Contractual Incompleteness as a Signal of Trust *

Florian Herold†

Preliminary Version

Abstract

This paper shows how the fear of signaling distrust can endogenously lead to incomplete contractual agreements. According to standard results in contract theory an optimal incentive contract should be conditional on all verifiable information containing statistical information about an agent’s action or type. Most real world contracts, however, condition only on few contingencies and often no explicit contract is signed at all. This paper argues that the proposal of a sophisticated complete contract including fines for misbehavior and other explicit incentives signals distrust to the partner. A trustworthy partner would choose the desired action anyway. Insisting on explicit contractual incentives thus means that the partner’s trustworthiness is called into question. Thus if it is important for the relation that an agent believes to be trusted, a principal may prefer to propose an incomplete contract rather than to signal her distrust by proposing a complete contract. Asymmetric information about how much one partner trusts the other one leads thus endogenously to contractual incompleteness for strategic reasons.

JEL-Classification: D86, C72, C9,

Keywords: Trust, Incomplete Contracts, Signaling, Fairness

*An earlier version of this paper can be found as a chapter of my dissertation. I would like to thank Roland Benabou, Dirk Bergemann, Armin Falk, Ernst Fehr, Georg Gebhardt, Klaus M. Schmidt and participants at seminars in Dresden, Munich and Toulouse for helpful comments and suggestions.

†Contacts: Kellogg School of Management, Northwestern University, 2001 Sheridan Rd., Evanston, IL 60208, Tel.: ++1 857 491-5305, email: f-herold@northwestern.edu
1 Introduction

This paper demonstrates how the fear to signal distrust can lead endogenously to incomplete contractual agreements.

According to standard results in contract theory an optimal contract should be conditional on all verifiable information containing statistical information about an agent’s action or type.\(^1\) Most real world contracts, however, condition only on few contingencies and often no explicit contract is signed at all. The costs of writing a complete contract or the limited ability to foresee all relevant contingencies can only partially explain the observed contractual incompleteness. There remain many relationships in which a simple contract could help to avoid potentially severe incentive problems at relatively low costs. Nonetheless, many people abstain from writing a complete contract. Why?

This paper argues that designing a sophisticated complete contract with fines, punishments and other explicit incentives signals distrust to the other party.

Consider the example of a scientist hiring a research assistant. Some potential incentive problems could be avoided by simple contractual arrangements. For instance, if one part of the assistant’s work consists of collecting some data, the scientist could give him the right incentives by announcing to spot-check his work and to fire the assistant in case she detects some faked data. The potential damage for the scientist in case her research relies on faked data is considerable and may far outweigh her costs of spot-checking. She may, nonetheless, abstain from such an announcement, because the research assistant is likely to interpret such checks as a signal of distrust regarding his scientific dedication. The feeling that the scientist distrusts him destroys the assistant’s motivation in other parts of the relationship. For instance, the assistant may expect a lower success from some potential, mutually beneficial, joint research projects if the scientist doubts his scientific dedication. He would therefore invest less effort in searching for such joint projects - to the disadvantage of the scientist.

More generally, consider a principal (“she”) who is interested in the success of a project

\(^1\) See e.g. Holmstrom [19] or Laffont and Tirole [?].
that she can only realize with the help of an agent (“he”). There are two types of agents. The *trustworthy* type has an intrinsic interest in the success of the project and is therefore willing to exert effort even without any contractual incentives.\(^2\) The *untrustworthy* type is not intrinsically interested in the success of the project. Only explicit contractual incentives can motivate him to exert effort.

The principal may hold different beliefs about the type of the agent depending on some private signals.\(^3\) We call the belief of the principal that the agent is trustworthy the “trust” of the principal in the agent. More trust means that the principal considers it more likely that the agent exerts effort even in absence of explicit incentives. The more a principal trusts the agent the lower are her expected costs (and her expected marginal costs) from contractual incompleteness. A principal can therefore try to separate herself from less trusting types by using contractual incompleteness as a signaling device.

Why should the principal have an interest in signaling trust? In our model trust is relevant in some parts of the relationship which are non-contractible by assumption. The more the principal trusts the agent, the more she is willing to rely on the agent. She may, for instance, follow his advice more often or give the agent more discretion in his decisions. This, in turn, increases the productivity of the agent’s effort in the joint project. Thus, if a trustworthy agent believes to be distrusted by the principal, he invests less effort and the project is less successful.

Our model focuses on the point that the concern to avoid a signal of distrust may spread contractual incompleteness from non-contractible parts to the contractible parts of a relationship.\(^4\) Even in a perfectly contractible world, however, the concern to signal distrust causes contractual incompleteness whenever the belief to be distrusted leads to a negative

---

\(^2\)Equivalently, the agent could be motivated by some fairness motives.

\(^3\)Such a signal could be private information about the agent, could stand for past experiences of the principal with other agents, or could symbolize a more or less optimistic nature of the principal.

\(^4\)Holmström-Milgrom [20] and Bernheim-Whinston [3] give two different arguments, why it can be optimal to leave some verifiable aspects of a relationship unspecified when other aspects cannot be verified. Holmström-Milgrom show in a multi-task setting, that it may be optimal to give no explicit incentives if the agent has some intrinsic motivation, tasks are substitutes and when the unverifiable task is sufficiently important. Bernheim-Whinstons argument is based on the observation, that writing no contract may give both sides more discretion to punish the other side. This can be important in a repeated games framework where harsh out of equilibrium punishments may be necessary to sustain the desired equilibrium.
reaction of the agent. We discuss other, more psychological, reasons for such a negative reaction after presenting our main model.

The literature on the foundations of incomplete contracts is extensive. A recent survey of this literature is Tirole [28]. Spier [25] points out most explicitly that signaling can cause contractual incompleteness. In her model a risk-averse principal hires a risk-neutral agent. The principal has private information on whether the probability that her project results in high profits is high or low. In the refined equilibrium the principal offers an unconditional wage and thereby (inefficiently) forgoes some insurance in order to signal to the agent that the success probability is high. Notice, however, that, in general, asymmetric information at the contracting stage can equally well lead to more instead of less complete contracts. For instance, slightly changing the setting of Spier to an informed risk-averse principal selling a risky asset to a less informed risk-neutral agent (potentially with some transaction cost for writing the contract conditional on outcome) results in a too complete contract: The principal can signal a good risk asset by conditioning the proposed contract on the outcome, although under symmetric information the agent should pay a fixed price for the asset independently of the asset’s outcome. Hence, in general, the effect of signaling concerns on contractual completeness is ambiguous.

In our paper, in contrast, there is a clear prediction that contracts should be less complete when the principal wants to signal trust: The trustworthy type is defined as someone who chooses the desired action even without contractual enforcement. Hence, the more the principal trusts, the lower she estimates the costs of an incomplete contract. More contractual incompleteness therefore signals more trust and the equilibrium contract is distorted towards less completeness whenever the principal wants to signal trust.

Furthermore, our paper adds a new perspective to the literature on the detrimental effects of sanctions and explicit incentives. A number of authors and recent experimental studies suggest that sanctions, control and explicit incentives can crowd out intrinsic motivation and

---

5See also Allen-Gale [1] for similar ideas in the context of financial economics.
6See also Tirole [28], p.764 for this point.
7See e.g. Etzioni [9], Deci-Ryan [6] and Frey [15].
8See Fehr-Rockenbach [13], Gneezy-Rustichini [17],[18], and Fehr-Klein-Schmidt [12] In section 3 we briefly discuss the experimental findings of Falk-Kosfeld [10] in the light of our model.
may even be counterproductive. Sanctions seem to have a particularly detrimental effect in cases where they are deliberately designed by one of the involved parties.

Bénabou-Tirole [2] and Fehr-Klein-Schmidt [12] suggest two different channels through which explicit incentives can negatively affect performance. In Bénabou-Tirole [2] a better informed principal wants to give explicit incentives to the agent if a task is unpleasant or the agent has a low ability. The less informed agent understands that explicit incentives are a signal for an unpleasant task or for his low ability. Explicit incentives therefore crowd out intrinsic motivation and are thus less powerful than under symmetric information. The model of Bénabou-Tirole requires that the principal has superior knowledge about the agent’s type (or his task). Our model, in contrast, focuses on the better knowledge of the principal about her own beliefs about the agent’s type. This seems a natural assumption in almost every setting with asymmetric information.

The paper by Fehr-Klein-Schmidt [12] addresses the question of how fairness concerns affect the choice of contracts. They show experimentally that it may be optimal for a principal to rely on implicit incentives (the promise of a bonus for good performance) rather than on explicit incentives (a commitment to a limited fine after poor performance). They demonstrate by a calibration that the experimental results are consistent with a heterogeneous population where some players are inequity-averse while others act selfishly. Our argument may complement their explanation: It is natural to define trust in their setting as the belief that the agent is of a fair type. The existence of some fair-types gives implicit incentives their strength. The heterogeneity of preference types and of beliefs about these types can, in addition, make the choice of explicit incentives counterproductive, as they signal distrust.

Recently and independently of this paper other authors developed models with related ideas. Discuss relation to papers by Ellingsen-Johanneson [7], Sliwka [24], Friebel?.

After presenting and analyzing our model in section 2 we discuss our results in section 3 before concluding in section 4.
2 The Model

2.1 Setting

Consider the following principal-agent relationship: A principal needs to hire an agent to realize a project. The agent is one of two types, trustworthy ($T$) or untrustworthy ($U$); the proportion of trustworthy types in the economy, $\pi$, is strictly between zero and one. The trustworthy type of agent is intrinsically interested in the success of the project, whereas the untrustworthy type does not care about the project per se.\(^9\)

The principal cannot observe the agent’s type. She receives, however, a binary, private signal $s \in \{+, -\}$ about the type of the agent. In case the agent is trustworthy, the principal receives a positive signal with the exogenously given probability $\sigma_T$. In case the agent is untrustworthy, the principal receives a positive signal with the exogenously given probability $\sigma_U$, with $\sigma_U < \sigma_T$. These two moves by “nature” are illustrated in figure 1

![Figure 1: Nature’s moves](image)

By Bayes’ rule a principal with a positive signal believes that she faces the trustworthy type with probability $\pi_+ = \frac{\pi}{\pi + \frac{\pi}{\pi_T}(1-\pi)} > \pi$ and a principal with a negative signal believes that she interacts with a trustworthy type with probability $\pi_- = \frac{\pi}{\pi + \frac{\pi}{\pi_T}(1-\pi)} < \pi$. Notice that $\pi_+ > \pi_-$, i.e. a principal with a positive signal has a stronger belief in the agent’s

\(^9\)Instead of being intrinsically interested in the project the agent may also have preferences for fairness and therefore, deliberately, exert high effort.
trustworthiness. In other words, the principal with a positive signal trusts the agent more strongly than the principal with a negative signal.

The project consists of two parts, the contractible part 1 and the non-contractible part 2. The project’s success in the contractible part 1, $B_1 \in \{0, \overline{B}_1\}$, depends only on an unobservable effort $e_1 \in \{0, \overline{e}_1\}$ by the agent. High effort $\overline{e}_1$ benefits the project deterministically by $\overline{B}_1 \equiv B_1(\overline{e}_1) > \overline{e}_1$. Low effort $e_1 = 0$ leads to $B_1 = 0$. A contract can condition on the outcome $B_1(e_1)$ which is realized at the very end of the relationship, i.e. after both players have chosen their actions in the second part of the relationship. Although effort $e_1$ is not directly observable, it can be inferred from the realized value of $B_1$.\(^{10}\) A sufficiently harsh punishment in case of $B_1 = 0$ therefore implements a high effort level $e_1 = \overline{e}_1$ in this contractible part 1.

