g. (Matthew Rabin, 1993; Martin Dufwenberg and Georg Kirchsteiger, 2004; Armin Falk and Urs Fischbacher, 2006) or for any mixture of these types. Tore Ellingsen and Magnus Johannesson (2007) provide a sufficiently rich model to account for hidden costs of control that includes different types of principals as well as an agent with belief-dependent preferences. Here, we follow an idea suggested by Falk and Kosfeld (2006) and simply suppose that the agent is guilt averse (see Dufwenberg and Charness (2006), Battigalli and Dufwenberg (2005)). The intuition for hidden costs then is the following. Control by the principal signals her low (pessimistic) expectations about the effort of the agent. If the agent is guilt averse, than he has incentives to meet these expectations and would indeed choose a lower effort than if left uncontrolled. We examine this argument more carefully and use an explicit model to demonstrate that this is not the only equilibrium behavior in control games with guilt-averse agents. We then propose a *legitimacy criterion* that reflects our idea that control is legitimate if it protects assets or is not directly aimed at the agent. This criterion helps us select what we believe is the most reasonable behavior and will imply hypothesis that we then test experimentally.

Applying the legitimacy criterion to the two situations that we described earlier, we obtain the following predictions. In these situations when controlling is legitimate, it should not be costly. The reason is that in contrast to the Falk and Kosfeld setup, where controlling action had a clear interpretation as a principal's concern of receiving a low effort from the agent, in our framework there is another (legitimate) purpose for controlling. Namely, the protection of one's own property in the first case and a protection against low effort from an external source (e.g., a different, lazy agent) in the second case. In both cases, if the agent finds controlling legitimate, i.e., does not perceive it as a signal of low expectations by the principal, then this would lead him to choose the same effort in both cases: when controlled and when not controlled. In other words, the hidden costs of control would vanish. This is indeed what we observe in our experiments.

The presence of a guilt-averse agent is not the only possible explanation for the observation that the agent provides lower effort when he is controlled. For example, the agent may exhibit some form of indirect reciprocity and punish the principal not because she lowered his payoff but because she restricted his options. Based on this idea, we would predict costs of control to be present even if it prevents theft or when it is not exclusively aimed at the agent. Another explanation is that control is a signal of the principal's type. Control may for example indicate that the principal herself would not provide voluntary effort as an agent. A non-selfish agent may want to punish such a selfish principal while rewarding a principal who –like him– is non-selfish. Notice that this idea cannot explain hidden costs as equilibrium behavior because selfish principals would have an incentive to imitate non-selfish principals and obtain more effort by not controlling.

The remainder of the paper is organized as follows. Section 2 provides a model to explain hidden costs of control, introduces a formal notion of legitimacy, and presents the resulting predictions. Section 4 describes the experimental and econometrical procedures while Section 5 lists our results. Finally, Section 6 summarizes our findings.

2 Explaining hidden costs of control

In this section, we describe a class of games that encompasses most of the treatments of Falk and Kosfeld (2006). We show that if the agent is sufficiently guilt-averse, there are two types of equilibria: open-door equilibria, where the principal does not control, and closed-door equilibria, where she controls. Based on the idea of legitimacy, we then introduce a refinement that for some parameter values eliminates the open-door equilibria. Given our refinement, we then predict no hidden costs of control for these values.

2.1 Players, strategy space, and monetary payoffs

The players in the game are a risk-neutral principal (she) and an agent (he). The principal decides whether to impose a minimal effort requirement. The agent chooses an effort, x, where x is limited by an initial positive endowment π_{A}^{0}: $x \leq \pi_{\mathrm{A}}^{0}$. Effort costs the agent x but leads to a benefit of $2x$ for the principal. Extending the game by Falk and Kosfeld, we want to allow for the possibility that the principal's action is not specifically aimed at the agent. We do this by introducing a chance move that determines whether the principal interacts with the agent or with a selfish robot that always chooses the minimal effort. Let us summarize the sequence. First, the principal decides on the minimal effort requirement $y \in \{x_{\mathrm{L}}, \underline{x}\}$, where $x_{\mathrm{L}} \leq 0$ and $x_{\mathrm{L}} < \underline{x} < \pi_{\mathrm{A}}^0$. We say that the principal *controls* if she requires $y = \underline{x}$ and does *not control* otherwise. Second, nature decides whether the principal interacts with the agent or receives the minimal effort requirement y, where $p > 0$ denotes the probability of interaction with the agent. If principal and agent interact, the agent chooses an effort level x from the set $\left\{y, \ldots, \pi_{\mathrm{A}}^0\right\}$. The strategy of the agent is a pair $(x^{\mathrm{nc}}, x^{\mathrm{c}})$ which specifies an effort choice for each of the two possible choices by the principal. In order to distinguish these choices, we define a label function:

$$s(y) = \begin{cases} \mathrm{nc} & \text{if} \quad y = \underline{x} \\ \mathrm{c} & \text{if} \quad y = x_{\mathrm{L}}. \end{cases}$$

The effort choice of the agent is then denoted by $x^{s(y)}$, i.e. x^{c} when the principal controlled the agent ($y = \underline{x}$) and x^{nc} when the principal did not control ($y = x_{\mathrm{L}}$). Given that principal and agent interact, the monetary payoffs are:

$$\pi_{\mathrm{A}}^1(y, x) = \pi_{\mathrm{A}}^0 - x \text{ for the agent and} \tag{1}$$
$$\pi_{\mathrm{P}}^1(y, x) = \pi_{\mathrm{P}}^0 + 2 \cdot x \text{ for the principal,} \tag{2}$$

where $\pi_{\mathrm{P}}^0 \geq 0$ is the initial endowment of the principal. If the principal does not interact with the agent, the agent has no effort costs and his payoff is the initial endowment. Since the selfish robot does not exert any voluntary, the principal gets only the payoff from the required effort. Summarizing, the monetary payoffs when principal and agent do not interact are:

$$\pi_{\mathrm{A}}^1(y, x) = \pi_{\mathrm{A}}^0 \text{ for the agent and} \tag{3}$$
$$\pi_{\mathrm{P}}^1(y, x) = \pi_{\mathrm{P}}^0 + 2 \cdot y \text{ for the principal.} \tag{4}$$

Given specific parameter values and a particular strategy of the agent, the principal may be indifferent between controlling or leaving the choice free. This occurs, for example, if an agent encounters the principal with certainty $p = 1$ and does not condition effort on control. As a rule to break the indifference, we assume that the principal controls when she is indifferent.[2]

Notice that the treatments considered by Falk and Kosfeld (2006) are special cases of the described setting.[3] In their treatments, principal and agent interact with certainty ($p = 1$) and the minimum requirement is larger than zero ($\underline{x} > 0$). This implies that control is always aimed specifically at the agent and that it always forces the agent's payoff below his endowment. The class of games considered here is broader and later allows us to describe and experimentally study determinants of costs of control.

