Contractual incompleteness as a signal of trust

Florian Herold
Kellogg School of Management, Northwestern University, 2001 Sheridan Rd., Evanston, IL 60208, United States

ARTICLE INFO

Article history:
Received 1 July 2008
Available online 15 May 2009

JEL classification:
D86
D82

Keywords:
Trust
Incomplete contracts
Signaling

ABSTRACT

This paper shows how the fear of signaling distrust can endogenously lead to incomplete contractual agreements. We consider a principal agent relationship where the agent may be trustworthy (dedicated to the project) or not. The principal may trust the agent (i.e. have a high belief of facing a trustworthy agent), or distrust him. The proposal of a complete contract, including fines and other explicit incentives, is shown to signal distrust. When trust is important in some non-contractible part of the relationship, a principal may prefer to leave the contract incomplete rather than to signal distrust by proposing a complete contract. Contractual incompleteness arises endogenously due to asymmetric information about how much one partner trusts the other side.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

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? In this paper I show that designing a sophisticated complete contract, which enforces certain behavior with fines, punishments, and other explicit incentives, can signal distrust to the other party – even in a setting with completely standard preferences. Trust, however, is of crucial importance for the functioning of many economic and noneconomic relationships. A principal may therefore prefer to leave a contract incomplete rather than to signal distrust by proposing a complete contract.

The perception that a contract proposal signals distrust seems to be widespread. Consider e.g. the following quote from a review of a book on the patient’s choices in cancer treatment. The reviewer (among other things) is skeptical about the author’s suggestion that before treatment patients should ask the oncologist to sign a contract. “[...] [The author] also suggests a contract between patient and physician [...] which specifies the responsibilities of the oncologist and the patient and requires the signature of both. Although such a contract may be pragmatic and useful, it suggests a mutual distrust between patient and physician and has all the negative implications of a prenuptial contract. In short, the author suggests that the inherent distrust is a wise posture in conventional therapy. [...]”2 Thus the reviewer sees trust as an important prerequisite for successful treatment and is worried that proposing a contract might erode this trust. The quote implies also that the same problem is often considered relevant for prenuptial agreements.

E-mail address: f-herold@northwestern.edu.

1 See e.g. Holmstrom (1982) or Laffont and Tirole (1993).

2 See Fitzgerald (1994).
As a further example consider a scientist hiring a research assistant. Simple contractual arrangements can solve some incentive problems. 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 she 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 her scientific dedication. The belief 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.

There are two ways how to model this idea. One way would be to make behavioral assumptions on trust, the perception of trust, and the value of feeling trusted. While I do think there is a psychological or behavioral appeal of the main argument, this paper focuses on showing that the main results can be derived under completely standard assumptions on the players’ preferences.

Consider a principal (“she”) who is interested in the success of a project that she can only realize with the help of an agent (“he”). There are two types of agents. One type has an intrinsic interest in the success of the project and is therefore willing to exert effort even without a contract prescribing high effort. I shall call this first type trustworthy.3 The other type, the untrustworthy agent, is not intrinsically interested in the success of the project. Only contractual arrangements can make him exert effort. The principal may hold different beliefs about the agent’s type. These different beliefs are either heterogeneous prior beliefs (representing variety in dispositional trust) or they are rational probability assessments which the principal derived by Bayes’ rule after receiving some private signals. I call the principal’s belief that the agent is trustworthy the “trust” of the principal in the agent.4 A more trusting principal has a stronger belief that the agent is the trustworthy type who exerts high effort even in the absence of a contract. So a trusting principal has lower expected costs from contractual incompleteness. She is therefore able to separate herself from less trusting types by using contractual incompleteness as a signaling device.

Why is the principal interested in signaling trust? In my model trust is relevant in a part of the relationship which is non-contractible by assumption. In this non-contractible part, principal and agent simultaneously choose actions that contribute to the success of the project. These actions are either complements or substitutes. In both cases the agent works harder if he believes he is trusted by the principal, although for different reasons. In both cases a trusting principal expects the agent to work harder. Under complementarity this means a trusting principal works harder, and so an agent who believes he is trusted works harder too. If efforts are substitutes, a trusting principal works less (or redirects her efforts to profit the project in ways unrelated to the agent’s effort). So an agent who believes he is trusted works harder, because he expects lower effort by the principal and efforts are substitutes. Thus, in both cases the principal wants to signal trust to encourage stronger efforts by the agent in the non-contractible part of the relationship, independently of whether efforts are complements or substitutes. In summary, we need that there is a non-contractible part of the relationship, that the agent’s type is his private information, and that the principal’s type, i.e. her belief about the agent’s type, is her private information. Then we get incomplete contracts even in the completely contractible part of the relationship, because the principal is afraid of sending a signal of distrust which would have negative consequences in the non-contractible parts of the relationship.5

The assumption of a non-contractible part of the relationship is not necessary for my argument if we are willing to depart from standard assumptions on the agent’s preferences. We only need that the agent reacts negatively if he believes he is distrusted. I discuss such psychological or behavioral reasons for reacting negatively when perceiving distrust after presenting the main model in which players have standard economic preferences.