The success of the project in part 2, $B_2(e_{2,a}, e_{2,p})$, depends on an unobservable effort choice by the agent, $e_{2,a}$, as well as on an unobservable effort choice by the principal, $e_{2,p}$. Higher effort by the agent or the principal increases the expected success of the project. Furthermore, the effort of the agent and the principal are complements, i.e. the higher the effort of one side, the higher the marginal productivity of one additional unit of effort by the other side. In the appendix we show that the main argument of the paper goes through as well if effort by the agent and effort by the principal are substitutes.

**Assumption 1** $B_2(\cdot, \cdot)$ is twice continuously differentiable with \(\frac{\partial B_2(\cdot, \cdot)}{\partial e_{2,a}} > 0\), \(\frac{\partial B_2(\cdot, \cdot)}{\partial e_{2,p}} > 0\), \(\frac{\partial^2 B_2(\cdot, \cdot)}{\partial e_{2,a}^2} < 0\), \(\frac{\partial^2 B_2(\cdot, \cdot)}{\partial e_{2,p}^2} < 0\) (i.e. $B_2()$ is increasing and concave in each variable); \(\frac{\partial^2 B_2(\cdot, \cdot)}{\partial e_{2,a} \partial e_{2,p}} > 0\) (complementarity); \(\frac{\partial B_2(0,0)}{\partial e_{2,a}} > 1\), \(\frac{\partial B_2(0,0)}{\partial e_{2,p}} > 1\), and $B_2$ is bounded (guaranteeing an interior solution) and ... (guaranteeing a unique solution).

The principal wants to signal her trust in the agent due to this second, non-contractible part of the relationship: the principal and a trustworthy agent are willing to exert some effort for the sake of the project. The higher their expectations about the other side’s effort, the more effort they are willing to invest themselves. A trustworthy agent works harder than

---

\(^{10}\)By assuming that $e_1$ is only ex post (indirectly) observable we avoid complication of $e_1$ signalling something about the agent’s type.
an untrustworthy agent (who does not work in part 2). The higher trust of the principal, the higher the probability she assigns to high effort of the agent and the harder she will work. The more the agent believes to be trusted, the higher his expectations about the principals effort and the more effort the trustworthy agent is willing to exert. The principal wants the agent to exert high effort and is therefore interested in signaling her trust in the agent.

The total success $B$ of the project is the sum of the successes, $B_1(e_1)$ and $B_2(e_{2a}, e_{2p})$, in both parts of the relationship, i.e.

$$B(e_1, e_2, \cdot) \equiv B_1(e_1) + B_2(e_{2a}, e_{2p}).$$

(1)

**Timing of events**

The timing of events is illustrated in Figure 2.

![Figure 2: Timeline](image)

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal receives private signal about the agent’s type.</td>
<td>Contracting Stage</td>
<td>Part 1: agent’s unobservable effort choice $e_1$.</td>
<td>Part 2: principal’s and agent’s unobservable, complementary effort choices $e_{2p}$ and $e_{2a}$.</td>
<td></td>
</tr>
</tbody>
</table>

After nature has randomly chosen the agent’s type and the principal’s signal, principal and agent sign a contract. Then, the agent chooses his effort level $e_1$. Finally, principal and agent choose there effort level in part 2. Both effort choices are unobservable. By assumption a contract cannot condition on the outcome in part 2.

**Contracting Stage**

At the contracting stage, the principal proposes a contract. This contract can, by assumption, only be conditional on $B_1(e_1)$, the project’s success in the first part. The most relevant feature of this contract is, whether the contract enforces high effort $\tau_1$ by a sufficiently harsh
punishment in case of $B_2(e_2) = 0$. In general, however, the principal can design a sophisticated wage scheme and the agent’s decision whether to accept or reject the proposal may potentially reveal information about his type.

Here, we want to demonstrate our main point as concise as possible. For the moment, we thus restrict the set of contractual choices of the principal to a binary choice $C \in \{\text{contract } (c), \text{no contract } (n)\}$. In appendix A.2, we demonstrate that our main argument remains valid if we allow for more general contracts and if we take care of the agent’s participation constraint.

The contract prescribes high effort $e_1 = \bar{e}_1$ in the contractible part 1. A court enforces the contract by the threat of a sufficiently harsh punishment in case of $B_1(e_1) = 0$. In case of writing no-contract the principal refrains from such an enforcement.

A possible interpretation for this simple setting is that principal and agent are working together already, and that there exists a binding agreement fixing the wage. In particular, let this existing agreement give the principal the discretion to enforce high effort of the agent in the contractible part 1 of the relationship through a contract or to abstain from doing so. Notice that in this simple setting, the agent takes no observable action. After the first exogenous signal $s$, the principal’s beliefs (i.e. her trust) about the agent’s type remain fixed at $\pi_\pm$.

**Preferences**

For simplicity, let the principal and the agent be risk neutral. The untrustworthy type of agent maximizes his monetary payoff $M$ minus his total effort costs $e \equiv e_1 + e_2$. He does not care about the project.

The utility-function of the untrustworthy-agent is given by

$$U_{u}(M, e, B(\cdot)) = M - e. \quad (2)$$

The trustworthy type of agent is intrinsically interested in the success of the project $B(\cdot)$. We allow for the possibility that the trustworthy agent puts a lower weight, $\kappa \leq 1$, on the
project’s success than the principal (in monetary units). The utility function of a trustworthy agent is

\[ U_T(M, e, B(\cdot)) = M - e + \kappa B(\cdot). \]  

(3)

We need, however, that the trustworthy agent sufficiently cares about the project to deliberately exert high effort \( \overline{e}_1 \).

Assumption 2

\[ 1 \geq \kappa \geq \frac{\overline{e}_1}{B_1}. \]  

(4)

The principal’s utility is given by

\[ V(W_P, e_{2P}, B(\cdot)) = B(\cdot) - e_{2P} - W_P, \]  

(5)

where \( W_P \) is the wage payed by the principal, with \( W_P \geq M \). In case \( W_P > M \) the principal is “burning” some money. Wage payments are relevant only for the extended version in appendix A.2. Here, we normalize \( W_P = 0 = M \). Hence, the principal maximizes simply the expected total success of the project \( B(\cdot) \) minus her effort costs \( e_{2P} \).

2.2 Analysis of the Principal-Agent Relationship

We analyze this principal-agent relationship backwards.

2.3 Part 2

In the non-contractible part 2 the untrustworthy agent will never invest any effort as he is not intrinsically interested in the projects success: \( e_{2, a}^{U} = 0 \).

The trustworthy agent is willing to invest some effort. How much he is willing to invest depends on his expectations about the principal’s effort choice. The more effort he expects the principal to exert, the higher his own effort choice. In equilibrium the effort choice of the trustworthy agent, of the trusting principal and of the distrusting principal are given by
the following first order conditions.

\[
\alpha_T \frac{\partial B_2 (e_{2a}^+, e_{2p}^+)}{\partial e_{2a}} + (1 - \alpha_T) \frac{\partial B_2 (e_{2a}^-, e_{2p}^-)}{\partial e_{2a}} = \frac{1}{\kappa},
\]

(6)

\[
\pi_+ \frac{\partial B_2 (e_{2a}^+, e_{2p}^+)}{\partial e_{2p}} + (1 - \pi_+) \frac{\partial B_2 (0, e_{2p}^+)}{\partial e_{2p}} = 1,
\]

(7)

\[
\pi_- \frac{\partial B_2 (e_{2a}^-, e_{2p}^-)}{\partial e_{2p}} + (1 - \pi_-) \frac{\partial B_2 (0, e_{2p}^-)}{\partial e_{2p}} = 1,
\]

(8)

where \(0 \leq \alpha_T \leq 1\) denotes the trustworthy agent’s belief that the principal received a positive signal and trust him. In general, \(\alpha_T\) can differ from \(\sigma_T\) as the principal’s choice of contract may signal her type to the agent.

Notice that for every given \(e_{2a}^T\), equation 7 defines a unique optimal \(e_{2p}^+ (e_{2a}^T)\) and equation 8 defines a unique optimal \(e_{2p}^- (e_{2a}^T)\). Furthermore \(e_{2p}^+ (e_{2a}^T) > e_{2p}^- (e_{2a}^T)\) as \(\pi_+ > \pi_-\).

By the implicit function theorem

\[
\frac{d}{de_{2a}} (e_{2p}^+ (e_{2a})) = -\frac{\pi_+}{\pi_-} \frac{\partial^2 B_2 (e_{2a}^+, e_{2p}^+)}{\partial e_{2a} \partial e_{2p}} + (1 - \pi_+) \frac{\partial^2 B_2 (0, e_{2p}^+)}{\partial e_{2p}^2} > 0.
\]

(9)

For every given \(\alpha_T, e_{2p}^+, \) and \(e_{2p}^-\) equation 6 defines a unique optimal \(e_{2a}^{T*} (e_{2p}^+, e_{2p}^-, \alpha_T)\).

**Claim 1** Under Assumption ··· there exists a unique solution to the system of equations 6, 7, and 8 for any given \(\alpha_T \in [0, 1]\).

Remark: Simplify analysis by a condition for uniqueness similar to dominant diagonal condition \(- \frac{\partial f_n}{\partial x_n} > \sum_{j \neq n} \frac{\partial^2 f_n}{\partial x_n \partial x_j}\). Compare e.g. Milgrom and Roberts, Econometrica 1990 (Example 2 in section 4).

We denote this solution by \(e_{2a}^{T*} (\alpha_T), e_{2p}^- (\alpha_T), \) and \(e_{2p}^+ (\alpha_T)\).

Furthermore, let \(V_2^\pm (\alpha_T) \equiv \pi_\pm B_2 (e_{2a}^{T*} (\alpha_T), e_{2p}^\pm (\alpha_T)) + (1 - \pi_\pm) B_2 (0, e_{2p}^\pm (\alpha_T)) - e_{2p}^\pm (\alpha_T)\) denote the expected net profit from part 2 to a ±-principal and for a given \(\alpha_T\).
Lemma 1  For any $\alpha_T \in [0, 1]$ holds

\[ V_2^+(\alpha_T) > V_2^-(\alpha_T) \quad \text{and} \quad \partial V_2^+ (\cdot) / \partial \alpha_T > \partial V_2^- (\cdot) / \partial \alpha_T > 0. \] (10)

In other words ...

Proof: to be written.

Analysis of the Contracting Stage

The principal’s decision of whether or not to write a contract is observed by the agent and may influence the trustworthy agent’s belief $\alpha_T$ of whether the principal trusts him. This belief, in turn, influences the agent’s effort decision $e^*_T(\alpha_T)$.

In a perfect Bayesian equilibrium the contractual choice $C_+ \in \{c, n\}$, of a principal with a positive signal, and $C_- \in \{c, n\}$, of a principal with a negative signal, must both be optimal given the principal’s type (i.e. his belief $\pi_{\pm}$) and given the trustworthy agent’s beliefs $\alpha_T^c$ and $\alpha_T^n$. On the other hand, the trustworthy agent’s beliefs $\alpha_T^c$ after observing the choice of a contract and $\alpha_T^n$ after observing the choice of no-contract must be rational given the equilibrium contractual choices $C_+$ and $C_-$. 

Consider a potential equilibrium candidate which we denote, in a slight abuse of notation, by $(C_+, C_-, \alpha_T^c, \alpha_T^n)$.

Let $V_n^c$ denote the expected payoff of a $(\pm)$-principal from the choice of writing no-contract, and $V_c^c$ the expected payoff from the choice of writing a contract. Then,

\[ V_n^c = \pi_+ B_1 + V_2^+ (\alpha_T^c) \] (12)
\[ V_c^c = B_1 + V_2^+ (\alpha_T^c) \] (13)
\[ V_n^- = \pi_- B_1 + V_2^- (\alpha_T^n) \] (14)
\[ V_c^- = B_1 + V_2^- (\alpha_T^n). \] (15)

\[\text{11}\]The beliefs of the untrustworthy agent are payoff irrelevant. The remaining equilibrium actions are the unique best responses given $(C_+, C_-, \alpha_T^c, \alpha_T^n)$. 