2.2 Hidden costs of control

Before we introduce behavioral assumptions that generate hidden costs of control, we need to define formally what we mean by hidden costs of control. If

[2] The assumption can be justified by a continuity argument in the example: for a probability of interaction that is slightly lower than one ($p < 1$), the principal strictly prefers to control.

[3] The only exception is the gift-exchange treatment GE10.

the principal controls the agent, the overall effect of control on the principal's payoff is $2 \cdot (x^{\mathrm{c}} - x^{\mathrm{nc}})$. It is possible to distinguish between two effects of control. The first effect is direct: when the principal controls the agent, the agent can no longer choose effort below the minimum requirement $\underline{x}$. This direct effect of control implies that the principal's payoff increases by $2 \cdot (\underline{x} - x^{\mathrm{nc}})$ but only if the agent exerts effort below the required level when he is not controlled ($x^{\mathrm{nc}} < \underline{x}$). Otherwise control has no direct effect on the payoff. The second effect is indirect or behavioral. It is indirect in the sense that the agent responds to control without being forced to do so. For example, the agent may exert an effort above the minimum requirement when he is not controlled ($x^{\mathrm{nc}} > \underline{x}$) and just fulfil the requirement when being controlled ($x^{\mathrm{c}} = \underline{x}$). Or he may voluntarily increase effort when being controlled ($x^{\mathrm{c}} > x^{\mathrm{nc}} > \underline{x}$). These and other voluntary responses by the agent make up for the remainder of the overall effect of control. The indirect effect is thus the overall effect of control on the principal's payoff minus the direct effect:

$$2 \cdot (x^{\mathrm{c}} - x^{\mathrm{nc}}) - 2 \cdot \left\{ \begin{array}{lll} \underline{x} - x^{\mathrm{nc}} & \text{if} & x^{\mathrm{nc}} < \underline{x} \\ 0 & \text{if} & x^{\mathrm{nc}} \geq \underline{x} \end{array} \right\}. \tag{5}$$

Unlike the direct effect of control, which is never harmful to the principal, this indirect effect may be positive, zero, but also negative. In the latter case, we speak of *hidden costs of control*. This definition leads us immediately to a necessary and sufficient condition for the presence of hidden costs of control:[4] the effort under no control has to be larger than with control ($x^{\mathrm{nc}} > x^{\mathrm{c}}$). Equipped with this condition, we can now examine under which circumstances hidden costs of control occur.

2.3 Guilt-aversion

Guilt aversion provides a possible explanation for why an agent works less when he is controlled: by controlling, the principal conveys her low expectations about the effort of the agent who then feels less guilty supplying such low effort. In what follows, we examine this argument more carefully. Our model is very basic but rich enough to explain costs of control.[5] Let $\alpha = (\alpha^{\mathrm{c}}, \alpha^{\mathrm{nc}})$ represent the principal's (first-order) belief about the agent's strategy conditional on control and no control. Let β represent the agent's second-order belief about the principal's first-order belief α.[6] Denote the first moments of β by $\mu = (\mu^{\mathrm{c}}, \mu^{\mathrm{nc}})$. So, μ^{c} is the effort that the agent believes the principal expects from him under control and μ^{nc} that under no control.

The presence of a guilt averse agent implies that the agent has a utility function which depends on his belief about what the principal expects him to do. If he does not meet the principal's expectations then he suffers from guilt.

[4] If $x^{\mathrm{nc}} > x^{\mathrm{c}}$ it follows that $x^{\mathrm{nc}} > \underline{x}$ and hence the indirect effect is negative. Conversely if $x^{\mathrm{c}} \geq x^{\mathrm{nc}}$ then the indirect effect is non-negative because $x^{\mathrm{c}} - x^{\mathrm{nc}} \geq \underline{x} - x^{\mathrm{nc}}$.

[5] For much deeper and more complete discussion of guilt-aversion, see Battigalli and Dufwenberg (2006).

[6] Higher order beliefs could be defined accordingly but are not needed for our analysis.

Let G_{A} represent the extent to which the agent is "hurt" if he does not meet the principal's expectations. More specifically, let the guilt of the agent for a given belief β and a strategy profile, $(y, x^{s(y)})$ be measured by:

$$\begin{aligned} G_{\mathrm{A}}\left((y, x^{s(y)}) \mid \beta\right) &= \theta \max\left[0, (\pi_{\mathrm{P}}^0 + 2\mu^{s(y)}) - (\pi_{\mathrm{P}}^0 + 2x^{s(y)})\right] \\ &= \theta \max\left[0, 2(\mu^{s(y)} - x^{s(y)})\right], \end{aligned}$$

where θ is a parameter to describe the degree of guilt-aversion which we assume to be sufficiently large:[7] $\theta > 1/2$. The overall utility from monetary payoffs and guilt is:

$$\begin{aligned} u_{\mathrm{A}}\left(y, (x^{\mathrm{nc}}, x^{\mathrm{c}}) \mid \alpha_{\mathrm{A}}\right) &= \pi_{\mathrm{A}}^1\left(y, x^{s(y)}\right) - G_{\mathrm{A}} \\ u_{\mathrm{P}}\left(y, (x^{\mathrm{nc}}, x^{\mathrm{c}}) \mid \alpha_{\mathrm{A}}\right) &= \pi_{\mathrm{P}}^1\left(y, x^{s(y)}\right). \end{aligned} \tag{6}$$

By introducing guilt into the agent's utility we have made his utility belief-dependent. This turns the game into a *psychological game* which was first formalized by Geanakoplos, Pearce, and Stacchetti (1989). Our model, however, goes beyond the scope of their framework since we allow not only the initial belief but also an updated belief (the belief after observing the decision of the principal) to enter the agent's utility function. In that sense our analysis is consistent with Battigalli and Dufwenberg (2005) who extend the framework of Geanakoplos, Pearce, and Stacchetti to dynamic games.