The literature on the foundations of incomplete contracts is extensive.6 Spier (1992) already points out that signaling can cause contractual incompleteness.7 In her model, a risk-averse principal hires a risk-neutral agent. The principal has private information on the probability that her project is profitable. 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 probability of success is high. In general, however, such asymmetric information at the contracting stage can equally lead to more complete rather than

---

3 In a more behavioral model, the trustworthy agent could also follow a norm of “decent” behavior, feel committed to unwritten agreements, or be motivated by some fairness consideration.

4 This notion of trust differs somewhat from the terminology used in the experimental literature on the trust game, where trust often refers to an action. Both notions of trust are closely related though, because trust in our sense (a high belief of facing a trustworthy type) typically results in the action “trust” in a trust game.

5 Holmström and Milgrom (1992) and Bernheim and Whinston (1998) give two different arguments, as to why it can be optimal to leave some verifiable aspects of a relationship unspecified when other aspects cannot be verified. Holmström and Milgrom (1992) shows 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 and Whinston’s (1998) 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 off-equilibrium punishments may be necessary to sustain the desired equilibrium.

6 A good survey of this literature is Tirole (1999).

7 See also Allen and Gale (1994) for similar ideas in the context of financial economics. See Spence (1973, 1974) for the first formal analysis of signaling games.
less complete contracts. See e.g. Tirole (1999), p. 764, for this point. In this paper, in contrast, there is a clear prediction that contracts are less complete when the principal wants to signal trust. The trustworthy type is defined by having preferences which make him choose the desired action anyway, 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, this 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, controls, and explicit incentives can crowd out intrinsic motivation and may even be counterproductive. Sanctions seem to have a particularly detrimental effect if they are deliberately designed by one of the involved parties.

Bénabou and Tirole (2003) is perhaps the work most closely related to the approach of this paper. In their model a principal has private information about the agent’s type. The agent tries to infer information about this own type from the principal’s actions. A potential conflict between extrinsic incentives and intrinsic motivation arises because giving an agent high-powered incentives may signal bad news about his ability or the task.

The mechanism and intuition why explicit incentives can be counterproductive in Bénabou and Tirole (2003) differs however substantially from this paper. First, the term “trust” has a very different meaning in the two papers. In Bénabou and Tirole (2003) “trust” refers to the knowledge that the agent has a high ability or low costs. Hence their model is really about managing the agent’s self-confidence. Here, in contrast, trust refers to the principal’s belief in the agent’s trustworthiness. It is an argument why formal contracts can erode trust in a relationship and can thus be counterproductive. A second, related difference between the two papers is that in Bénabou and Tirole (2003) the principal needs to have private information about the agent’s type. Correspondingly, their applications focus on child rearing, education, or a task which is completely new to the agent. In contrast, this paper works with the standard informational assumption that each player is better informed about his own type and her own beliefs. Typically a principal hires an agent because of his expertise and superior knowledge how to handle the task in question. An analysis along the lines of Bénabou and Tirole (2003) would not raise any concerns about providing explicit incentives in such situations, while, from the current paper’s perspective, explicit contracts are counterproductive when trust is important for the relationship. Trust in the other party often influences the probability of having future joint business and thereby the incentives to make e.g. unobservable, relation specific investments. Consider e.g. an entrepreneur who identified a customers’ need for a new gadget. While she understands the marketing side of the new product she wants to hire an engineer who understands the technical viability of potential features of this new gadget. The engineer is hired for his expertise and unlikely to learn about his ability from the entrepreneur. Yet, designing a useful gadget that satisfies customers’ needs with a viable cost efficient gadget requires trust between entrepreneur and engineer to enable a free exchange of their ideas and assessments. The more strongly the engineer believes to be trusted, the more he expects the entrepreneur to use these ideas and to exert the necessary complementary effort. Thus there are different important channels how explicit incentives can have adverse effects: either, à la Bénabou and Tirole (2003), by reducing the agent’s self-esteem, or by eroding trust between the principal and the agent, as highlighted in this paper. Finally, in Bénabou and Tirole (2003) explicit incentives can signal trust or distrust in the agent’s ability, depending on whether explicit incentives are more effective for the high or low type of agent. In the current setting, in contrast, contractual incompleteness signals always trust. Note also that nothing in the current setting prescribes ex ante that “no contract” should be a signal of trust and “contract” should be a signal of distrust – and yet signaling in the opposite way (i.e., “contract” as a signal of trust) is never equilibrium behavior.