12
The following relations are useful for the further analysis:

\[
V^n_v \geq V^c_v \iff V^+_2 (\alpha^+_T) - V^+_2 (\alpha^+_T) \geq (1 - \pi_+) \bar{B}_1
\]  
(16)

\[
V^n_\nu \geq V^c_\nu \iff V^-_2 (\alpha^-_T) - V^-_2 (\alpha^-_T) \geq (1 - \pi_-) \bar{B}_1
\]  
(17)

The right hand side is smaller for the trusting principal than for the distrusting one. Whenever the left hand side is positive, the l.h.s. is greater for the trusting principal than for the distrusting principal. Thus, the distrusting principal is less willing to choose no contract compared to the trusting principal.

Even a trusting principal only considers to refrain from the option to enforce high effort in part 1 by means of a contract if the expected costs \((1 - \pi_+) \bar{B}_1\) are below the maximal possible expected gains from such a signal \((V^+_2 (1) - V^+_2 (0))\). Otherwise, both types of principal always pool on writing a contract. Thus, we concentrate on the interesting case:

**Condition 1**

\[
\bar{B}_1 < \frac{V^+_2 (1) - V^+_2 (0)}{1 - \pi_+}
\]  
(18)

**Refinement by the Intuitive Criterion:** Signaling games suffer from a multiplicity of perfect Bayesian equilibria that are often sustained by pessimistic out of equilibrium beliefs. We mainly focus on equilibria that are consistent with the intuitive criterion, an equilibrium refinement introduced by Cho and Kreps [5].\(^{12}\) It constrains the set of out-of-equilibrium-beliefs by the following equilibrium domination argument: Consider an out-of-equilibrium action in a perfect Bayesian equilibrium of the game. If one type A can never expect to profit from this deviation (given that all players play best-responses to some out of equilibrium belief) whereas a different type potentially could profit, then any “intuitive” belief should assign probability 0 to type A. The precise definition of the intuitive criterion is in the appendix. The following definition tightens the notation.

**Definition 1** An intuitive equilibrium is a perfect Bayesian Equilibrium that is consis-

\(^{12}\)In our context we need the somewhat more elaborate definition of the intuitive criterion used by Maskin-Tirole [22], because we deal with more stages than the standard two-stage signaling game.
tent with the intuitive criterion.

In the appendix we specify and derive all perfect Bayesian equilibria of this game. Here, we concentrate on our main point and summarize the most relevant results in the following propositions.

**Proposition 1** Under assumption 1, 2, and condition 1, there always exists an intuitive equilibrium in which the trusting principal chooses to write no contract.

**Remark 1** These intuitive equilibria are

a) for \( B_1 \geq \frac{V_0^-(1) - V_0^-(0)}{1 - \pi_-} \):

the separating equilibrium in which the trusting principal chooses to write no contract
and the distrusting principal chooses to write a contract.

b) for \( \frac{V_0^-(\sigma_T) - V_0^-(0)}{1 - \pi_-} < B_1 \leq \frac{V_0^-(1) - V_0^-(0)}{1 - \pi_-} \):

the hybrid equilibrium in which the trusting principal chooses to write no contract and
the distrusting principal chooses to write no contract with probability
\[
q = \sigma_T^{1-\alpha_n} \frac{1-\alpha_n}{\alpha_n^2}.
\]

c) for \( B_1 \leq \frac{V_0^-(\sigma_T) - V_0^-(0)}{1 - \pi_-} \):

the pooling equilibria on writing no contract, in which both types of principal forbear
from writing a contract.

To see the intuition for proposition 1 and remark 1, notice that the expected costs of
refraining from writing a contract are smaller for the trusting principal, \((1 - \pi_+)B_1\), than for
the distrusting principal, \((1 - \pi_-)B_1\). In addition, the expected gains from signaling trust
are higher for the trusting principal as she considers it more likely that she interacts with a
trustworthy agent, the type who invests more effort when he believes in being trusted.

Consider a very large \( B_1 \), such that condition 1 does not hold. Then, the expected losses
from signaling trust by writing no contract are too large compared to the potential gains;
the trusting as well as the distrusting principal pool on writing a contract.

Decreasing \( B_1 \) until condition 1 just holds, a second equilibrium emerges, the separating
equilibrium of remark 1 a. In the range of remark 1 a the trusting principal takes the risk
to forgo $B_1$ in part 1 of the project to separate herself from the distrusting type. For a
distrusting principal writing no contract, to imitate the trusting type, is too expensive.

As we decrease $B_1$ further the costs of signaling trust, by forbearing from writing a con-
tract, become smaller. Perfect separation of the trusting and distrusting principal becomes
unsustainable when the expected costs of writing no contract for the distrusting type fall
below the gains from being mistaken for a trusting type.\textsuperscript{13} Then, a fraction of distrusting
principals starts imitating the signal. The larger this fraction of distrusting imitators, the
lower, in equilibrium, the belief of the rational agent after observing the signal “no contract”.
In the hybrid equilibrium of remark 1 b the fraction of distrusting principals imitating the
signal takes exactly the value that makes a distrusting principal indifferent between choosing
“contract” or “no contract”.

Decreasing $B_1$ further, decreases, in this hybrid equilibrium, the probability that a dis-
trusting principal reveals his type by writing a contract. When this probability reaches zero,
we end up in the pooling equilibrium of remark 1 c, sustained by an out of equilibrium belief,
that a “contract” is a signal of a distrusting type.

In addition to the equilibria in proposition 1 and remark 1 there exist further perfect
Bayesian equilibria, in particular the pooling equilibria on writing a contract. For an inter-
mediate range of $B_1$, however, these additional equilibria do not pass the intuitive criterion
and intuitive equilibrium actions are unique.

\textbf{Proposition 2} Under assumption 1, 2, condition 1, and in case of

$$\frac{V_2^- (1) - V_2^- (\sigma_T)}{1 - \pi_-} < B_1 < \frac{V_2^+ (1) - V_2^+ (\sigma_T)}{1 - \pi_+}$$

the equilibria of remark 1 are the only intuitive equilibria. In particular, the trusting principal
chooses to write no contract in any intuitive equilibrium.

\textbf{Corollary 1} Under assumption 1, 2, condition 1, and inequality 19 there is a unique out-
come in all intuitive equilibria. More precisely,

\textsuperscript{13}Notice that this threshold, $\frac{\pi}{1 - \pi} \epsilon (B_h^1 - B_h^0)$, is close to zero if $\epsilon$ is small.
a) For $\overline{B}_1 \geq \frac{V_2^{1(-1)} - V_2^{0(0)}}{1 - \pi}$: the separating equilibrium, in which the trusting principal chooses to write “no contract”, is the unique intuitive equilibrium.

b) For $\frac{V_2^{1(-1)} - V_2^{0(0)}}{1 - \pi} < \overline{B}_1 < \frac{V_2^{1(-1)} - V_2^{0(0)}}{1 - \pi}$: the hybrid equilibrium in which the trusting principal chooses to write no contract and the distrusting principal chooses to write no contract with probability $q = \frac{\sigma_T}{1 - \sigma_T} \frac{1 - \alpha_T^n}{\alpha_T^n}$ is the unique intuitive equilibrium.

c) For $\overline{B}_1 \leq \frac{V_2^{1(-1)} - V_2^{0(0)}}{1 - \pi}$: all intuitive equilibria are pooling equilibria on writing no contract; these equilibria differ only in their out-of-equilibrium beliefs, $\alpha_T^n$.

The intuition for proposition 2 follows from two observations. Firstly, under condition 19 there exists no hybrid equilibrium in which the trusting principal uses a mixed strategy. Secondly, the pooling equilibrium on writing a contract requires that the out of equilibrium belief for writing no contract, $\alpha_T^n$, is smaller than 1. Under condition 19, however, the intuitive criterion requires an out of equilibrium belief of 1, as the distrusting principal can never expect to profit, whereas the trusting principal does profit for $\alpha_T^n = 1$.

These propositions confirm the main point of this paper; the principal may choose to leave contracts incomplete to avoid a signal of distrust. Proposition 1 states that, under condition 1, there always exists an equilibrium in which at least the trusting principal abstains from writing a contract, and that these equilibria pass the intuitive criterion. Under condition 19 these are the only equilibria that are not excluded by the intuitive criterion - as stated in proposition 2. Then, there is a unique intuitive equilibrium outcome, in which the trusting principal forbears from writing a contract.

3 Discussion

Our model demonstrates that the fear to signal distrust can endogenously cause contractual incompleteness. The trusting principal, in particular, prefers to to write no contract, although under symmetric information she would strictly prefer to write such a contingent
contract. She is more afraid of being mistaken for the distrusting principal, than of being exploited by an untrustworthy agent.

In the simple model, with a binary contractual choice, *contract or no contract*, such an equilibrium, in which the trusting type refrains from writing a contract, exits only if the exogenously given costs of contractual incompleteness are not to high. The range for uniqueness of the intuitive equilibrium is even smaller.

In appendix A.2 the principal can design more general wage schemes. This complicates the analysis and requires additional assumptions to ensure that the principal can not separate the trustworthy and the untrustworthy agent by a screening contract. In return, however, we get a clear cut prediction about the intuitive equilibrium outcome: In any intuitive equilibrium the distrusting principal chooses the complete contract and the trusting principal chooses the least-cost separating contract. In other words, the trusting principal chooses a degree of contractual incompleteness just sufficient to separate herself from the distrusting type.

Intuitively, the principal can choose any degree of contractual incompleteness by designing the contract properly, i.e. she can choose any loss in case the agent turns out to be untrustworthy. In an intuitive equilibrium, the trusting principal never proposes a contract that is chosen by the mistrusting principal with a strictly positive probability. Marginal and absolute costs of contractual incompleteness are lower for a trusting principal while the gains from signaling trust are larger. In a candidate equilibrium contract chosen by both types of principal the trusting type can separate herself by choosing an additional degree of contractual incompleteness just sufficiently high that the distrusting type strictly prefers the candidate equilibrium contract. By the intuitive criterion the agent should assign a belief of $\alpha_T = 1$ when observing such a deviating contract. Then, however, the trusting principal has a strict preference for this deviation and the candidate equilibrium could not have been an intuitive equilibrium in the first place.

\footnote{Losses smaller than $B_1 > 0$ can be generated by a lottery over contracts - with some probability complete, with the counter probability incomplete. Losses larger than $B_1 > 0$ can be generated by giving the agent no explicit incentives and a commitment by the principal to burn some money in case the agent turns out to be untrustworthy, i.e. in case of $B_1 = 0$.}
In our model, the value of signaling trust comes from an under-investment of the agent in a non-contractible part of the relationship in case he believes to be distrusted. If we are willing to depart further from standard economic preferences and take a more sociological or psychological standpoint then the negative consequences of signaling distrust become even more relevant. The proposal of a detailed complete contract with sophisticated fines and rewards basically states “I believe you are one of those types who exploits me if he can”. Most people would perceive such a statement as an insult and lose any motivation to invest in this relationship. In addition, trust seems to have a strong mutual component. How can I trust you if you do not trust me? In a relation of similar partners this inference can be rational to some degree and may be amplified by the false consensus effect.\footnote{To see this consider the following stylized model: The world can be either good or bad. In a good world most people are trustworthy while in a bad world most people are untrustworthy. If someone knows only his own preferences he should assign a higher probability to a bad world if he is untrustworthy, himself. Hence, people who distrust others may be more likely to be untrustworthy themselves. See also Englemann [8].}

In particular sociologist emphasize that human behavior and well-being is strongly influenced by social status and what other people think about them. Feeling distrusted lowers one’s utility directly and a signal of distrust may destroy the potential surplus from a relationship. Only if we feel trusted we attach positive emotions to a relationship, we are willing to invest in it and to forgo some instant profits to maintain the impression of being trustworthy.\footnote{Recent studies suggest (see Sunnafrank-Ramirez [27]) that the first impressions are decisive for the long-term nature of a relationship. Contract proposal are often the starting point in a relation.} Luhmann [21] coined the expression of the “self-fulfilling prophecy of distrust”. This effect is demonstrated neatly in a recent experimental study by Falk-Kosfeld [10]. An agent can spend any amount from his endowment of 120 token in an investment which benefits only the principal (by the double amount of the investment). Upfront, the principal can choose whether she wants to force the agent to invest at least 10 token, or whether she abstains from any control (then the minimum investment is 0). The large majority of principals chose not to control the agent and, in fact, on average agents invested significantly more when the principal chose to not control. An additional control-treatment where the minimum investment was exogenously given demonstrates that it is not the control per se that leads to lower investments of the agent, but the fact that the principal has deliberately
chosen to control. This clearly suggests that a principal choosing to control the agent signals distrust and that this crowds out trustworthiness of the agent.