2.4 Open and closed-door equilibria

In this section, we analyse behavior in the class of games described in Section 2.1 when the agent is guilt-averse. As an equilibrium concept, we employ the sequential equilibrium introduced by Battigalli and Dufwenberg (forthcoming).[8] There are two different types of equilibria that can occur.

Open-door equilibrium: The principal expects the agent to exert (substantially) more effort if she leaves the choice free and requires no minimum effort. The agent meets the principal's expectations and chooses a higher effort if he is not controlled:

$$x^{\mathrm{nc}} - x^{\mathrm{c}} = \mu^{\mathrm{nc}} - \mu^{\mathrm{c}} > \frac{1-p}{p}(\underline{x} - x_{\mathrm{L}}). \tag{7}$$

In the treatments of Falk and Kosfeld, where the probability of interaction is one, $p = 1$, the condition on beliefs reduces to

$$x^{\mathrm{nc}} - x^{\mathrm{c}} = \mu^{\mathrm{nc}} - \mu^{\mathrm{c}} > 0. \tag{8}$$

[7] If $\theta \leq 1/2$ marginal costs of exerting effort always exceed the marginal guilt and guilt-averse agents act like selfish agents.

[8] This equilibrium is an extension of the sequential equilibrium by Kreps and Wilson for games with guilt-averse players. In any such equilibrium, beliefs are correct. Here, this means $\alpha = \beta$.

Closed-door equilibrium: The principal does not expect the agent to exert (substantially) more effort if she leaves the choice free and requires a minimum effort. The agent meets the principal's expectations and his effort under no control is not much larger than if he is controlled:

$$x^{\mathrm{nc}} - x^{\mathrm{c}} = \mu^{\mathrm{nc}} - \mu^{\mathrm{c}} \leq \frac{1-p}{p}(\underline{x} - x_{\mathrm{L}}). \tag{9}$$

Again, the condition simplifies if principal and agent always interact:

$$x^{\mathrm{nc}} - x^{\mathrm{c}} = \mu^{\mathrm{nc}} - \mu^{\mathrm{c}} \leq 0. \tag{10}$$

What is the relationship between these equilibrium candidates and hidden costs of control? For the case of a certain interaction between principal and agent, we compare conditions (8) and (10) to see that hidden costs of control are only present in the open-door equilibrium. More generally, it can be shown that costs of control are always larger in the open-door than in the closed-door equilibrium (see Proposition 4 in the appendix).

In order to study the determinants of costs of control, we have to answer which equilibrium type arises under which conditions? Unfortunately, both types simultanously exist whenever the probability of interaction is sufficiently large (see Proposition 3 in the appendix). Thus, we have no theoretical guidance with respect to the presence and extent of costs of control.

From inequalities (7) and (9), it can be seen that off-equilibrium beliefs play a crucial role for the type of equilibrium. In the next section, we restrict these off-equilibrium beliefs following our intuition that a certain behavior of the principal may be considered legitimate. This enables us later to eliminate the open-door equilibrium for certain parameter values.

2.5 Legitimacy

We have seen that a guilt-averse agent leads to multiple equilibria that differ in their off-equilibrium beliefs and in the hidden costs of control, respectively. The purpose of this section is to formalize our concept of legitimacy and use it to restrict off-equilibrium beliefs. It is, however, not our aim to provide a general refinement. Rather, we focus on the class of games introduced earlier.

Let us reconsider the Verizon example and our intuition what renders control legitimate in a world where there is no control yet. A preliminary condition seems to be that a worker who does not misuse company's assets for private purposes should not be directly affected by control. We formalize this by requiring that a legitimate action should not force the agent's payoff below the equilibrium payoff, i.e. the payoff without control. While this restricts the direct effect of control on the agent, legitimacy also seems to be driven by indirect effects. Even if the worker is willing to put in effort rather than shirk, control means that his freedom to do with his endowment as he pleases is limited by control. Such a limitation may be regarded as harassment unless Verizon has a specific reason to impose it. This idea is reflected in our two main conditions for legitimacy: either the action does not touch the agent's endowment or it must increase the principal's payoff. We summarize these considerations in the following definition.

Definition 1 (Legitimate Deviation) *Let (y^*, x^*) be a profile of equilibrium actions and associated payoffs π^*_A, π^*_P. Then, a deviation of the principal $\tilde{y}$ is legitimate if and only if there is an action $x \in A(\tilde{y})$ such that*

$$\pi^1_A(\tilde{y}, x) \geq \pi^*_A \text{ and} \tag{11}$$

$$\text{(i) } \pi^0_A \leq \pi^1_A(\tilde{y}, x) \text{ or (ii) } \pi^1_P(\tilde{y}, x) > \pi^*_P. \tag{12}$$

When $\tilde{y}$ violates at last one of these conditions it is not a legitimate deviation.

How does the principal affect the beliefs of the agent if she unexpectedly controls? We belief the answer depends on the legitimacy of control. Suppose the legitimacy conditions are not met and that the agent is simply harassed. Then, the principal can hardly expect the agent to exert the same effort as without control. In other words, beliefs of the agent about the principal's expectations fall when control is enforced without legitimation. On the other hand, the agent probably reacts quite differently if he regards control to be legitimate. Then, the principal may well expect the agent to exert the same effort as before. Accordingly, the beliefs of the agent about the principal's expectations do not fall. We formalize this idea in the following definition.

Definition 2 (Legitimacy Criterion) *Consider an equilibrium (x^*, y^*). This equilibrium fulfills the legitimacy criterion if and only if the agent's beliefs about the principal's expectations $\alpha_A(\tilde{y})$ following a deviation $\tilde{y}$, weakly first-order stochastically dominates the equilibrium belief $\alpha_A(y^*)$ whenever $\tilde{y}$ is legitimate:*

$$\tilde{y} \text{ legitimate } \Leftrightarrow \alpha_A(\tilde{y}) \geq_{FSD} \alpha_A(y^*).$$

Appealing to the moments of the agent's beliefs, we can rephrase the criterion in the following way: whenever a deviation $\tilde{y}$ is legitimate, the expected belief about effort cannot be lower $\mu^{s(\tilde{y})} \geq \mu^{s(y^*)}$. In words, the agent does not believe that the principal expects more voluntary effort of him if (and only if) the principal takes a non-legitimate action. In the next section, we examine under which conditions the open- and closed-door equilibria meet the legitimacy criterion.