There exist further, different arguments in the literature why explicit incentives can be counterproductive. Bénabou and Tirole (2006) considers a model with heterogeneity in individual altruism and concerns for social reputation. Explicit incentives can then crowd out prosocial behavior as they create doubt about the true motives for which good deeds are performed. In a similar spirit, recent work by Ellingsen and Johannesson (2008) considers a model in which agents care about social esteem, i.e. being perceived as altruistic. Furthermore, altruists care more about social esteem from other altruists. In a situation where limited control is possible at some costs, the choice not to control can signal altruism, and can therefore induce a responder to reciprocate in the hope of gaining the esteem of this altruist. These approaches provide a quite different, more behavioral mechanism why explicit incentives can be counterproductive. They apply to circumstances where gaining social esteem by doing good deeds is an important motive, whereas my model provides an argument concerned with interpersonal trust in a setting of standard preferences.

Furthermore, Sliwka (2007) considers a model with three types of agents: selfish agents, fair agents, and conformists. Conformists cooperate only if they believe to be in an environment with more fair than selfish agents. Explicit incentives by the better informed principal signal an environment of mostly selfish agents and may result in defection of the conformists. My model differs in that the principal signals her trust in the agent’s intrinsic motivation and works with perfectly standard
preferences whenever there is heterogeneity in intrinsic motivation. A recent working paper, Friebl and Schnedler (2007), focuses on organizational design and compares empowerment and hierarchical control in team governance. Low hierarchies signal that co-workers are considered to be motivated and may increase deliberate team work. Furthermore, Friebl and Schnedler (2007) adds also to the literature by considering dynamic aspects of the effects of the organizational design.

Finally, Fehr, Klein and Schmidt (2007) address the question of how fairness concerns affect the choice of contracts. They find in experiments 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 that the experimental results are consistent with a heterogeneous population of subjects, where some players are inequity-averse while others act selfishly. My argument can complement their explanation, if we are willing to go beyond purely standard economic preferences make additional behavioral assumptions which ensure that a signal of distrust leads to a negative reaction even in the absence of a non-contractible part. I discuss such behavioral approaches in Section 3. It is natural in their setting to define trust 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.11

The main model is presented in Section 2. Section 3 provides a discussion of the results and of further extensions.

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 (D); the proportion of trustworthy types in the economy, \( \pi \), is strictly between zero and one. A trustworthy agent is intrinsically interested in the success of the project, whereas an untrustworthy agent does not care about the project per se.

The principal cannot observe the agent’s type. She holds a belief about the agent’s type which can be low or high, i.e. \( \pi_1 < \pi_0 \). These beliefs are private information of the principal and can be interpreted as the principal’s type. This heterogeneity in prior beliefs could just represent variety in the principal’s dispositional trust. Here, I assume instead that this heterogeneity in trust comes from a private, binary signal \( s \in \{l, h\} \) about the agent’s type which the principal received and from which she rationally derived her beliefs about the agent’s type. When the agent is trustworthy, the principal receives a high signal \( h \) with the exogenously given probability \( \sigma_T \). When the agent is untrustworthy, the principal receives a high signal with the exogenously given probability \( \sigma_D \), with \( \sigma_D < \sigma_T \). By Bayes’ rule a principal with a high signal believes that she faces the trustworthy type with probability \( \pi_h = \pi / (\pi + 1 - \sigma_D (1 - \pi)) > \pi \), and a principal with a low signal believes that she interacts with a trustworthy type with probability \( \pi_l = \pi / (\pi + \sigma_D (1 - \pi)) < \pi \). Notice that \( \pi_h > \pi_l \), i.e. a principal with a high signal has a stronger belief in the agent’s trustworthiness. In other words, a principal with a high signal trusts the agent more than a principal with a low signal.

The project consists of two parts, the contractible part 1 and the non-contractible part 2. The total success \( B \) of the project is the sum of the successes, \( B_1(\alpha) \) and \( B_2(e, f) \), in both parts of the relationship, i.e. \( B(\alpha, e, f) = B_1(\alpha) + B_2(e, f) \).

The project’s success in the contractible part 1, \( B_1(\alpha) \in [0, B_1] \), depends, for simplicity, only on an unobservable effort \( \alpha \in [0, \bar{\alpha}] \) by the agent. High effort \( \bar{\alpha} \) benefits the project deterministically by \( B_1 \equiv B_1(\bar{\alpha}) > \bar{\alpha} \). Low effort \( \alpha = 0 \) leads to \( B_1 = 0 \). A contract can be written contingent on the outcome \( B_1(\alpha) \) 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 \( \alpha \) is not directly observable, it can be inferred from the realized value of \( B_1 \).12 A sufficiently harsh punishment in case of \( B_1 = 0 \) therefore implements a high effort level \( \alpha = \bar{\alpha} \) in this contractible part 1.

The success of the project in part 2, \( B_2(e, f) \), depends on an unobservable effort choice by the agent, \( e \in [0, \infty) \), as well as on an unobservable effort choice by the principal, \( f \in [0, \infty) \). Higher efforts by the agent or the principal increase the expected success of the project.

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 levels \( \alpha \) and \( e \) and the principal chooses her effort level \( f \). Effort choices are unobservable, yet \( B_1(\alpha) \) is observable after all effort choices have been made. By assumption, a contract cannot be written contingent on the outcome \( B_2(e, f) \) in the non-contractible part 2. The timing of events is illustrated in Fig. 1.