4 Conclusions

This paper demonstrates that the fear to signal distrust can lead to endogenous contractual incompleteness, even when there are no costs of writing a contract and all types of principal would prefer to write a complete contract under symmetric information about their beliefs. Our results are driven, firstly, by asymmetric information on the principal’s beliefs about the agent’s type and, secondly, by the importance of trust, and the belief to be trusted, for the relationship.

This fear to signal distrust by proposing to write a contract can cause considerable inefficiencies. Carefully designed policies may therefore become relevant and can help to mitigate these inefficiencies. Many couples are reluctant to sign a prenuptial agreement as they are afraid of signaling distrust to their spouse. A well designed, fair, standard regulation, at least for the case that no contract is written, may therefore be very important. In fact, most states do regulate the consequences in case of a divorce to some degree.

Similarly, parents who funded their children’s education should, in return, receive support from their children in case they need help when becoming elderly. Most parents, however, will be very reluctant to insist on any explicit contract with their children - for not showing distrust. Again, some prudential regulation by a government may help to mitigate such problems.

THE REMAINDER OF THE PAPER REFERS TO AN OLD VERSION OF THE MODEL AND IS CURRENTLY UNDER REVISION.
A Appendix

A.1 Proofs and all Perfect Bayesian Equilibria

The Perfect Bayesian Equilibria

Separating Equilibrium: Under condition 1, there exists a separating, perfect Bayesian equilibrium if and only if

\[ B_1^1 \leq B_0^0 + \frac{1 - \pi - B_1}{\pi -} \frac{1}{\epsilon}. \]  

(20)

In this equilibrium the trusting principal with belief \( \pi_+ \) chooses to write no contract, (i.e. \( C_+ = n \)), and the distrusting principal with belief \( \pi_- \) chooses to write a contract, (i.e. \( C_- = c \)).

Pooling on writing No Contract: There exist perfect Bayesian pooling equilibria on writing no-contract if and only if

\[ B_{\sigma T}^\sigma \geq B_0^0 + \frac{1 - \pi - B_1}{\pi -} \frac{1}{\epsilon}. \]  

(21)

On the equilibrium path the belief of the trustworthy agent is \( \alpha_T^n = \sigma_T \) and the equilibrium is sustained by an out of equilibrium belief \( \alpha_T^c \leq (B_h^*)^{-1} \left( B_h^*(\sigma_T) - \frac{1 - \pi - B_1}{\pi -} \right) \).

Pooling on writing a Contract: There always exist perfect Bayesian pooling equilibria on writing a contract. On the equilibrium path, the belief of the trustworthy agent is \( \alpha_T^c = \sigma_T \) and the equilibrium is sustained by an out of equilibrium belief

\[ \alpha_T^n \leq (B_h^*)^{-1} \left( B_h^*(\sigma_T) + \frac{1 - \pi + B_1}{\pi +} \right). \]

Hybrid Equilibria with a random contractual choice of the distrusting principal:

Under condition 1, there exists a hybrid perfect Bayesian equilibrium in which the trusting principal chooses to write no-contract with certainty and the distrusting principal randomly
mixes between both contractual choices if and only if

\[ B_h^{\sigma_T} \leq B_h^0 + \frac{1 - \pi - \overline{B}_1}{\pi+ \epsilon} \leq B_h^1. \]  

(22)

In this equilibrium the trustworthy agent’s beliefs are \( \alpha_i^c = 0 \) and \( \alpha_i^n = (B_h^*)^{-1} \left( \frac{1 - \pi - \overline{B}_1}{\pi+ \epsilon} + B_h^0 \right) \). The mixing probability for the distrusting principal to choose no-contract is \( q = \frac{\sigma_T}{1 - \sigma_T} \frac{1 - \alpha_i^2}{\alpha_i^2} \).

**Hybrid Equilibria with random contract choice of the trusting principal:**

Under condition 1 there exists a hybrid perfect Bayesian equilibrium in which the distrusting principal writes a contract with certainty and the trusting principal randomly mixes between “contract” and “no contract” if and only if

\[ B_h^1 - B_h^{\sigma_T} < \frac{1 - \pi+ \overline{B}_1}{\pi+} \]  

(23)

In this equilibrium the trustworthy agent’s beliefs are \( \alpha_i^c = (B_h^*)^{-1} \left( B_h^1 - \frac{1 - \pi+ \overline{B}_1}{\pi+} \right) \) and \( \alpha_i^n = 1 \). The mixing probability for the trusting principal to choose to write no-contract is \( q' = 1 - \frac{1 - \sigma_T}{\sigma_T} \frac{\alpha_i^c}{1 - \alpha_i^c} \).

**Refinement by the Intuitive Criterion:** Which of the perfect Bayesian equilibria are affected by the refinement of the intuitive criterion? The separating and hybrid equilibria have no out-of-equilibrium belief and therefore pass the intuitive criterion, anyway. Pooling equilibria on writing no contract pass the intuitive criterion too, since both types of principal could potentially profit from a deviation to the complete contract.

The only equilibria that potentially fail to pass the intuitive criterion are the pooling equilibria on the complete contract. The intuitive criterion requires the out-of-equilibrium belief after the choice of no contract to be \( \alpha_i^n = 1 \) if for this belief\(^ {17} \) the trusting principal expects to profit from the deviation, whereas the distrusting principal does not. In fact, no

\(^ {17} \text{Clearly, } \alpha_i = 1 \text{ leads to maximal profits for the principal.} \)
pooling equilibrium on the complete contract passes the intuitive criterion if (and only if)\(^\text{18}\)

\[
\frac{1 - \pi_+}{\pi_+} B_1 < B_{h}^1 - B_{h}^{\sigma_T} < \frac{1 - \pi_-}{\pi_-} \delta B_1.
\]

\(\text{(24)}\)

In the remainder of this paper we denote a perfect Bayesian equilibrium that passes the intuitive criterion as an **intuitive equilibrium**.

**Proofs and Derivation of the Equilibria**

**Perfect Bayesian Equilibria and Intuitive Equilibria**

The following definition is useful in the further analysis:

**Definition 2**

\[
e^\alpha_T \equiv e^\alpha_T(\alpha_T), \quad \text{in particular,}
\]

\[
e^0_2 \equiv e^\alpha_T(\alpha_T = 0) \quad \text{and} \quad e^1_2 \equiv e^\alpha_T(\alpha_T = 1).
\]

\(\text{(25)}\)

\(\text{(26)}\)

**Separating Equilibrium** In the only possible separating equilibrium\(^\text{19}\) the trusting principal chooses to write no contract and the distrusting principal chooses to write a contract. In this perfect-Bayesian-equilibrium the agent has the right beliefs the principal’s signal, i.e. he holds the beliefs \(\alpha^n_t = 1\) and \(\alpha^c_t = 0\).

The trustworthy agent therefore invests \(e^1_2\) if and only if the principal writes no contract. \(C_+ = n\) and \(C_- = c\) are best responses for both types of principal if and only if

\[
(\text{IC}_+) \quad V^n_+ \geq V^c_+
\]

\[
(\text{IC}_-) \quad V^n_- \leq V^c_-,
\]

\(\text{(27)}\)

\(\text{(28)}\)

i.e. by equivalences 17 and 17

\(\text{18}\)In case of \(\frac{1 - \pi_+}{\pi_+} \delta B_1 = 1 - \sigma_T\), the intuitive criterion still requires \(\alpha_t(\text{cc} = 0) = 1\), but pooling on the complete contract remains incentive compatible.

\(\text{19}\)There can be no separating equilibrium in which the distrusting principal chooses to write no contract since incompleteness is costly for her and she prefers to not being separated from the trusting type.
Equation 29 is already implied by condition 1. Summarizing, under condition 1, there exists a perfectly separating perfect Bayesian equilibrium if and only if condition 30 holds.

**Pooling Equilibria on “no contract” (n):** If both types of principal choose in equilibrium to write no contract then the agent’s belief remains unchanged when he observes this action, i.e. \( \alpha_T^n = \sigma_T \) (and \( \alpha_U^n = \sigma_u \)). Out of equilibrium the trustworthy agent has some belief \( \alpha_T^c \in [0, 1] \). This forms a perfect-Bayesian-Equilibrium if and only if

\[
(I_C^+) \quad V^u_n \geq V^c_n \\
(I_C^-) \quad V^u_n \geq V^c,
\]

which by equivalence 17 and 17 corresponds to

\[
(I_C^+) \quad B^\sigma_T - B^*_h (\alpha_T^c) \geq \frac{1 - \pi_+}{\pi_+} B_1 \\
(I_C^-) \quad B^\sigma_T - B^*_h (\alpha_T^c) \geq \frac{1 - \pi_-}{\pi_-} B_1.
\]

\( (I_C^+) \) is already implied by \( (I_C^-) \).Pooling on no contract can be sustained as a perfect-Bayesian-Equilibrium by an out-of-equilibrium-belief \( \alpha_T^c \in [0, 1] \) if and only if the equilibrium can be sustained by \( \alpha_T^c = 0 \). This is the case if and only if

\[
\sigma_T \geq (B^*_h)^{-1} \left( B^0_h + \frac{1 - \pi_-}{\pi_-} B_1 \right).
\]

**Pooling Equilibria on “contract” (c):** If both types of principal choose in equilibrium to write a contract (c) then the belief of the agent remains unchanged when observing this
action, i.e. $\alpha_T^\ell = \sigma_T$ (and $\alpha_T^u = \sigma_u$). Out of equilibrium the agent has some belief $\alpha_T^\ell \in [0,1]$. These beliefs and contractual choices form a perfect-Bayesian-Equilibrium if and only if

\begin{equation}
(\text{IC}_+) \quad V_+^n \leq V_+^\ell
\end{equation}

\begin{equation}
(\text{IC}_-) \quad V_n^u \leq V_n^\ell,
\end{equation}

which by equivalence 17 and 17 corresponds to

\begin{equation}
(\text{IC}_+) \quad B_h^\ell (\alpha^n) - B^\sigma_T \leq \frac{1 - \pi_+ B_1}{\pi_+}
\end{equation}

\begin{equation}
(\text{IC}_-) \quad B_h^\ell (\alpha^n) - B^\sigma_T \leq \frac{1 - \pi_- B_1}{\pi_-}.
\end{equation}

$(\text{IC}_-) \text{ is already implied by } (\text{IC}_+)$. Hence, pooling on writing a contract can always be sustained as a perfect-Bayesian-Equilibrium, e.g. by the out-of-equilibrium-belief $\alpha_T^\ell = 0$.