2.6 Legitimate behavior by the principal

Deviating from control is always legitimate because the agent can ensure to keep his endowment by playing the lowest possible effort x_{L} (for a fully-fledged argument see Lemma 2 in the appendix). The legitimacy definition thus reflects the idea that it is legitimate if the principal does *not* control the agent. More interesting is the question when it is legitimate that the principal controls the agent. This is the case either if control does not reduce the endowment of the agent ($\underline{x} \leq 0$) or if there is some probability that the principal does not interact with the agent ($p < 1$) (details are in Lemma 3 in the appendix). Recall that for the setting examined by Falk and Kosfeld, the principal always interacts with the agent ($p = 1$) and that control forces the monetary payoff to the agent below his endowment π_{A}^{0}. We can hence use the above considerations to characterize equilibria in their setting (for the complete proof see Appendix A.4):

Proposition 1 (Legitimate equilibria in Falk and Kosfeld's setting) *Suppose that $p = 1$ and $\underline{x} > 0$. Then, the beliefs that satisfy the open-door equilibrium fulfill the legitimacy criterion. The beliefs that supports the closed-door equilibrium satisfy the legitimacy criterion only if $\mu^{c} = \mu^{nc}$.*

In the setting of Falk and Kosfeld, the legitimacy criterion is thus no help to determine which equilibrium is played. However, it allows us to establish that the indirect effect of control on effort is either negative (in the open-door equilibrium) or absent (in the closed-door equilibrium) but never positive. Next, we want to leave the setting of Falk and Kosfeld.

Proposition 2 (Legitimate equilibria) *Suppose that $p < 1$ or $\underline{x} \leq 0$. Then, the open-door equilibrium does not satisfy the legitimacy criterion while the closed-door equilibrium does.*

The complete proof is in Appendix A.5. The proposition implies that beyond Falk and Kosfeld's setting, the legitimacy criterion leads to a unique prediction: the principal controls, any conditioning of effort on control is small and so are costs of control. In the next section, we suggest treatments to test this prediction.

3 Treatments and Predictions

The question whether the open-door or closed-door equilibrium describes the behavior of subjects better and whether the legitimacy criterion can help us select amongst them is an empirical one. The treatments by Falk and Kosfeld (2006) indicate that the open-door equilibrium is better suited to describe behavior in their setting, where principal and agent interact with certainty and control implies that the agent looses some of his endowment. Since beliefs are crucial for the type of equilibrium and the cultural context may hence matter, we replicate their main treatment as our BASELINE treatment. In this treatment (which is called C10), the agent has an initial endowment of $\pi_{\mathrm{A}}^{0} = 120$, while that of the principal is $\pi_{\mathrm{P}}^{0} = 0$. Control ensures an effort of $\underline{x} = 10$ and the agent cannot take from the principal $x_{\mathrm{L}} = 0$ with whom he always interacts $p = 1$. The legitimacy criterion does not yield a unique prediction for the BASELINE treatment but from the experiment of Falk and Kosfeld, we expect the following hypotheses to be true.

Hypothesis 1
(a) There are costs of control in the BASELINE treatment.
(b) Principals leave the choice free in the BASELINE treatment.

Our main focus is on parameter settings where the legitimacy has a unique prediction which we can test. For our second treatment, called ROBOT, we take exactly the same parameter values as in the BASELINE treatment with the exception that interaction no longer takes place with certainty: with probability $p = 0.5$ the principal faces a selfish robot. If the respective parts of Hypothesis 1 are valid, we can use the BASELINE treatment as a benchmark and expect the following cross-treatment effect.

Hypothesis 2
(a) Costs of control are lower in the ROBOT than in the BASELINE treatment.
(b) Principals control more often in the ROBOT than in the BASELINE treatment.

While the ROBOT treatment deviates from the BASELINE treatment by the possiblity that the principal interacts with a selfish robot. Our third treatment, called ENDOWMENT, explores a simple change in endowments. In this treatment, the principal starts out with an endowment of $\pi_{\mathrm{P}}^0 = 20$ while the agent begins with $\pi_{\mathrm{A}}^0 = 110$. The agent can now steal from the principal $x_{\mathrm{L}} = -10$ and control prevents stealing $\underline{x} = 0$. The ENDOWMENT treatment and the BASELINE are identical with the exception of a mere shift in the effort label: an effort of x in the ENDOWMENT treatment leads to the same payoffs as an effort of $x + 10$ in the BASELINE treatment; control by the principal in both cases ensures a payoff of 20. Again, we can use the BASELINE treatment as a benchmark if the respective parts of Hypothesis 1 are valid. This leads to the following cross-treatment effect.

Hypothesis 3
(a) Costs of control are lower in the ENDOWMENT than in the BASELINE treatment.
(b) Principals control more often in the ENDOWMENT than in the BASELINE treatment.

In the next section, we take these hypotheses to the data.

4 Procedures

We ran a total of 12 sessions: 4 for each treatment. All sessions were run in the experimental laboratory at the University of Mannheim. Subjects were primarily undergraduate students who were randomly recruited from a pool of approximately 1000 subjects using an E-mail recruitment system. The software was written in Visual Basic 6 and the experiment lasted for approximately 60 minutes (including reading the instructions and final payments).

After the subjects arrived at the laboratory, they were randomly matched in anonymous pairs and seated at the computer terminals. They were handed instructions (included in the appendix) and had 15 to 20 minutes to study them. After everyone had finished reading, they were asked to complete a series of control questions designed to verify their understanding of the experiment.

When all questions were answered successfully, the experiment began. At the end of the experiment, we paid each subject privately in cash. Each subject received about 10% of their experimental earnings plus €4 show up fee. The average earning was about €10 for the whole experiment.