**Contracting stage.** At the contracting stage, the principal proposes a contract. By assumption this contract can only be written contingent on \( B_1(\alpha) \), the success of the project in the first part. The most relevant feature of this contract is whether the contract enforces high effort \( \bar{\alpha} \) by a sufficiently harsh punishment in case of \( B_1(\alpha) = 0 \). I want to demonstrate the main point as concisely as possible and restrict the set of contractual choices of the principal to a binary choice.

---

11 In the related work Fehr and Schmidt (2007) principals can either choose to make only a non-binding bonus promise or to add an imperfect effort enforcement. The later instrument does not significantly increase agents effort, which is difficult to understand by inequity aversion alone and makes this complimentary argument even more relevant.

12 By assuming that \( \alpha \) is only ex post (indirectly) observable I avoid complication of \( \alpha \) signaling something about the agent’s type.
\(\mathcal{C} \in \{\text{contract} (c), \text{no contract} (n)\}\). This paper neglects more sophisticated mechanisms that the principal might try to use, e.g. in the hope of screening the agent’s type. The possibility of such sophisticated contracts causes difficulties in most of the applied contracting literature and the technical problems that arise with the general mechanism design problem are beyond the scope of this paper. A somewhat more general setting in which the principal can propose any wage scheme which depends only on \(B_1\) is analyzed in detail in the working paper version of this paper, Herold (2008). The results are briefly discussed in Section 3.

The contract pays the exogenous wage \(\bar{w}\) if \(B_1(a) = B_1\) and \(\bar{w} - \Delta w\) if \(B_1(a) = 0\) whereas no contract pays the wage \(\bar{w}\) in either case. Let \(\Delta w > \bar{a}\). Then, if there is a contract, the threat of the wage reduction in case of low effort is sufficient to enforce high effort \(a = \bar{a}\) in the contractible part 1. We can interpret this as a situation where principal and agent are already working together. They have a binding agreement, and this existing agreement gives the principal the discretion to enforce the high effort of the agent in the contractible part 1 of the relationship through a contract or to abstain from doing so. Notice that here the agent takes no observable action. After the first exogenous signal \(s\), the principal’s belief \(\pi_i\), \(i \in \{l, h\}\), about the agent’s type (i.e. her trust) does not change.

**Preferences.** For simplicity, let the principal and the agent be risk neutral. The untrustworthy type of agent maximizes his monetary payoff \(w\) minus his total effort costs \((a + e)\). He does not care about the project. The utility function of the untrustworthy agent is given by

\[
U_U(w, e, B(\cdot)) = w - a - e, \tag{1}
\]

where \(w\) is the wage paid by the principal. In case of a success in the contractible part \((B_1(a) = B_1)\), the wage \(w\) is given by \(\bar{w}\), which, here in the simplest model, is an exogenously given level. If, on the other hand \(B_1 = 0\), in which case the agent must have shirked, then the wage \(w\) drops to \(\bar{w} - \Delta w\) if the principal decided to write a contract and stays at \(\bar{w}\) if the principal did not write a contract.

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

\[
U_T(w, e, B(\cdot)) = w - a - e + \kappa B(\cdot). \tag{2}
\]

The principal’s utility is given by

\[
V(w, f, B(\cdot)) = B(\cdot) - f - w. \tag{3}
\]

The principal maximizes simply the expected total success of the project \(B(\cdot)\) minus the wage payment \(w\) and minus her effort costs \(f\).

**Assumptions.** The structure of the game including \(\sigma_T, \sigma_U, \pi\), and the payoff functions of the different types are assumed to be common knowledge. The trustworthy agent needs to care sufficiently about the project to exert deliberately high effort \(\bar{a}\).

**Assumption 1.** \(\kappa \geq \frac{\sigma}{\bar{a}}\).

Furthermore, a few technical assumptions simplify the analysis.

**Assumption 2.** \(B_2(\cdot, \cdot)\) is twice continuously differentiable with

(a) \(\frac{\partial B_2(\cdot, \cdot)}{\partial e} > 0\), \(\frac{\partial B_2(\cdot, \cdot)}{\partial f} > 0\), \(\frac{\partial^2 B_2(\cdot, \cdot)}{(\partial e)^2} < 0\), \(\frac{\partial^2 B_2(\cdot, \cdot)}{(\partial f)^2} < 0\) (thus \(B_2(\cdot, \cdot)\) is increasing and concave in each variable); \(\frac{\partial B_2(0,0)}{\partial e} > 1\), \(\frac{\partial B_2(0,0)}{\partial f} > 1\), and \(B_2\) is bounded;

(b) \(|\frac{\partial^2 B_2(\cdot, \cdot)}{(\partial e)^2}| > |\frac{\partial^2 B_2(\cdot, \cdot)}{(\partial e)(\partial f)}|\) and \(|\frac{\partial^2 B_2(\cdot, \cdot)}{(\partial f)^2}| > |\frac{\partial^2 B_2(\cdot, \cdot)}{(\partial e)(\partial f)}|\).