**Hybrid-Equilibrium with $C_+ = n$ and $q \equiv \text{prob}(C_+ = n) \in (0,1)$:** In this equilibrium the agent knows, when observing a contract that the principal is distrusting, i.e. $\alpha_T^\ell = 0$. In case he observes no contract he updates his beliefs by Bayes-rule:

\begin{equation}
\alpha^n_T \equiv \text{prob}(+|C = n) = \frac{\text{prob}(C = n|+) \text{prob}(+)}{\text{prob}(C = n)} = \frac{\sigma_T}{\sigma_T + (1 - \sigma_T)q}.
\end{equation}

Notice that $\sigma_T < \alpha^n_T < 1$ for every $q \in (0,1)$. Vice versa, for any $\sigma_T < \alpha^n_T < 1$ there exists a $q \in (0,1)$ that leads to this $\alpha^n_T$, namely $q = \frac{\sigma_T - \alpha^n_T}{1 - \sigma_T}$. These beliefs and contractual choices
form a perfect Bayesian equilibrium if and only if

\[(IC_+) \quad V^n_+ \geq V^c_+ \quad (43)\]

\[(IC_-) \quad V^n_- = V_- c, \quad (44)\]

which by equivalence 17 and 17 corresponds to

\[(IC_+) \quad B^*_h(\alpha^+_n) - B^0_h \geq \frac{1 - \pi_+}{\pi_+} B_1 \quad (45)\]

\[(IC_-) \quad B^*_h(\alpha^+_n) - B^0_h = \frac{1 - \pi_-}{\pi_-} B_1. \quad (46)\]

\((IC_+)\) is implied already by \((IC_-)\). Hence, there exists a perfect Bayesian hybrid equilibrium in which the trusting principal chooses “no contract” and the distrusting principal mixes between “contract” and “no contract” if and only if

\[B^{\sigma_T}_h < B^0_h + \frac{1 - \pi_-}{\pi_-} B_1 < B^1_h. \quad (47)\]

**Hybrid-Equilibrium with \(q' \equiv \text{prob}(C = n) \in (0, 1)\) and \(C = c\):** In this equilibrium the agent knows, when observing “no contract”, that the principal is of the trusting type, i.e. \(\alpha^n_T = 1\). In case he observes a contract he updates his beliefs by Bayes-rule:

\[\alpha^c_T = \frac{\text{prob}(C = c | +) \text{prob}(+) \text{prob}(C = c)}{\text{prob}(C = c)} = \frac{(1 - q') \sigma_T}{(1 - q') \sigma_T + (1 - \sigma_T)}. \quad (48)\]

Notice that \(0 < \alpha^c_T < \sigma_T\) for every \(q' \in (0, 1)\). These beliefs and contractual choices form an equilibrium if and only if

\[(IC_+) \quad V^n_+ = V^c_+ \quad (49)\]

\[(IC_-) \quad V^n_- \leq V_- c, \quad (50)\]
which by equivalence 17 and 17 corresponds to

\[
\begin{align*}
(IC_+) & \quad B_h^1 - B_h^*(\alpha_t^i) = \frac{1 - \pi_+}{\pi_+} B_1 \\
(IC_-) & \quad B_h^1 - B_h^*(\alpha_t^i) \leq \frac{1 - \pi_-}{\pi_-} B_1.
\end{align*}
\]

\[(51)\]  \[(52)\]

(IC_-) is implied already by (IC_+). Hence, there exists a perfect-Bayesian hybrid equilibrium in which the distrusting principal chooses to write a contract and the trusting principal mixes between “contract” and “no contract” if and only if

\[
B_h^0 < B_h^1 - \frac{1 - \pi_+}{\pi_+} B_1 < B_h^{\sigma_T}.
\]

\[(53)\]

The first inequality is guaranteed already by condition 1.

It is straightforward to check that there do not exist any further perfect Bayesian equilibria.

**Equilibrium Refinement by the Intuitive Criterion**  In our context of only two types of principals and more than the standard two stages of a signaling game the intuitive criterion by Cho-Kreps [5] (CK) takes the following form (see also Maskin-Tirole [22]):

Let \( T = \{ +, - \} \) denote the set of the two types of principals. Let \( BR(w, \alpha_t) \) denote the (unique) equilibrium strategies of the continuation game between the principal and the agent after \( w \) has been offered and has led the trustworthy agent\(^{20}\) to update his belief to \( \alpha_t \).

Consider a candidate perfect Bayesian equilibrium that leads in equilibrium to an expected utility \( V_i^* \) for a principal of type \( i \).

We denote an out-of-equilibrium contract proposal \( \hat{w} \) as equilibrium dominated for type \( i \in T \), if and only if

\[
V_i^* > \max_{\alpha_t \in [0,1]} V_i \left( BR \left( \hat{w}, \alpha_t \right) \right).
\]

\[(54)\]

**Definition 3**  A perfect Bayesian equilibrium passes the intuitive criterion if and only if

\(^{20}\)Notice that the equilibrium strategies are independent of the belief \( \alpha_u \) of the untrustworthy agent.
the out-of-equilibrium beliefs \( \alpha_t(\tilde{w}) \) assign zero probability to type \( i \) (i.e. \( \alpha_i(\tilde{w}) = 0 \) if \( i = + \) and \( \alpha_i(\tilde{w}) = 1 \) if \( i = - \)) whenever \( \tilde{w} \) is equilibrium dominated for type \( i \) and not equilibrium dominated for the other type \( j \).

All separating or hybrid equilibria are not affected by this additional constraint on the out of equilibrium beliefs since “contract” and “no contract” are both played in equilibrium by some type with a strictly positive probability. Only the pooling equilibria need to be analyzed.

**Pooling Equilibria on “Contract” (c):** In equilibrium the expected payoffs for the trusting principal and distrusting principal are

\[
V^{c,eq}_+ = \bar{B}_1 + \pi_+ B^{\sigma_T}_h - (1 - \pi_+) B_d \\
V^{c,eq}_- = \bar{B}_1 + \pi_- \epsilon B^{\sigma_T}_h - (1 - \pi_-) \epsilon B_d.
\]

In case a principal deviates to writing “no contract”, the trustworthy agent plays a best response to some belief. For both types of principal the best they can hope for is that this belief is \( \alpha^n_T = 1 \) and the trustworthy agent invests \( e_2 = e_1^2 \). The maximal resulting expected payoffs for the trusting principal and for the distrusting principal are therefore:

\[
V_{+}^{n,max} = \pi_+ (\bar{B}_1 + B^1_h) - (1 - \pi_+) B_d \\
V_{-}^{n,max} = \pi_- (\bar{B}_1 + \epsilon B^1_h) - (1 - \pi_-) \epsilon B_d.
\]

Writing “no contract” is equilibrium dominated for the trusting principal if and only if

\[
V^{c,eq}_+ > V_{+}^{n,max} \\
\iff \quad B^1_h - B^{\sigma_T}_h < \frac{1 - \pi_+}{\pi_+} B_1.
\]
For the distrusting principal writing “no contract” is equilibrium dominated if and only if

\[ V_{c,eq}^{C} > V_{n,max}^{n} \]

\[ \Leftrightarrow B_{h}^{1} - B_{h}^{\sigma} < \frac{1 - \pi_{-} \bar{\beta}_{1}}{\epsilon}. \]  

Hence, for

\[ \frac{1 - \pi_{+}}{\pi_{+}} \bar{B}_{1} < B_{h}^{1} - B_{h}^{\sigma} < \frac{1 - \pi_{-}}{\pi_{-}} \bar{\beta}_{1}, \]

the intuitive criterion demands \( \alpha_{nT}^{\sigma} = 1 \) and a pooling-equilibrium on writing a “contract” does not pass this refinement due to \((IC_{+})^{21}\).

The **Pooling Equilibria on “No Contract” (n):** all pooling equilibria on “no contract” pass the intuitive criterion, since both types could, potentially, profit from a deviation to “contract” (e.g., when \( \alpha_{T}^{\sigma} = \alpha_{T}^{n} \)).

### A.2 Wage Scheme Contracts

In the second scenario we allow for a more general class of contracts and take the participation constraint of the agent into account. Payments can condition on \( B_{1} \) and thereby indirectly on the agents’ effort \( e_{1} \). We still assume that the contract can not condition on anything in the non-contractible part 2 of the relationship.

We have a problem of two-sided asymmetric information. The focus of this paper is on a signaling story: The principal signals her trust by proposing some contractual incompleteness. In general, however, the principal might also try to screen between both types of agents. Firstly, she could try to propose a contract, that fulfills only the participation constraint of the trustworthy agent. In other words the principal would make the trustworthy agent pay for the right to work for him. In most settings, this is completely implausible and just an artefact of how we modeled the trustworthy agent’s preferences. We therefore

\[ 21 \text{In case of } \frac{1 - \pi_{+}}{\pi_{+}} \bar{B}_{1} = B_{h}^{1} - B_{h}^{\sigma} \text{ equilibrium domination still requires } \alpha_{T}^{\sigma} = 1, \text{ but } (IC_{+}) \text{ is nonetheless satisfied.} \]
change our modeling of the trustworthy agent’s preferences slightly: Instead of gaining from a successful projects we assume that a trustworthy agent suffers from working for a project that is less successful than it could be. Then the trustworthy agent still works deliberately, but he is not willing to to pay for his job.\textsuperscript{22} The trustworthy agent’s utility is then given by

\[ U_T(m, e, B) = m - e + \kappa(B - B_{\text{max}}), \]

where \( B_{\text{max}} = B_1 + B_h \) is the maximal success the project could ever have. Furthermore, we assume that both types of agent have an outside option of 0.

A second way the principal could try to screen between both types of agent is by offering a menu of contracts. In case of perfect screening, the principal would learn the type of the agent by his choice of contract and change her belief (i.e. her trust) correspondingly. Then the agent would understand that the principal knows his type and all asymmetric information would be resolved in equilibrium. Such complications can not arise, however, if being trusted is sufficiently important for the untrustworthy agent (i.e. if \( M_d \) is sufficiently large). Then screening is impossible. The untrustworthy type always mimics the behavior of the trustworthy agent at the contracting stage. Under the following assumption the principal can not screen the type of the agent\textsuperscript{23}:

**Assumption 3 (No Screening Condition (sufficient))**

\[ M_d > \frac{\kappa (B_1 + B_h)}{\min\{\sigma u, 1 - \sigma u\}} \]

**Lemma 2** Under assumption 3 the untrustworthy agent chooses the same contract as the trustworthy agent in any perfect Bayesian equilibrium.

The proof is at the end of the appendix.

\textsuperscript{22}The following to arguments lead to the same result: either we can assume that the trustworthy agent has a higher outside option, or that the principal needs to hire always the agent and cannot risk that the untrustworthy agent rejects the offer.

\textsuperscript{23}Even if this assumption does not hold, screening may well be too expensive for being optimal.
The Contract Proposal  The contract-proposal is designed by an informed party: The principal has relevant private information when proposing the contract and may therefore signal something about her private information to the agent.

In general, the principal may propose any probability distribution over a set of contracts\textsuperscript{24}. Each of these contracts can only condition on the realization of $B_1$, i.e. a contract specifies a tuple $w \equiv (w_P, w_P, w_A, w_A)$ with $w_P \geq w_A$ and $w_P \geq \overline{w}_A$. $w_P$ is the principal’s payment in case of $B_1 = 0$ and $\overline{w}_P$ her payment in case of $B_1 = B_1$. $w_A$ is the wage the agent receives in case of $B_1 = 0$ and $\overline{w}_A$ the wage the agent receives if $B_1 = B_1$. The requirements $w_P \geq w_A$ and $\overline{w}_P \geq \overline{w}_A$ capture that the wage of the agent has to be payed by the principal. In case of $w_P > \overline{w}_A$ the principal commits to “burn some money”.

After the principal’s contract proposal the agent updates his beliefs about the principal’s type. Then, he accepts the contract if and only if his expected utility under the contract equals at least his outside option. Otherwise, the agent receives his outside option of 0. We assume that the principal always wants to hire the agent.