An innovation of this article is that we actually estimate the indirect effect of control and test whether it is negative, i.e. whether they are hidden costs of control. From the econometricians viewpoint, the choices of the subjects x^{c} and x^{nc}. are random variables: X^{c} and X^{nc}. Accordingly, the overall effect of control on the principal's payoff is:

$$2 \cdot (X^{\mathrm{c}} - X^{\mathrm{nc}})$$

and the indirect effect, defined in equation (5), becomes

$$2 \cdot (X^{\mathrm{c}} - X^{\mathrm{nc}}) - 2 \cdot \left\{ \begin{array}{lll} \underline{x} - X^{\mathrm{nc}} & \text{if} & X^{\mathrm{nc}} < \underline{x} \\ 0 & \text{if} & X^{\mathrm{nc}} \geq \underline{x} \end{array} \right\}.$$

Consequently, the expected indirect effect is:

$$2 \cdot \mathrm{E}(X^{\mathrm{c}} - X^{\mathrm{nc}}) - 2 \cdot \mathrm{E}(\underline{x} - X^{\mathrm{nc}} | X^{\mathrm{nc}} < \underline{x}) \cdot P(X^{\mathrm{nc}} < \underline{x}). \tag{13}$$

A point estimate for the indirect effect can readily be obtained by replacing the theoretical by empirical moments and the probability by the respective frequency. If we want to know whether hidden costs are present, we have to check whether the indirect effect is significantly smaller than zero. At first glance, it is not obvious how such a test can be performed: the indirect effect comprises expected values but also a probability and thus does not directly lend itself to non-parameteric tests. Fortunately, the indirect effect can be written using a transformed variable:

$$2 \cdot (X^{\mathrm{c}} - X^{\mathrm{nc}}) - 2 \cdot \left\{ \begin{array}{lll} \underline{x} - X^{\mathrm{nc}} & \text{if} & X^{\mathrm{nc}} < \underline{x} \\ 0 & \text{if} & X^{\mathrm{nc}} \geq \underline{x} \end{array} \right\} = 2 \cdot (X^{\mathrm{c}} - \tilde{X^{\mathrm{nc}}}), \tag{14}$$

where

$$\tilde{X}^{\mathrm{nc}} := \left\{ \begin{array}{lll} \underline{x} & \text{if} & X^{\mathrm{nc}} < \underline{x} \\ X^{\mathrm{nc}} & \text{if} & X^{\mathrm{nc}} \geq \underline{x} \end{array} \right\}.$$

This transformed variable is constructed to eliminate the direct effect of control and leads to the following representation of the expected indirect effect:

$$2 \cdot \mathrm{E}(X^{\mathrm{c}} - \tilde{X}^{\mathrm{nc}}). \tag{15}$$

This simple representation now enables us to use standard tests, such as the Wilcoxon signed rank test, in order to determine whether the expected indirect effect is significantly different from zero.[9]

[9] Falk and Kosfeld (2006) suggest to use a two-sided Wilcoxon signed rank test and a respective transformation to test whether control has a "behavioral impact". Our considerations show that the formal hypothesis behind this test is that the expected indirect effect from equation (13) is significantly different from zero.

5 Results

Our first finding replicates an important aspect of the work by Falk and Kosfeld (2006).

Result 1 *In the BASELINE treatment, there are hidden costs of control.*

There are two necessary conditions for costs of control to exist. First, there must be subjects which exert more than the required level when they are not controlled. Indeed half of the subjects in the BASELINE treatment do so. Second, at least some of these subjects must exert a lower effort when controlled. In other words, they have to condition their effort negatively on control. The left panel in Figure 1 shows the distribution of the effort choices under control and no control. The distribution under control is stochastically smaller which indicates that some subjects provide more effort without control. The right panel in the same figure depicts the distribution of the individual difference between the effort under control and no control for each subject. It shows that around 37% of the subjects give less when they are controlled and 18% reduce their effort by ten or more.

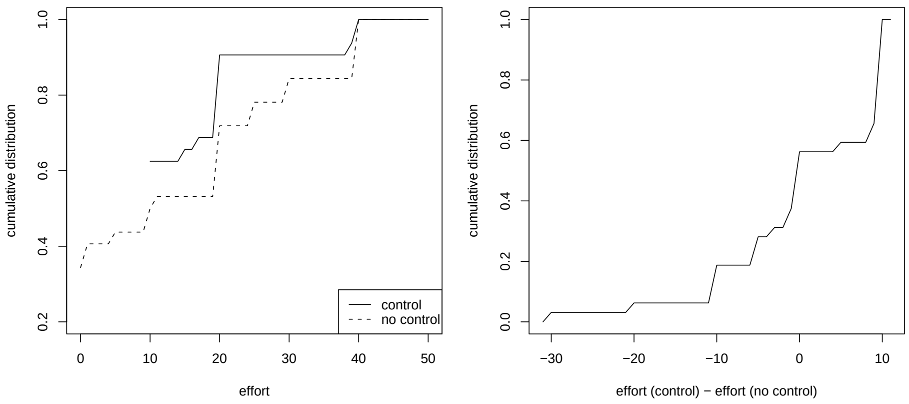

Figure 1: Cumulative distribution of effort (left panel) and of within-subject differences in effort choices for the BASELINE treatment

When we estimate the indirect effect of control on effort as described in the previous section, we find that it amounts to -6.875. On average, principals thus incur hidden costs of 7 points when controlling. Is this loss statistically significant? In line with Falk and Kosfeld (2006), we find that hidden costs of control are significant (two-sided exact Wilcoxon signed rank test as well as exact permutation test have p-values< 0.001). Falk and Kosfeld infere the presence of hidden costs of control indirectly from the observation that the overall effect of control is negative. Notice that our approach allows to directly identify the hidden costs of control and test whether they are significant. Indeed, the direct and indirect effect cancel each other in our data: the direct effect amounts to 8.313 points so that the overall effect is 8.313-6.875=1.4375 and we cannot reject the hypothesis that the overall effect is zero (The two-sided Wilcoxon signed rank test has a p-value of 0.38 and the permutation test of 0.70).

After having established the existence of costs of control, we now examine how these costs change across treatments.