Part (a) of Assumption 2 guarantees an interior solution. For given beliefs, the dominant diagonal condition in part (b) guarantees a unique equilibrium of the non-contractible part of the interaction. The uniqueness of the solution is not essential, yet it simplifies the analysis.

Fig. 1. Timeline.
The agent's and the principal's efforts in the non-contractible part of the interaction could be either complements or substitutes. In both cases, the desire to signal trust leads to contractible incompleteness and the results go through.

Assumption 3. $\frac{\partial^2 B_2(\cdot, \cdot)}{\partial e \partial f} > 0$, i.e. the effort $f$ of the principal and the effort $e$ of the agent are complements.

Alternative Assumption 3. $\frac{\partial^2 B_2(\cdot, \cdot)}{\partial e \partial f} < 0$, i.e. the effort $f$ of the principal and the effort $e$ of the agent are substitutes.

All lemmata and propositions are valid if either Assumption 3 or Alternative Assumption 3 holds.

2.2. Analysis of the principal-agent relationship

I analyze this principal-agent relationship via backward induction. In the non-contractible part 2, the untrustworthy agent will never invest any effort, as he is not intrinsically interested in the project's success: $e_{lh} = 0$. The trustworthy agent works harder if he expects stronger efforts by the principal. In equilibrium the effort choices of the trustworthy agent, of the trusting principal, and of the distrusting principal are given by the following first order conditions,

$$\alpha_T \frac{\partial B_2(e_T, f_h)}{\partial e} + (1 - \alpha_T) \frac{\partial B_2(e_T, f_l)}{\partial e} = \frac{1}{\kappa},$$

$$\pi_h \frac{\partial B_2(e_T, f_h)}{\partial f} + (1 - \pi_h) \frac{\partial B_2(0, f_h)}{\partial f} = 1,$$

$$\pi_l \frac{\partial B_2(e_T, f_l)}{\partial f} + (1 - \pi_l) \frac{\partial B_2(0, f_l)}{\partial f} = 1,$$

where $\alpha_T \in [0, 1]$ denotes the trustworthy agent's belief that the principal received a high signal and trusts him. In general, $\alpha_T$ can differ from $\alpha_T$ as the principal's choice of contract may signal her type (i.e. her belief $\pi_i$, $i \in \{l, h\}$) to the agent. Assumption 2 guarantees a unique solution to this system of equations.

Lemma 1. Let Assumptions 1, 2, and either Assumption 3 or Alternative Assumption 3 hold. For any given belief $\alpha_T \in [0, 1]$ the continuation game in the non-contractible part 2 of the interaction has a unique equilibrium, characterized by Eqs. (4)–(6).

All proofs are in Appendix A. I denote this unique solution by $e^*_T(\alpha_T)$, $f^*_l(\alpha_T)$, and $f^*_h(\alpha_T)$. For a given $\alpha_T$ I denote the principal's expected net profit from part 2 by

$$v_1(\alpha_T) = \pi_l B_2(e^*_T(\alpha_T), f^*_l(\alpha_T)) + (1 - \pi_l) B_2(0, f^*_l(\alpha_T)) - f^*_l(\alpha_T).$$

where subindex $i \in \{l, h\}$ indicates whether the principal’s belief is low or high.

Lemma 2. Let Assumptions 1, 2, and either Assumption 3 or Alternative Assumption 3 hold. For any $\alpha_T \in [0, 1]$ the following holds: $v_h(\alpha_T) > v_l(\alpha_T)$ and

$$\frac{dv_h(\cdot)}{d\alpha_T} > \frac{dv_l(\cdot)}{d\alpha_T} > 0.$$
expects from the principal and the higher the effort level he will choose. Again, the principal wants to signal trust to make the agent work harder.

Lemma 2 states in addition that the trusting principal expects to profit more from an increase in \( \alpha_T \) than the distrusting principal. Intuitively, a higher \( \alpha_T \) increases the trustworthy agent’s effort and the trusting principal expects to profit more from such an increase because she assigns a higher probability to the trustworthy type of agent.

**Analysis of the contracting stage.** The agent observes whether the principal writes a contract. This influences the agent’s belief \( \alpha_T \) that 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 \( C_h \in \{c, n\} \) of a trusting principal and the contractual choice \( C_l \in \{c, n\} \) of a distrusting principal are both optimal given their beliefs \( \pi_i, i \in \{l, h\} \), and given the trustworthy agent’s optimal action derived from his beliefs \( \alpha_T \) and \( \alpha_T' \). On the other hand, the trustworthy agent’s beliefs \( \alpha_T' \) after observing the choice of a contract and \( \alpha_T'' \) after observing the choice of no-contract must be consistent given the equilibrium contractual choices \( C_l \) and \( C_h \).