When the agent accepted a court draws a realization from the proposed probability distribution. The court is committed to enforce this realized deterministic contract $(w_P, \overline{w}_P, w_A, \overline{w}_A)$.

The Reduced Form of the Contract Proposal  We simplify the analysis by treating only those contract proposals as different, that lead to different payoffs for at least one of the relevant types.

When the principal offers a contract, she has to care only about the belief and the participation constraint of the trustworthy agent. The untrustworthy agent always accepts the contract, when the trustworthy does and since the untrustworthy agent exerts no effort anyway, his beliefs do not affect the principal. The only way the untrustworthy agent reacts

\textsuperscript{24}Maskin and Tirole \cite{22} discuss in detail the problem of contract design by an informed principal with common values. They allow the principal to design (almost) any contracting mechanism. In particular the principal can propose a menu of contracts from which she herself chooses one contract after the agent accepted. This gives the signaling game screening properties. However, their analysis is does not directly apply to our setting since here the agent’s beliefs on the principal’s type do matter even after the contract is written. Nonetheless, we conjecture that our results do still hold in this more general setting, since their techniques select also the least-cost separating equilibrium (as we do).

\textsuperscript{25}Thus we avoid any potential problems of ex post incentives to renegotiate the contract.
to a contract is that he plays \( e_1 = \bar{e}_1 \) if and only if \( \Delta w_A \equiv \bar{w}_A - w_A > \bar{e}_1 \).\(^{26}\) For any given contract the principal can directly incorporate this into her calculation of her expected payoff.

Only the expected payoffs of the trusting principal, the distrusting principal and the trustworthy agent should therefore be relevant for the equilibrium analysis. When the trustworthy agent accepted the contract proposal of the principal, he has a certain belief \( \alpha_T \) about the principal’s type and we can calculate the expected equilibrium payoffs for each type, given this belief \( \alpha_T \). We simplify the analysis by

**Definition 4** Two contract-proposals \( \mu(w) \) and \( \mu'(w) \) are **payoff equivalent** if for every \( \alpha_T(\mu) = \alpha_T(\mu') \) both contract proposals lead in equilibrium to the same expected payoffs for the trusting principal, the distrusting principal and the trustworthy agent.

In appendix A.2 we demonstrate that each equivalence class can be described by a tuple \((w, l, l_u)\) where \( w \in \mathbb{R}, l \geq 0 \) and \( l_u \geq \min\{0, B_1 - \bar{e}_1 - l\} \). W.l.o.g., consider only contracts in which the trustworthy agent has an incentive to chose\(^{27}\) \( e_1 = \bar{e}_1 \): \( w \) is the expected wage payment from the principal to the trustworthy agent, \( l \) is an unconditional loss of utility for the principal and \( l_u \) is a loss in utility for the principal from contractual incompleteness in case the agent turns out to be untrustworthy. The possibility to choose \( l > 0 \) (i.e. to commit to always burn some money) will play only a minor role for the further analysis and is only included for completeness. The possibility to commit to a loss conditional on meeting an untrustworthy type, however, is crucial. The expected costs of such a commitment are lower for the trusting principal than for the distrusting principal. Choosing a high \( l_u \) can therefore serve as a signal of trust.

The expected payoffs from a contract \((w, l, l_u)\) conditional on the belief \( \alpha_T(w) \equiv \alpha_T(w, l, l_u) \) are for the trusting and distrusting principal, respectively:

\[
V_+(w, l, l_u, \alpha_T(w)) = -(w + l) - (1 - \pi_+) l_u + \pi_+ B_h^{\alpha_T(w)} + \left[ B_1 - (1 - \pi_+) B_d \right],
\]
\[
V_-(w, l, l_u, \alpha_T(w)) = -(w + l) - (1 - \pi_-) l_u + \pi_- \epsilon B_h^{\alpha_T(w)} + \left[ B_1 - (1 - \pi_-) \epsilon B_d \right].
\]

\(^{26}\)For convenience, we assume the tie breaking rule that the untrustworthy agent plays \( e_1 = 0 \) if \( \Delta w_A = \bar{e}_1 \).
\(^{27}\)This no real restriction as we show in appendix A.2 that there is such a contract in each equivalence class.
The respectively last terms in rectangular brackets are independent of the contract (and independent of the belief resulting from the contract-proposal). It is convenient to re-normalize the principals utilities to

\[ V_+(w, l, l_u, \alpha_T(w)) = -(w + l) - (1 - \pi_+) l_u + \pi + B_h^{\alpha_T(w)} , \]

\[ V_-(w, l, l_u, \alpha_T(w)) = -(w + l) - (1 - \pi_-) l_u + \pi - \epsilon B_h^{\alpha_T(w)} . \]

(66)

(67)

The expected utility of the trustworthy agent is

\[ U_T(w, l, l_u, \alpha_T(w)) = w - \bar{\epsilon}_1 - \epsilon_2 (\alpha_T(w)) + \kappa \left( (\alpha_T(w) (1 - \epsilon) + \epsilon) B_h^{\alpha_T(w)} - \bar{B}_h \right) , \]

where we used that \( B_{\max} = \bar{B}_1 + \bar{B}_h \). Thus, if the trustworthy agent has the belief \( \alpha_T(w) \) after a contract-proposal \( w \), then he accepts the contract if and only if

\[ w \geq \bar{\epsilon}_1 + \epsilon_2 (\alpha_T(w)) + \kappa \left( \bar{B}_h - (\alpha_T(w) (1 - \epsilon) + \epsilon) B_h^{\alpha_T(w)} \right) . \]

(68)

The right hand side decreases in \( \alpha_T \), it is the stronger the belief of the agent that he is trusted, the lower the minimum wage offer that he requires to accept the offer. Thus, a wage \( w \geq \bar{\epsilon}_1 + \epsilon_2 + \kappa (\bar{B}_h - \epsilon B_h^{\alpha_T}) \) assures acceptance of the t-agent for any belief \( \alpha_T \).

It is typical for signaling games to suffer from a multiplicity of perfect Bayesian equilibria. We have a (3 dimensional) continuum of contracts, correspondingly many possible out of equilibrium beliefs and, therefore, a large number of equilibria. We derive and specify them in the appendix. Given this multiplicity it is remarkable that only one of this equilibria passes the intuitive criterion - the least cost separating equilibrium:

**Proposition 3** The only perfect Bayesian equilibrium passing the intuitive criterion is the least cost separating equilibrium:

\[ w^+ = \left( w^+ = \bar{\epsilon}_1 + \epsilon_2 + \kappa (\bar{B}_h - B_h^1) , l^+ = 0 , l_u^+ = \frac{\epsilon_2^0 - \epsilon_2^1 + (\kappa + \epsilon \pi_-) B_h^1 - (\kappa + \pi_+) \epsilon B_h^{0}}{1 - \pi_-} \right) \]

\[ w^- = \left( w^- = \bar{\epsilon}_1 + \epsilon_2^0 + \kappa (\bar{B}_h - \epsilon B_h^{0}) , l^- = 0 , l_u^- = 0 \right) . \]

(69)
Before we prove proposition 3, we briefly discuss the result. Intuitively the bite of the intuitive criterion in this more general set of possible contracts comes from the fact that marginally increasing incompleteness of the contract (i.e. the costs $l_u$ of meeting the untrustworthy type) is less costly for the trusting principal. Whenever the agent might mistake her for the distrusting type, she could gradually increase incompleteness until the bad type would not follow her.

$l_u$ was defined as the loss of the principal compared to a complete contract in case the agent turns out to be untrustworthy (for a fixed belief $\alpha_T$). Such a loss exists only when the contract provides insufficient incentives for the untrustworthy type to exert high effort $e_1$. An $l_u > 0$ means therefore that the contract is incomplete (at least with some strictly positive probability).

In the unique intuitive equilibrium the trusting principal chooses to propose an incomplete contract to signal her trust in the agent’s trustworthiness.

Derivation of the perfect Bayesian equilibria in pure strategies and proof of proposition 3 in scenario 2:

Separating equilibria  Let $w^+ \equiv (w^+, t^+, l^+_u)$ denote the contract proposal of the (+)-principal and $w^- \equiv (w^-, t^-, l^-_u)$ the contract proposed by the (-)-principal. Then in any perfect Bayesian separating equilibrium: $\alpha_T(w^+) = 1$ and $\alpha_T(w^-) = 0$.

The proposal of $w^-$ can only be optimal for the (-)-principal in such a separating equilibrium if

$$w^- = (w = \bar{c}_1 + \epsilon^0_2 + \kappa (\bar{B}_h - \epsilon B^0_h), l = 0, l_u = 0).$$

(70)

There exist some out of equilibrium beliefs that sustain $w^+$ and $w^-$ as a separating equilibrium if and only if the beliefs $\alpha_T(w) \equiv 0 \forall w \neq w^+$ sustain the equilibrium. Notice that with these out of equilibrium beliefs $w^-$ is more attractive for both types of principal than any out of equilibrium contract proposal.
In the separating equilibrium the t-agent accepts the proposal \( w^+ \) if and only if
\[
  w^+ \geq \overline{e}_1 + e_2^1 + \kappa \left( B_h - B^1_{h,1} \right). \tag{71}
\]

Furthermore, it must optimal for each type of principal to choose her own proposal, i.e.
\[
  (IC_+) \quad V_+(w^+, l^+, l^+_u, \alpha_T(w^+)) \geq V_+(w^-, l^-, l^-_u, \alpha_T(w^-)) \tag{72}
\]
\[
  (IC_-) \quad V_-(w^+, l^+, l^+_u, \alpha_T(w^+)) \leq V_-(w^-, l^-, l^-_u, \alpha_T(w^-)). \tag{73}
\]

Rearranging leads to
\[
  (IC_+) \quad (w^+ + l^+) + (1 - \pi_+) l^+_u \leq \overline{e}_1 + e_2^0 + \kappa \left( B_h - \epsilon B^0_{h,1} \right) + \pi_+ \left( B^1_{h,1} - B^0_{h,1} \right) \tag{74}
\]
\[
  (IC_-) \quad (w^+ + l^+) + (1 - \pi_-) l^+_u \geq \overline{e}_1 + e_2^0 + \kappa \left( B_h - \epsilon B^0_{h,1} \right) + \epsilon \pi_- \left( B^1_{h,1} - B^0_{h,1} \right). \tag{75}
\]

Hence, \( w^+ \) and \( w^- \) can be sustained as an perfect Bayesian separating equilibrium if and only if \( w^- = (w = \overline{e}_1 + e_2^0 + \kappa \left( B_h - \epsilon B^0_{h,1} \right), l = 0, l_u = 0) \) and \( w^+ \) fulfills conditions 71, 74 and 75.

**Separating equilibria passing the intuitive criterion**  The intuitive criterion selects the least cost separating equilibrium with \( w^+ = \overline{e}_1 + e_2^1 + \kappa \left( B_h - B^1_{h,1} \right), l^+ = 0 \) and \( l^+_u \) having the value just sufficiently high to fulfill condition 75, i.e.
\[
  l^+_u = \frac{e_2^0 - e_1^1 + (\kappa + \epsilon \pi_-) B^1_{h,1} - (\kappa + \pi_-) \epsilon B^0_{h,1}}{1 - \pi_-}. \tag{76}
\]

First we proof by contradiction that all perfect Bayesian separating equilibria with \( w^+ + l^+ > \overline{e}_1 + \overline{e}_2 \) do not pass the intuitive criterion. The intuition is that it is simply cheaper for the trusting principal to separate via \( l^+_u \) than by \( w^+ \) or \( l^+ \) as the latter two losses have for both types of principal the same expected value, whereas in case of \( l^+_u \) the trusting principal expects a lowers loss than the distrusting one. More formally, suppose \( (w^+, l^+, l^+_u) \) fulfills conditions 74 and 75 and \( \xi \equiv w^+ + l^+ - \overline{e}_1 - \overline{e}_2 > 0 \). Then the contract \( \hat{w}^+ \equiv
\((w^+ = \bar{w}_1 + \bar{w}_2, \hat{l}^+ = 0, \hat{l}_u^+ = t_u^+ + \frac{\xi}{1 - \frac{\pi_+ \sigma}{\pi_+ - \pi_-}})\) must, by the intuitive criterion lead to the belief \\
\(\alpha_T(\hat{w}^+) = 1\) and thus \(V_+ (\hat{w}^+) > V_+ (w^+)\) which contradicts optimality of \(w^+\).