Result 2 *In comparison to the BASELINE treatment, costs of control are lower in the ROBOT treatment.*

In the ROBOT treatment a third of the subjects voluntarily exerts an effort above the minimum requirement–even if they are not controlled. Those subjects can in principle reduce effort when they are controlled. The left panel in Figure 2 illustrates that agents still condition effort on control, while the right panel indicates that the differences between effort under control and no control are considerably smaller in the ROBOT treatment than in the BASELINE treatment. In the ROBOT treatment, 9% of the subjects choose to exert less effort under control while about 37% did so in the BASELINE treatment. Likewise, the share of subjects who "punish" control by reducing effort by 10 or more points drops from 18% in the BASELINE to only 6% in the ROBOT treatment. The estimate for the indirect effect of control amounts to -2.06. they are no longer significantly different from zero (the one-sided Wilcoxon signed rank test has p-value 0.78 and the permutation test has a p-value of 0.81). Hidden costs in the ROBOT treatment are only about a third of those in the BASELINE treatment; they are 5.84 points lower. We can verify whether this drop is significant by using representation (15) and testing whether the expected value of $X^{\mathrm{c}} - \tilde{X}^{\mathrm{nc}}$ in the ROBOT treatment is greater than or equal to that in the

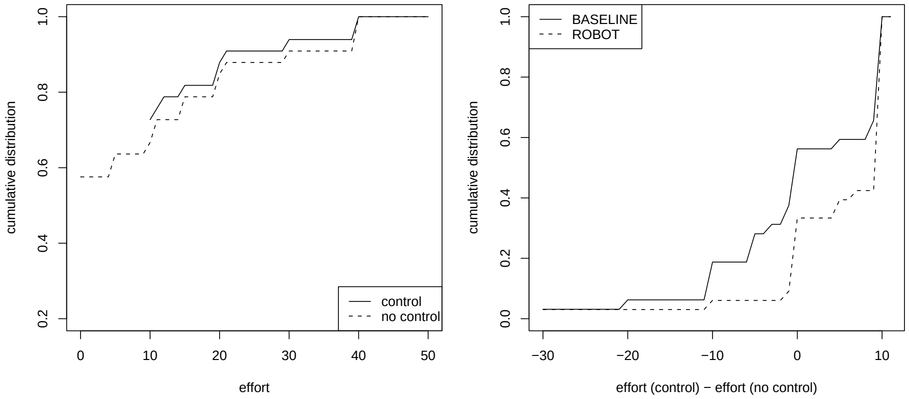

Figure 2: Cumulative distribution of effort for the ROBOT treatment (left panel) and of within-subject differences in effort choices for the BASELINE and ROBOT treatment (right panel)

BASELINE treatment. Based on two-sample one-sided tests, this hypothesis is rejected (the Wilcoxon signed rank test has p-value < 0.01 and the permutation test has a p-values of 0.06). The hidden costs of control thus decrease when moving to an environment, where control is less likely to be aimed at the agent. This observation resonates with the idea that guilt-averse agents condition less when control teaches them less about the expectation of the principal.

Next, we turn to the differences between the BASELINE and the endowment treatment.

Result 3 *In comparison to the BASELINE treatment, hidden costs of control are lower in the ENDOWMENT treatment.*

A third of the subjects gives voluntarily more than the required level and can thus potentially choose a lower effort leven under control. Differently from the BASELINE treatment, the distribution of effort under control is not smaller than without control in the ENDOWMENT treatment—see left panel in Figure 3. The share of subjects that punish the principal for controlling is lower than in the BASELINE treatment and the punishments too: only 11% exert less effort when being controlled and only about 5% reduce there effort by more than 10—see right panel in Figure 3. Recall that the indirect effect can be positive

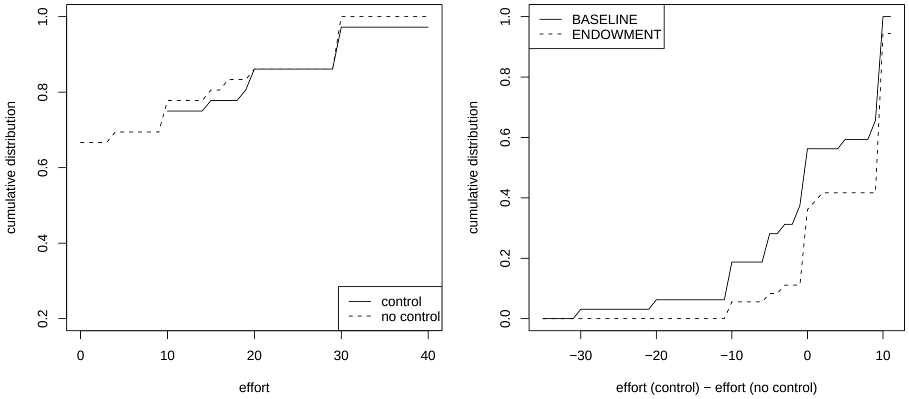

Figure 3: Cumulative distribution of effort (left panel) and of within-subject differences in effort choices for the ENDOWMENT treatment

if agents choose to exert more effort when they are controlled. If the indirect benefit is positive there are, of course, no hidden costs of control. This is indeed the case for the ENDOWMENT treatment: the estimate for the indirect effect amounts to $6.\overline{11}$. If we use the same testing approach as before, we find that the indirect effect in the ENDOWMENT treatment (and hence the hidden costs of control) are significantly lower than in the BASELINE treatment at any conventional level (Wilcoxon signed rank test and permutation test have p-values below < 0.01).[10]

The findigns from the ENDOWMENT treatment mesh well with the idea of legitimacy: if control is legitimate (according to our definition) principals are not punished for controlling.

Next, we turn to the behavior of principals.

Result 4 *In comparison to the BASELINE treatment, principals control more often in the ROBOT treatment.*

In the BASELINE treatment, the principal controls in 23 of 33 cases. These numbers contrast with the ROBOT treatment, where she controls in 32 out of 33 cases. Figure 4 depicts the shares of principals that control for the three treatments. The share in BASELINE is significantly lower than in ROBOT (p-value for Pearson's χ^2-test: < 0.01, p-value for Fisher's exact test: 0.074). Principal's seem to be aware of the fact that controlling leads to lower costs when control is not only concerning the agent but also a selfish ROBOT.