Consider a potential equilibrium candidate which I denote, in a slight abuse of notation, by \((C_l, C_h, \alpha_T', \alpha_T'')\). Let \( V^n_i \) denote the expected payoff of a principal of type \( i \in \{l, h\} \) from the choice of writing no-contract, and \( V^c_i \) the expected payoff after writing a contract. Then,

\[
V^n_h = \pi_h B_1 + v_h(\alpha_T'),
V^n_l = \pi_l B_1 + v_l(\alpha_T'),
V^c_h = \pi_l B_1 + v_l(\alpha_T''),
V^c_l = \pi_l B_1 + v_l(\alpha_T'').
\]

The following relations are useful for the further analysis:

\[
V^n_h \geq V^n_l \iff v_h(\alpha_T') - v_h(\alpha_T') \geq (1 - \pi_h) B_1,
V^c_h \geq V^c_l \iff v_l(\alpha_T'') - v_l(\alpha_T'') \geq (1 - \pi_l) B_1.
\]

In equilibrium \( \alpha_T'' \geq \alpha_T' \) holds. Together with Lemma 2, inequality \( \alpha_T'' \geq \alpha_T' \) implies that \( v_h(\alpha_T') - v_h(\alpha_T') \geq v_l(\alpha_T'') - v_l(\alpha_T'') \geq 0 \), i.e. the trusting principal gains more from signaling trust. Furthermore, \( (1 - \pi_l) B_1 \geq (1 - \pi_h) B_1 \geq 0 \), i.e. the trusting principal expects to have smaller losses from writing no contract. Thus, the distrusting principal is less willing to choose a no-contract compared to the trusting principal.

Even a trusting principal only considers refraining from the option to enforce high effort in part 1 by means of a contract if the expected costs \((1 - \pi_h) B_1\) are below the maximal possible expected gains from such a signal \((v_h(1) - v_h(0))\). Otherwise, both types of principal would always write a contract. Thus, I concentrate on the interesting case:

**Condition 1.**

\[
B_1 < \frac{v_h(1) - v_h(0)}{1 - \pi_h}.
\]

**Refinement: Intuitive equilibria.** Signaling games suffer from a multiplicity of perfect Bayesian equilibria that are often sustained by pessimistic off-equilibrium beliefs. I focus on equilibria that are consistent with the intuitive criterion, an equilibrium refinement introduced by Cho and Kreps (1987). In the current context, a slightly more demanding definition of the intuitive criterion is needed because there are more than the two stages of the standard signaling game. I use a refinement from Maskin and Tirole (1992). It constrains the set of off-equilibrium beliefs by the following equilibrium domination argument. Consider an off-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 off-equilibrium belief), he assigns a higher probability to the type of agent.

**Definition 1.** An intuitive equilibrium is a perfect Bayesian equilibrium that is consistent with the intuitive criterion.

In Appendix A I specify and derive all perfect Bayesian equilibria of this game. Here, I concentrate on the main point and summarize the most relevant results in the following proposition.

---

13 The beliefs of the untrustworthy agent are payoff irrelevant. The remaining actions are the equilibrium responses given \((C_h, C_l, \alpha_T', \alpha_T'')\).

14 Proof: Suppose not. Then \( \alpha_T'' < \alpha_T' \). This implies firstly that in equilibrium the distrusting principal chooses \( n \) with positive probability. She is willing to do so only if \( V^n_i \geq V^c_i \). By equivalence (14) this implies \( v_l(\alpha_T'') - v_l(\alpha_T'') \geq (1 - \pi_l) B_1 \). Secondly, \( \alpha_T'' < \alpha_T' \) implies by Lemma 2 that \( v_l(\alpha_T'') - v_l(\alpha_T'') < 0 < (1 - \pi_l) B_1 \), a contradiction.
Proposition 1. Let Assumptions 1, 2, Condition 1, and either Assumption 3 or Alternative Assumption 3 hold.

(a) There exists an intuitive equilibrium in which at least the trusting principal writes no-contract.

(b) Suppose the following condition holds in addition:

$$\frac{v_l(1) - v_l(\sigma_T)}{1 - \pi_l} < B_1 < \frac{v_h(1) - v_h(\sigma_T)}{1 - \pi_h}. \tag{16}$$

Then, for a given value of $B_1$, each player type chooses a unique strategy in all intuitive equilibria. In particular, the trusting principal chooses to write no contract in any intuitive equilibrium.

Remark 1. These intuitive equilibria are

a) for $B_1 \geq \frac{v_l(1) - v_l(0)}{1 - \pi_l}$, the separating equilibrium in which the trusting principal chooses to write no contract and the distrusting principal chooses to write a contract;

b) for $\frac{v_l(\sigma_T) - v_l(0)}{1 - \pi_l} < B_1 < \frac{v_l(1) - v_l(0)}{1 - \pi_l}$, the hybrid equilibrium in which the trusting principal chooses to write no contract and the distrusting principal chooses to write no contract with probability $q = \frac{\sigma_T}{1 - \sigma_T} \frac{1 - \pi_l}{\pi_l}$;

c) for $B_1 \leq \frac{v_l(\sigma_T) - v_l(0)}{1 - \pi_l}$, the pooling equilibrium on writing no contract, in which both types of principal forgo writing a contract.