Hence, we must have \(w^+ = \bar{w}_1 + e_1^1 + \kappa (B_h - B_h^1)\), \(l^+ = 0\), and \\
\[
\frac{\epsilon_2^0 - \epsilon_2^1 + (\kappa + \pi_+) B_h^1 - (\kappa \epsilon + \pi_+) B_h^0}{1 - \pi_+} (IC^+)
\]
\[
\frac{\epsilon_2^0 - \epsilon_2^1 + (\kappa + \epsilon \pi_-) B_h^1 - (\kappa + \pi_-) \epsilon B_h^0}{1 - \pi_-}.
\]

Under the intuitive criterion a contract with \(l_u^+ > \frac{\epsilon_2^0 - \epsilon_2^1 + (\kappa + \epsilon \pi_-) B_h^1 - (\kappa + \pi_-) \epsilon B_h^0}{1 - \pi_-}\) cannot be an equilibrium, as for sufficiently small \(\epsilon > 0\) the contract with \(\hat{l}_u^+ = l_u^+ - \epsilon > \frac{\epsilon_2^0 - \epsilon_2^1 + (\kappa + \epsilon \pi_-) B_h^1 - (\kappa + \pi_-) \epsilon B_h^0}{1 - \pi_-}\) has to lead to \(\alpha_T(\hat{l}_u^+) = 1\) and leads therefore to a higher payoff to the (+)-principal.

Hence \(l_u^+ = \frac{\epsilon_2^0 - \epsilon_2^1 + (\kappa + \epsilon \pi_-) B_h^1 - (\kappa + \pi_-) \epsilon B_h^0}{1 - \pi_-}\).

**Lemma 3** The only separating equilibrium passing the intuitive criterion is the least cost separating equilibrium:

\[
w^+ = \left( w^+ = \bar{w}_1 + e_2^1 + \kappa (B_h - B_h^1), l^+ = 0, l_u^+ = \frac{\epsilon_2^0 - \epsilon_2^1 + (\kappa + \epsilon \pi_-) B_h^1 - (\kappa + \pi_-) \epsilon B_h^0}{1 - \pi_-} \right)
\]
\[
w^- = \left( w^- = \bar{w}_1 + e_2^0 + \kappa (B_h - B_h^0), l^- = 0, l_u^0 = 0 \right).
\]

**Pooling Equilibria** Here, a contract \(w = (w, l, l_u)\) can be sustained by some beliefs as a pooling equilibrium if and only if it can be sustained by the out of equilibrium beliefs \(\alpha_T(w') = 0 \ \forall w' \neq w\). In case all out of equilibrium beliefs are 0 the most attractive alternative to the pooling contract \(w\) is the contract \\
\((w' = (\bar{w}_1 + e_2^0 + \kappa (B_h - \epsilon B_h^0)), l' = 0, l_u = 0)\). In equilibrium the belief of the trustworthy agent is \(\alpha_T(w) = \sigma_T\). The equilibrium contract is therefore accepted by the t-agent if and only if \\
\[
(PC_T) \quad w \geq \bar{w}_1 + e_2^*(\sigma_T) + \kappa (B_h - (\sigma_T (1 - \epsilon) + \epsilon) B_h^{\sigma_T})
\]

35
The incentives to deviate from the equilibrium contract are higher for the distrusting principal as the trusting principal expects lower costs from an \( l_u > 0 \) and values more when the agent beliefs to be trusted. The (-)-principal has no incentive to deviate from the equilibrium contract if and only if

\[
(IC_-) \quad V_-(w, \sigma_T) \geq V_-(w' = (\bar{e}_1 + \bar{e}_2^0 + \kappa (B_h - \epsilon B_h^0)), l' = 0, l'_u = 0, \alpha_T = 0).
\]

Hence a contract \( w = (w, l, l_u) \) is sustainable as a perfect Bayesian pooling equilibrium if and only if

\[
(PC_T) \quad w \geq \bar{e}_1 + \bar{e}_2^0 (\sigma_T) + \kappa (B_h - (\sigma_T (1 - \epsilon) + \epsilon) B_h^\sigma_T),
\]

\[
(IC_-) \quad (w + l) + (1 - \pi_-)l_u \leq \bar{e}_1 + \bar{e}_2^0 + \pi_- \epsilon (B_h^\sigma_T - B_h^0) + \kappa (B_h - \epsilon B_h^0).
\]

In such a pooling equilibrium the payoffs for the (+) and (-) principal are respectively

\[
\hat{V}_+^{pool}(w, l, l_u) = -(w + l) - (1 - \pi_+)l_u + \pi_+ B_h^\sigma_T
\]

\[
\hat{V}_-^{pool}(w, l, l_u) = -(w + l) - (1 - \pi_-)l_u + \pi_- \epsilon B_h^\sigma_T.
\]

None of the pooling equilibria passes the intuitive criterion  In these pooling equilibria the intuitive criterion demands that an out of equilibrium belief \( \alpha_T(w', l', l'_u) = 1 \) if the (-)-principal does for any belief strictly worse under the alternative contract \( (w', l', l'_u) \) compared to her equilibrium payoff and if the (+)-principal may profit for some beliefs. In other words \( \alpha_T(w', l', l'_u) = 1 \) if

\[
(IC_-) \quad \hat{V}_-^{pool}(w, l, l_u) > \hat{V}_-(w', l', l'_u, \alpha_T = 1)
\]

\[
(IC_+) \quad \hat{V}_+^{pool}(w, l, l_u) \leq \hat{V}_+(w', l', l'_u, \alpha_T = 1),
\]

36
or equivalently
\[
(IC_-) \quad (w' + l') - (w + l) + (1 - \pi_-)(l'_u - l_u) > \pi_- \epsilon (B_1^1 - B_1^\alpha) \\
(IC_+) \quad (w' + l') - (w + l) + (1 - \pi_+)(l'_u - l_u) \leq \pi_+ (B_1^1 - B_1^\alpha). 
\]

Consider e.g. the contract \((w' \equiv w, l' \equiv l, l'_u \equiv l_u + \pi_+ + 1 - \pi_- + \epsilon (B_1^1 - B_1^\alpha))\). The intuitive criterion demands \(\alpha_T(w', l', l'_u) = 1\) and then the trusting principal does strictly better when proposing contract \(w'\) then in the pooling equilibrium. Hence, none of the pooling equilibria survives the intuitive criterion.

**No Hybrid equilibrium passes the intuitive criterion**  The argument is similar to the pooling-equilibria. In a hybrid equilibrium there must exist at least one contract \((w, l, l_u)\) that is chosen by both types of principal with a strictly positive probability, and hence \(0 < \alpha_T(w, l, l_u) < 1\). Consider e.g. the contract \((w' \equiv w, l' \equiv l, l'_u \equiv l_u + \pi_+ + 1 - \pi_- + \epsilon (B_1^1 - B_1^\alpha))\). The intuitive criterion demands \(\alpha_T(w', l', l'_u) = 1\) and then the trusting principal does strictly better when proposing contract \((w', l', l'_u)\) then in the equilibrium contract \((w, l, l_u)\). Hence, none of the hybrid equilibria passes the intuitive criterion.

**Proof of Lemma 2**

Let \(w \equiv (w, \Delta w)\) denote a contract that pays the agent a wage of \(w\) in case of \(B_1 = 0\) and a wage of \(\overline{w} = w + \Delta w\) in case of \(B_1 = \overline{B}_1\). We want to show that under the No-Screening Condition 3 the principal can not even partially separate the trustworthy from the untrustworthy agent by a menu of contracts \(\{(w_U, \Delta w_U), (w_T, \Delta w_T)\}\) (from which the agent can choose his preferred contract). Let \(\pi_{\pm}(w_i)\) denote the belief of the \((\pm)\)-principal that the agent is trustworthy when the agent chose the contract \(w_i\) from the menu. Furthermore, let \(\alpha_{T/U}(w_i)\) denote the belief of the \((T/U)\)-agent, that the principal trusts him (in the trust-game at the last stage) when he has chosen contract \(w_i\). The expected utility of the
agent from choosing contract \( w_1 \) are in dependence of his type

\[
U_{\ell}(w_1) = w_1 + \max\{0, \Delta w_1 - \bar{e}_1\} + (\alpha_{\ell}(w_1)(1 - \epsilon) + \epsilon) M_d \\
U_T(w_1) = w_1 + \max\{0, \Delta w_1 - \bar{e}_1 + \kappa B_1\} - e_2^*(\alpha_T(w_1)) + \kappa B_h^{\alpha_T(w_1)}(\alpha_T(w_1)(1 - \epsilon) + \epsilon) - \kappa B_{\max}.
\]

Necessary conditions for the menu \( \{(w_\ell, \Delta w_\ell), (w_T, \Delta w_T)\} \) to screen (or partially screen) between both types of agents are

\[
(\text{IC}_\ell) \quad U_{\ell}(w_\ell) \geq U_{\ell}(w_T) \tag{85}
\]

\[
(\text{IC}_T) \quad U_T(w_\ell) \leq U_T(w_T). \tag{86}
\]

Equivalently

\[
\max\{0, \Delta w_T - \bar{e}_1\} - \max\{0, \Delta w_\ell - \bar{e}_1\} + (1 - \epsilon) (\alpha_{\ell}(w_T) - \alpha_{\ell}(w_\ell)) M_d^{(\text{IC}_\ell)} \leq w_\ell - w_T
\]

\[
\leq \max\{0, \Delta w_T - \bar{e}_1 + \kappa B_1\} - \max\{0, \Delta w_\ell - \bar{e}_1 + \kappa B_1\} - (e_2^*(\alpha_T(w_T)) - e_2^*(\alpha_T(w_\ell))) + (\alpha_T(w_T)(1 - \epsilon) + \epsilon) \kappa B_h^{\alpha_T(w_T)} - (\alpha_T(w_\ell)(1 - \epsilon) + \epsilon) \kappa B_h^{\alpha_T(w_\ell)}
\]

A necessary condition for the existence of an \( w_T \) and \( w_\ell \) for which \( (\text{IC}_\ell) \) and \( (\text{IC}_T) \) both hold is that

\[
(\alpha_{\ell}(w_T) - \alpha_{\ell}(w_\ell)) M_d \leq \max\{0, \Delta w_T - \bar{e}_1 + \kappa B_1\} - \max\{0, \Delta w_\ell - \bar{e}_1\} + \max\{0, \Delta w_\ell - \bar{e}_1\} - \max\{0, \Delta w_T - \bar{e}_1 + \kappa B_1\} - (e_2^*(\alpha_T(w_T)) - e_2^*(\alpha_T(w_\ell))) + (\alpha_T(w_T)(1 - \epsilon) + \epsilon) \kappa B_h^{\alpha_T(w_T)} - (\alpha_T(w_\ell)(1 - \epsilon) + \epsilon) \kappa B_h^{\alpha_T(w_\ell)}.
\]

The first line on the right hand side is always greater or equal to \( B_1 \) and the second line is always greater or equal to 0. A necessary condition for that \( (\text{IC}_\ell) \) and \( (\text{IC}_T) \) can both hold
\[ (\alpha_U(w_T) - \alpha_u(w_U)) M_d \leq \kappa B_1 + (\alpha_T(w_T) (1 - \epsilon) + \epsilon) \kappa B_h^{\alpha_T(w_T)} \]
\[ - (\alpha_T(w_U) (1 - \epsilon) + \epsilon) \kappa B_h^{\alpha_T(w_U)} \]
\[ - (c^*_2(\alpha_T(w_T)) - c^*_2(\alpha_T(w_U))) \]
\[ \leq \kappa B_1 + \kappa B_h, \tag{87} \]

where we used \( 0 \leq \alpha_T(w_U) \leq \alpha_T(w_T) \leq 1 \) and \( B_h^1 < B_h \) for the last inequality. If the trustworthy agent would be fully separated from the untrustworthy, then Bayesian updating requires: \( \alpha_U(w_T) = 1 \) and \( \alpha_U(w_U) = 0 \) and therefore \( (\alpha_U(w_T) - \alpha_U(w_U)) = 1 \).