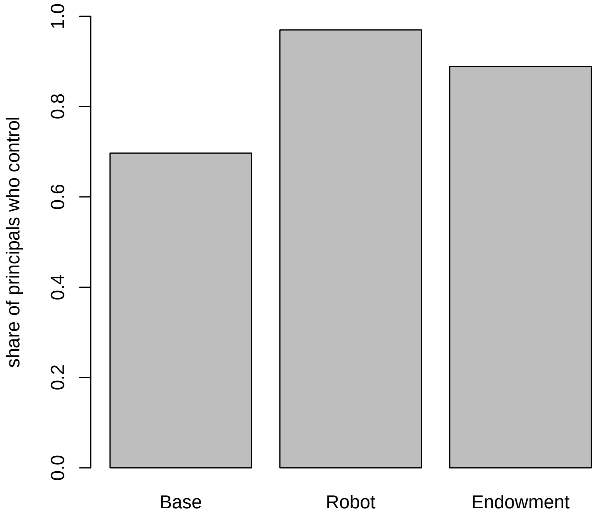

Figure 4: Control by principals across treatments

When interpreting this result, it is important to note that it is a dominant strategy to control in all treatments given the behavior of agents.[11] So a significantly larger proportion of principals controls in the ROBOT treatment *although* control already maximizes surplus in the BASELINE treatment. Moreover, agents anticipate that principals control more:[12] in the BASELINE

[10] The positive sign of the indirect effect is driven by an agent who exerts the maximal effort when being controlled and no effort otherwise. Excluding this agent from the data does not change the results: hidden costs remain significantly lower in the ENDOWMENT treatment; Wilcoxon signed rank test and permutation test have p-values below < 0.01).

[11] Based on the C10 treatment by Falk and Kosfeld, we expected no control to be dominant in the BASELINE treatment.

[12] We asked agents what they believe the principal will do. The question was not salient.

treatment 28 of 32 agents expect the principal to control while all 33 agents in the ROBOT treatment expect to be controlled; these differences are weakly significant (p-value for Pearson's χ^2-test: 0.06, p-value for Fisher's exact test: 0.05). This finding resonates with the idea how guilt-aversion affects the behavior of agents: in the ROBOT treatment, more agents expect the principal to control than in the BASELINE treatment. As it is the same subject pool from which agents are drawn in both cases, the anticipated increase in control is independent from the subjects. Accordingly, the anticipated increase in control in the ROBOT treatment must result from a source differently from the agent. Accordingly, if the principal acts in accordance with these expectations her actions are less revelatory about the effort she expects. Across all treatments there are seven agents who anticipate no control by the principal; these seven agents seem to belief that the principal expects higher effort from no control. Once controlled they have to update this belief. The fact that five of these agents indeed exert lower effort when controlled is additional, albeit anecdotal, confirmation that costs of control result from guilt aversion.

Result 5 *In comparison to the BASELINE treatment, principals control more often in the ENDOWMENT treatment.*

While about 70% of the principals control in the BASELINE treatment, ca. 90% control in the ENDOWMENT treatment. These differences are significant at a 5% level (p-value for Pearson's χ^2-test: 0.0464, p-value for Fisher's exact test: 0.0457). Principals seem to be aware that control is less costly in the ENDOWMENT treatment relative to the BASELINE treatment.

6 Conclusion

We have re-examined the idea that controlling entails costs in the light of the theory of guilt aversion. Together with a legitimacy refinement, this theory gives us predictions on how costs of control change in different contexts. We have studied two variations of the orginal experiment by Falk and Kosfeld (2006). In the first, control is not specifically aimed at the agent. In the second, control does not reduce the agent's endowment but is in line with the social norm of property. In both cases, we expect costs of control to be lower because the control by the principal is legitimate. We find these predictions confirmed. Our findings highlight that costs of control are highly specific to the context in which control is exerted. If control is expected for reasons independent of the individual agent, the costs of control are significantly lower. Falk and Kosfeld (2006) suggest that control may be an attractive option for the principal when it is particularly effective. Our study uncovers two other reasons: the agent may not take control personally or may show understanding for the principal's behavior.

References

Battigalli, P., and M. Dufwenberg (2005): "Dynamic Psychological Games," Working Paper 287, IGIER.

——— (forthcoming): "Guilt in Games," *American Economic Review*, Papers and Proceedings.

Charness, G., and M. Dufwenberg (2006): "Promises and Partnership," *Econometrica*, 74, 1579–1601.

Dufwenberg, M., and G. Kirchsteiger (2004): "A theory of sequential reciprocity," *Games and Economic Behavior*, 47(2), 268–298.

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.

Fehr, E., and K. M. Schmidt (1999): "A Theory Of Fairness, Competition, And Cooperation," *The Quarterly Journal of Economics*, 114(3), 817–868.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989): "Psychological Games and Sequential Rationality," *Games and Economic Behaviour*, 1(1), 60–79.

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

A Appendix

A.1 Auxiliary results

Lemma 1 *In all equilibria, the agent meets the expectations of the principal.*

Proof. Consistency requires that the expected values of the agent's second order beliefs are identical to that of the principal's first order beliefs. The principal's expectations are thus $\mu^{\mathrm{nc}}, \mu^{\mathrm{c}}$ in equilibrium. Suppose the agent chooses a strategy such that $\tilde{x}^i > \mu^i$ instead of $x^i = \mu^i$. Then, effort costs increase by $\tilde{x}^i - x^i$ and the deviation is not profitable. On the other hand, if the agent chooses a strategy $\tilde{x}^i < \mu^i$ instead of $x^i = \mu^i$. Guilt increases by more than $\tilde{x}^i - x^i$ and the deviation is also not worthwhile. Overall, the agent thus chooses $x^i = \mu^i$ and meets the expectations. ■

Lemma 2 *Deviating to 'no control' is always legitimate.*

Proof. Consider for instance $\hat{x} = x_{\mathrm{L}}$, which is available under the deviation. It satisfies Condition (11) because $\pi^1_{\mathrm{A}}(x_{\mathrm{L}}, x_{\mathrm{L}}) = \pi^0_{\mathrm{A}} > \pi^0_{\mathrm{A}} - \underline{x} \geq \pi^0_{\mathrm{A}} - x^* = \pi^*_{\mathrm{A}}$. Moreover, it also satisfies part (i) of Condition (12) as $\pi^1_{\mathrm{A}}(x_{\mathrm{L}}, x_{\mathrm{L}}) = \pi^0_{\mathrm{A}}$. ■

Lemma 3 *Deviation to control is not legitimate if and only if* $p = 1$ *and* $\underline{x} > 0$.

Proof. First note, that Condition (11) is satisfied by any $\hat{x} \in \mathcal{X}$, where $\mathcal{X} := \{x | \underline{x} \leq x \leq x^{\mathrm{nc}}\}$, because the agent is then at least as well off under deviation as in the equilibrium, $\pi^1_{\mathrm{A}}(\underline{x}, \hat{x}) = \pi^0_{\mathrm{A}} - \hat{x} \geq \pi^0_{\mathrm{A}} - \mu^{\mathrm{nc}} = \pi^*_{\mathrm{A}}$.