This proposition confirms 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 the condition (16) in Proposition 1(b) these are the only equilibria that are not excluded by the intuitive criterion. Then, there is a unique intuitive equilibrium outcome, in which the trusting principal always forgoes writing a contract.

Note that nothing in the setting prescribes ex ante that no contract should be the signal of trust and contract a signal of distrust – and yet signaling in the opposite way (i.e., trusting principal as a signal of trust) cannot be an equilibrium. The intuitive reason is that leaving a contract incomplete is less expensive for a principal who trusts the agent. Trust, by definition, implies that the principal considers it unlikely that the agent would exploit her if he can. Thus leaving him the possibility of shirking is less costly for the trusting principal than for the distrusting one.

Which equilibria exist depends on the potential costs of contractual incompleteness, which are captured by the parameter $B_1$. If these costs are extremely high no principal would be willing to pay these signaling costs. Lowering the potential costs of contractual incompleteness the trusting principal is willing to pay the signaling costs to separate herself from the distrusting type. A separating equilibrium results for a range of $B_1$ until the potential costs from contractual incompleteness are so low that the distrusting type starts imitating the trusting type with some probability in a hybrid equilibrium. The lower the potential costs from contractual incompleteness become, the higher the proportion of distrusting types who imitate the signal. For sufficiently small signaling costs pooling on no contract becomes the intuitive equilibrium.

Further perfect Bayesian equilibria can exist in addition to the equilibria described in Proposition 1 and Remark 1, e.g. the pooling equilibrium on writing a contract. Yet, for the intermediate range of $B_1$ defined by condition (16) of Proposition 1(b), these additional equilibria do not pass the intuitive criterion, and intuitive equilibrium actions are unique.

3. Discussion

This model demonstrates how the fear of signaling distrust can endogenously cause contractual incompleteness. The trusting principal, in particular, prefers 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, exists if the costs of contractual incompleteness are not too high. The range for uniqueness of the intuitive equilibrium is somewhat smaller.

The longer working paper version, Herold (2008), provides an extension where the principal can write more general wage schemes, the agent’s effort in part 1 of the relationship is continuous, and so the principal can choose the degree of contractual incompleteness. This complicates the analysis and requires additional assumptions. In return, the model provides a clear cut prediction about the equilibrium outcome: in any intuitive equilibrium, the distrustful principal chooses the complete contract, and the trusting principal chooses a contract with a least-cost separating degree of incompleteness. In other words, the trusting principal chooses a degree of contractual incompleteness just sufficient to separate herself from the distrustful type.15
In the main model, in which players have standard economic preferences, the value of signaling trust comes from an under-investment of the agent in a non-contractible part of the relationship when he believes he is distrusted. Contractual incentives and the associated signal of distrust can turn out to be counterproductive even without the non-contractible part of the relationship, if we are willing to take a more behavioral standpoint and consider behavior and preferences beyond the standard economic setting.

Firstly, a signal of mistrust may be perceived as a hostile act and lead the agent to reciprocate negatively (or to stop reciprocating positively). If the feeling of being mistrusted lowers an agent’s utility directly and if the agent attaches positive emotions to a relationship only when he feels trusted, then a signal of distrust destroys directly potential surplus from the relationship. The agent may decide not to participate in the relationship or only for a higher compensation. Perhaps even more important, the signal of trust can be seen as a hostile act, and the agent may stop to invest in the relationship or reciprocate negatively in other ways. Recent experimental findings by Falk and Kosfeld (2006) are consistent with such an interpretation.16

A second way how a signal of distrust can have detrimental effects for the relationship is given if trust has a mutual component: the agent finds it hard to trust someone who mistrusts him. In a relation of similar partners, this inference can be rational to some degree17 and may be amplified by the false consensus effect, i.e. a tendency to overestimate the probability that other people will act like oneself. Similarly, if the relationship will become eventually either a mutual friendship or a relation of mutual hostility (or at least of mutual indifference), then a signal of distrust makes the relation less likely to become a friendship. Thus the agent behaves less friendly.18 In fact, an experimental study by Malhotra and Murnighan (2002) on the effects of contracts on interpersonal trust suggests that a significant higher proportion of agents distrust principals who decided to implement a contract in previous rounds.19 These findings are consistent with such an interpretation.

Such behavioral considerations only reinforce the main argument. The relevance of these behavioral considerations is in the end an empirical question, which is left for future research. The focus of this paper is to derive the results under standard assumption on the agent’s preferences. The paper demonstrates that the fear of signaling 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. The 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 of the agent’s belief that he is trusted.

Acknowledgments

I thank the editor and the anonymous referees for helpful comments and suggestions. Furthermore I thank Roland Benabou, Dirk Bergemann, Peter Eso, Armin Falk, Ernst Fehr, Georg Gebhardt, Christoph Kuzmics, Klaus M. Schmidt, and participants at seminars at the Canadian Economic Theory Conference, Conference on Economics and Psychology at IDEI Toulouse, ESEM, NASM, Northwestern University, University of Munich, and TU-Dresden for their feedback to earlier drafts of this paper.