In the hybrid case in which the trustworthy agent chooses contract \( w_T \) for sure and the untrustworthy agent is indifferent between both contracts and mixes with some probability we have \( \alpha_U(w_U) = 0 \) and \( \alpha_U(w_T) \geq \sigma_U \), hence \( (\alpha_U(w_T) - \alpha_U(w_U)) \geq \sigma_U \).

In the hybrid case in which the untrustworthy agent chooses contract \( w_U \) for sure and the trustworthy agent is indifferent between both contracts and mixes with some probability we have \( \alpha_U(w_U) \leq \sigma_U \) and \( \alpha_U(w_T) = 1 \), hence \( (\alpha_U(w_T) - \alpha_U(w_U)) \geq 1 - \sigma_U \).

In any case holds \( (\alpha_U(w_T) - \alpha_U(w_U)) \geq \max\{\sigma_U, (1 - \sigma_U)\} \). Together with condition 87 we have shown that screening is impossible if

\[ M_d > \frac{\kappa (B_1 + B_h)}{\min\{\sigma_U, (1 - \sigma_U)\}}. \tag{88} \]

### Mapping on the Reduced Form Contract Proposal

After writing down the expected payoffs for the different contracts for the (++)-principal, the (--)principal, and the trustworthy agent we show in step 1 that for any contract-proposal \( \mu((w_P, w_P, w_A, w_A)) \) there exist a payoff equivalent reduced-form contract \( (w, l, l_u) \). In step 2 we show that for every reduced-form contract \( (w, l, l_u) \), there is at least one payoff equivalent contract-proposal \( \mu((w_P, w_P, w_A, w_A)) \) that gives at least the trustworthy agent the incentive to choose \( e_1 = \tau_1 \). In step 3 we show that (for a given \( \alpha_T \)) the expected payoffs under
two reduced-form contracts \((w, l, l_u)\) and \((w', l', l'_u)\) are equal for the trusting principal, the distrusting principal, and the trustworthy agent only if \((w, l, l_u) = (w', l', l'_u)\).

Within this subsection we only want to establish payoff equivalences for given beliefs \(\alpha_T\). Then neither the second investment \(e_2\) nor the behavior in the trust game are influenced by the contract. For this section we can therefore simplify the analysis by re-normalizing each players utility by substracting the expected payoff resulting from investment \(e_2\) and the trust-game. Then the expected (re-normalized) payoffs of an original-form deterministic contract \((w_p, w_p, w_A, w_A)\) (with \(\Delta w_A \equiv w_A - w_A\)):

1. If \(\Delta w_A > \bar{\tau}_1\) (complete contract: both type of agents: \(e_1 = \bar{\tau}_1\)):

\[
\begin{align*}
\check{V}_+(w_p, w_p, w_A, w_A) &= -w_p + B_1 \\
\check{V}_-(w_p, w_p, w_A, w_A) &= -w_p + B_1 \\
\check{U}_T(w_p, w_p, w_A, w_A) &= w_A - \bar{\tau}_1
\end{align*}
\]  

2. If \(\bar{\tau}_1 \geq \Delta w_A > \bar{\tau}_1 - B_1\) (incomplete contract: t-agent: \(e_1 = \bar{\tau}_1\), u-agent: \(e_1 = 0\)):

\[
\begin{align*}
\check{V}_+(w_p, w_p, w_A, w_A) &= \pi_+(-w_p + B_1) - (1 - \pi_+)w_p \\
\check{V}_-(w_p, w_p, w_A, w_A) &= \pi_-(-w_p + B_1) - (1 - \pi_-)w_p \\
\check{U}_T(w_p, w_p, w_A, w_A) &= w_A - \bar{\tau}_1
\end{align*}
\]  

3. If \(\Delta w_A < \bar{\tau}_1 - B_1\) (both type of agents choose \(e_1 = 0\)):

\[
\begin{align*}
\check{V}_+(w_p, w_p, w_A, w_A) &= -w_p \\
\check{V}_-(w_p, w_p, w_A, w_A) &= -w_p \\
\check{U}_T(w_p, w_p, w_A, w_A) &= w_A - B_1.
\end{align*}
\]

Notice any contract \((w_p, w_p, w_A, w_A)\) of the last category \((\Delta w_A < \bar{\tau}_1 - B_1)\) has always a corresponding payoff-equivalent contract \((w'_p, w'_p, w'_A, w'_A)\) in the first category \((\Delta w_A > \bar{\tau}_1)\),
e.g. \( \bar{w}_P \equiv w_p + B_1, \bar{w}_A \equiv w_A + \bar{v}_1 - B_1 \) and \( \bar{w}_P' \equiv \bar{w}_A' < \bar{w}_A - \bar{v}_1 \). Hence, we can restrict the analysis to contracts of categories 1 and 2. By “complete contract” we denote a contract of category 1 and by “incomplete contract” we denote a contract of category 2.

Now consider the general case of a probability distribution \( \mu \) over a set of contracts of category 1 or 2. Let \( \mu^c \) denote the total mass of complete contracts and therefore \( 1 - \mu^c \) the total mass of incomplete contracts. Let \( \bar{w}_A \) denote the expected value of the \( \bar{w}_A \) over all contracts. Furthermore, let \( \bar{w}_P^i \) and \( \bar{w}_P \) denote the expected values of \( w_P \) and \( \bar{w}_P \) conditional on having an incomplete contract. Correspondingly, let \( \bar{w}_P^c \) denote the expected values of \( w_P \) conditional on having a complete contract. Then the expected payoffs of a random contract-proposal \( \mu(w_p, \bar{w}_P, \bar{w}_A, \bar{w}_A) \) are

\[
\begin{align*}
\hat{V}_+(\mu) &= \mu^c(-\bar{w}_P + B_1) + (1 - \mu^c) (\pi_+ ( -\bar{w}_P + B_1 ) - (1 - \pi_+) \bar{w}_P^i) \\
\hat{V}_-(\mu) &= \mu^c(-\bar{w}_P + B_1) + (1 - \mu^c) (\pi_- ( -\bar{w}_P + B_1 ) - (1 - \pi_-) \bar{w}_P^i) \\
\hat{U}_T(\mu) &= \bar{w}_A - \bar{v}_1,
\end{align*}
\]

or equivalently

\[
\begin{align*}
\hat{V}_+(\mu) &= -(\mu^c \bar{w}_P + (1 - \mu^c) \bar{w}_P^i) - (1 - \pi_+) (1 - \mu^c) (B_1 - \Delta \bar{w}_P^i) + B_1 \\
\hat{V}_-(\mu) &= -(\mu^c \bar{w}_P + (1 - \mu^c) \bar{w}_P^i) - (1 - \pi_-) (1 - \mu^c) (B_1 - \Delta \bar{w}_P^i) + B_1 \\
\hat{U}_T(\mu) &= \bar{w}_A - \bar{v}_1.
\end{align*}
\]

The expected (re-normalized) payoffs from a reduced-form contract \((w, l, l_u)\) are for the (+)-principal, (-)-principal and the trustworthy agent are

\[
\begin{align*}
\hat{V}_+(w, l, l_u) &= -(w + l) - (1 - \pi_+) l_u + B_1, \\
\hat{V}_-(w, l, l_u) &= -(w + l) - (1 - \pi_-) l_u + B_1, \\
\hat{U}_T(w, l, l_u) &= w - \bar{v}_1.
\end{align*}
\]
**Step 1:** For a given original-form contract proposal $\mu$ we choose $w \equiv \overline{w}_A$, 
\[ l \equiv (\mu c \overline{w}_P + (1 - \mu c) \overline{w}_P) - \overline{w}_A, \] 
and $l_u \equiv (1 - \mu^c)(\overline{B}_1 - \Delta w'_P)$. Then $l \geq 0$ and $l_u \geq 0$ and the payoffs are equal to the original-form contract for both types of principal and the trustworthy agent.

**Step 2:** For a given reduced-form contract $(w, l, l_u)$ distinguish two cases

**Case 1:** If $l_u \geq \overline{B}_1 - \overline{c}_1$
we can choose the deterministic incomplete contract $\overline{w}_A \equiv w$, $\overline{w}_P \equiv w + l$, $\overline{w}_P \equiv w + l_u - \overline{B}_1$
and $w_A \equiv w - \overline{c}_1 - \epsilon$ with $\epsilon \in ]0, \overline{B}_1 - \overline{c}_1[$.

**Case 2:** If $l_u < \overline{B}_1 - \overline{c}_1$
we can choose a stochastic contract proposal mixing between two contracts:
With probability $\mu^c$ the complete contract $\overline{w}_A \equiv w$, $\overline{w}_P \equiv w + l$, $\overline{w}_P \equiv w_A \equiv w - \overline{c}_1 - \epsilon$ with $0 < \epsilon < \overline{B}_1 - \overline{c}_1$ is drawn.
With the counter probability $(1 - \mu^c) \equiv \frac{l_u}{(1 - \mu^c) \overline{B}_1 - l_u + \overline{c}_1}$ the incomplete contract $\overline{w}_A \equiv w$, $\overline{w}_P \equiv w + l$, $\overline{w}_P \equiv w + \overline{B}_1 - \overline{c}_1$, and $w_A \equiv w - \overline{c}_1 + \epsilon$ with $0 < \epsilon < \overline{B}_1$.

**Step 3:** Consider two reduced-form contracts $(w, l, l_u)$ and $w', l', l'_u$ with equal expected payoffs for (+)-principal, (-)-principal and (t)-agent. Then $w = w'$ due to equation 106 and $l' + (1 - \pi_+) (l'_u - l_u) = l = l' + (1 - P - \overline{\pi}_-) (l'_u - l_u)$, due to equations 104 and 105. Since $\pi_+ \neq \pi_-$ this equations can only hold if $l'_u = l_u$ and $l' = l$, q.e.d.

**References**


[14] Fehr, Ernst and Schmidt, Klaus (2000), “Theories of Fairness and Reciprocity - Evidence and Economic Applications”, (paper prepared for the invited session of the 8th World Congress of the Econometric Society)


[16] Friebel Guido, ????


B Notation

\( p \) : \( \text{prob}(\text{agent trustworthy}) \)

\( (1 - p) \) : \( \text{prob}(\text{agent untrustworthy}) \)

\( p_+ \) : \( \text{prob}(\text{agent trustworthy} \mid \text{signal} = +) \)

\( p_- \) : \( \text{prob}(\text{agent trustworthy} \mid \text{signal} = -) \)

\( \Pi_+ \) : Principals belief that agent is trustworthy given signal = +

\( \Pi_- \) : Principals belief that agent is trustworthy given signal = -

\( a_t \) : \( \text{prob}(\text{Principal’s signal} = + \mid \text{agent trustworthy}) \)

\( a_u \) : \( \text{prob}(\text{Principal’s signal} = + \mid \text{agent untrustworthy}) \)

\( \alpha_t \) : \( \text{prob}(\text{Principals belief} \ \Pi \geq d \mid \text{agent trustworthy}) \)

\( \alpha_u \) : \( \text{prob}(\text{Principals belief} \ \Pi \geq d \mid \text{agent untrustworthy}) \)