If $\underline{x} \leq 0$, part (i) of Condition (12) is satisfied for $\hat{x} = 0$ because $\pi^1_{\mathrm{A}}(\underline{x}, \hat{x}) = \pi^0_{\mathrm{A}}$. Conversely, if $\underline{x} > 0$, there is no $\hat{x} \in \mathcal{X}$ that fulfills part (i) because control then forces the agent to give up some of the initial endowment: $\pi^0_{\mathrm{A}} > \pi^0_{\mathrm{A}} - \underline{x}$. Hence, part (i) is fulfilled if and only if $\underline{x} \leq 0$.

If $p < 1$, part (ii) of Condition (12) is satisfied for $\hat{x} = x^{\mathrm{nc}}$ because $\pi^1_{\mathrm{P}}(\underline{x}, \hat{x}) = 2 \cdot (p x^{\mathrm{nc}} + (1-p)\underline{x}) \geq \pi^1_{\mathrm{P}}(\underline{x}, \hat{x}) = 2 \cdot (p x^{\mathrm{nc}} + (1-p) x_{\mathrm{L}})$. Conversely, if $p = 1$, there is no $\hat{x} \in \mathcal{X}$ that fulfills part (ii) because the principal gets at most the equilibrium payoff: $\pi^1_{\mathrm{P}}(\underline{x}, \hat{x}) \leq \pi_{\mathrm{P}}(x_{\mathrm{L}}, x^{\mathrm{nc}}) = \pi^*_{\mathrm{P}}$. Accordingly, part (ii) is fulfilled if and only if $p < 1$.

Summarizing, control is not legitimate if and only if $p = 1$ and $\underline{x} > 0$.

■

A.2 Proof of Proposition 3

Proposition 3 (Existence of closed- and open-door equilibria) *If the probability of interaction* p *is sufficently large,* $\frac{1-p}{p} < \frac{\pi^0_A - \underline{x}}{\underline{x} - x_L}$, *there is an open-door equilibrium. There is always a closed-door equilibrium.*

Proof. First, we prove existence of closed-door equilibria. Consider the following candidate. By definition, the principal controls. The whole probability mass of the principal's first-order beliefs α is focused on the strategy $x^{\mathrm{nc}} = x^{\mathrm{c}} = \underline{x}$. Define second-order beliefs of the agent consistently, such that $\mu^{\mathrm{nc}} = \mu^{\mathrm{c}} = \underline{x}$. By Lemma 1, the agent meets these expectations and plays $x^{\mathrm{nc}} = x^{\mathrm{c}} = \underline{x}$. On the other hand, a deviation by the principal to 'no control' yields the same payoff and is hence not profitable.

Second, we show that no open-door equilibria can exist if p is too small. For an open-door equilibrium to exist, the principal must have an incentive to leave the choice free: $p(x^{\mathrm{nc}} - x^{\mathrm{c}}) > (1-p)(\underline{x} - x_{\mathrm{L}})$, where strictness results from our assumption that the principal controls when she is indifferent. The largest value that can be attained by the costs of control on the left-hand side is $\pi^0_{\mathrm{A}} - \underline{x}$. Accordingly, open-door equilibria can only exist if

$$\frac{1-p}{p} < \frac{\pi^0_{\mathrm{A}} - \underline{x}}{\underline{x} - x_{\mathrm{L}}}. \tag{16}$$

Third, we prove that there is an open-door equilibrium if the preceding condition is met. Consider the following candidate. The principal expects the agent to contribute the complete endowment under no control and nothing else: $x^{\mathrm{nc}} = \pi^0_{\mathrm{A}}$

and $x^{\mathrm{c}} = \underline{x}$. Take consistent second-order beliefs of the agent such that $\mu^{\mathrm{nc}} = \pi^0_{\mathrm{A}}$ and $\mu^{\mathrm{c}} = \underline{x}$. Again, the agent will meet expectations by Lemma 1. For the principal, a deviation to 'control' is not profitable because of the inequality (16). The candidate is hence an equilibrium. ■

A.3 Proof of Proposition 4

Proposition 4 (Equilibrium type and beliefs) *Consider a sufficiently large probability of interaction, $(1-p)/p < (\pi^0_A - \underline{x})(\underline{x} - x_L)$, and an arbitrary equilibrium. Then, the equilibrium is an open-door equilibrium if and only if the second-order beliefs of the agent α_A satisfy*

$$\mu^{nc} - \mu^{c} > \frac{1-p}{p}(\underline{x} - x_L). \tag{17}$$

Proof. As the agent meets expectations by Lemma 1, the principal gets $p\mu^{\mathrm{nc}} + (1-p)x_{\mathrm{L}}$ if she leaves the choice free and $p\mu^{\mathrm{c}} + (1-p)\underline{x}$ if she controls. Using again our assumption that the principal controls if she is indifferent, no control his hence profitable if and only if

$$\mu^{\mathrm{nc}} - \mu^{\mathrm{c}} > \frac{1-p}{p}(\underline{x} - x_{\mathrm{L}}).$$

Accordingly, the equilibrium is only open-door if and only if this inequality is met. ■

A.4 Proof of Proposition 1

By Lemma 3, control is not legitimate if $p = 1$ and $\underline{x} > 0$. Hence, the legitimacy criterion requires that $\mu^{\mathrm{c}} \leq \mu^{\mathrm{nc}}$. This inequality is fulfilled by the beliefs that support the open-door equilibrium (see Proposition 4). Any open-door equilibrium thus fulfills the legitimacy criterion. However, the inequality is not fulfilled for all beliefs that support closed-door equilibria (see again Proposition 4). In particular, it rules out beliefs of the type $\mu^{\mathrm{c}} > \mu^{\mathrm{nc}}$. Hence, the only closed-door equilibria fulfilling the legitimacy criterion must be supported by beliefs $\mu^{\mathrm{c}} = \mu^{\mathrm{nc}}$.

A.5 Proof of Proposition 2

By Lemma 3, control is legitimate if $p < 1$ or $\underline{x} \leq 0$. Suppose there is an open-door equilibrium that satisfies the legitimacy criterion. Then, the principal can deviate and be certain that the belief of the agent will not drop so that the agent's action will be $\hat{x} \geq x^{\mathrm{nc}}$. This means the principal is at least as well off when controlling. By our indifference assumption, the principal will thus control.