---

16 An agent can spend any amount from his endowment of 120 tokens 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 tokens, or whether she abstains from any control, in which case 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 not to control. An additional control treatment in which 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 suggests that a principal choosing to control the agent signals distrust, and that this crowds out the trustworthiness of the agent.

17 Consider the following stylized model to see the rationale for such an inference. 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 Engelmann (2000).

18 Recent studies suggest (see Sunnafrank and Ramirez, 2004) that the first impressions are decisive for the long-term nature of a relationship. Contract proposals are often the starting points in a relation.

19 More precisely, compare behavior under Conditions 1 and 2 in Experiment 1 in Malhotra and Murnighan (2002). Under Condition 1 participants decide whether to trust in a trust game. Under Condition 2 in the first two rounds a principal (actually a computer, but participants are told it is a person) decided to write a contract, which implemented the actions “trust” and the response “honor trust” automatically. Then in round three the agent is told that the principal was not any more allowed to write a contract. The percentage of trusting agents drops significantly from 86.4% under Condition 1 to 60.7% under Condition 2. Note also that the mere mentioning of contracts, without the option to implement one, drops the percentage of trusting agents to 69.6% (or 78.3% if you consider Round 1 instead of Round 3). The drop is not significant (the number of participants is only 22–28 for each condition) but if such a drop could be confirmed with more data this would indicate that merely mentioning contracts already creates an atmosphere of distrust.
Appendix A. Definitions and proofs

A.1. Definition of the intuitive criterion

The original intuitive criterion by Cho and Kreps (1987) was defined in a context of two types and two stages. Here the sender (the principal) has also two types. Yet, we have more than two stages. So we need a slightly more demanding refinement (compare also Maskin and Tirole, 1992). Let \( T = \{ l, h \} \) denote the set of the two types of principals. Let \( BR(C, \alpha_T) \) denote the (unique) equilibrium of the continuation game between the principal and the agent after the principal selected \( C \in \{ n, c \} \) and the trustworthy agent updated 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 \( i \). I denote an off-equilibrium contract proposal \( \tilde{C} \) as equilibrium dominated for type \( i \) and not equilibrium dominated for the other type.

A perfect Bayesian equilibrium passes the intuitive criterion if and only if the off-equilibrium beliefs \( \alpha_T(\tilde{C}) \) assign zero probability to type \( i \) (i.e. \( \alpha_T(\tilde{C}) = 0 \) if \( i = h \) and \( \alpha_T(\tilde{C}) = 1 \) if \( i = l \)) whenever \( \tilde{C} \) is equilibrium dominated for type \( i \) and not equilibrium dominated for the other type \( j \).

Proof of Lemma 1. Lemma 1 is proven by defining a supermodular three player game of complete information which corresponds to the Bayesian game in part 2 of the relationship. The existence of a unique solution can then be proven by arguments used in Milgrom and Roberts (1990), Example 2. Define

\[
\begin{align*}
&f_1(x_1, x_2, x_3) = \kappa (\alpha_T B_2(x_1, x_2) + (1 - \alpha_T) B_2(x_1, x_3)) - x_1, \\
&f_2(x_1, x_2, x_3) = \pi_h B_2(x_1, x_2) + (1 - \pi_h) B_2(0, x_2) - x_2, \\
&f_3(x_1, x_2, x_3) = \pi_l B_2(x_1, x_3) + (1 - \pi_l) B_2(0, x_3) - x_3.
\end{align*}
\]

Under Assumption 3 (complements) these define a three player supermodular game. Assumption 2 guarantees that the dominant diagonal condition \(-\frac{\partial^2 f_n}{\partial x_n \partial x_k} > \sum_{i \neq n} \frac{\partial^2 f_n}{\partial x_n \partial x_k}\) holds for \( n \in \{1, 2, 3\} \) and the proof of uniqueness of the Nash equilibrium by Milgrom and Roberts (1990), Example 2 applies to our setting: since the game is supermodular there exist a smallest equilibrium \((x_1, x_2, x_3)\) and a largest equilibrium \((\hat{x}_1, \hat{x}_2, \hat{x}_3)\), with \( x_i \leq \hat{x}_i \) for all \( i \in \{1, 2, 3\} \). Proof by contradiction: Suppose \( x < \hat{x} \). Let \( n \) be the player for whom \( \hat{x}_i - x_i \) is largest. Then,

\[
\frac{\partial f_n}{\partial x_n}(\hat{x}) - \frac{\partial f_n}{\partial x_n}(x) = \int_0^1 \left\{ \sum_{i \in \{1,2,3\}} (\hat{x}_j - x_j) \frac{\partial^2 f_n}{\partial x_n \partial x_i} \left[ t\hat{x} + (1 - t)x \right] \right\} dt.
\]

The assumption that \( x \) and \( \hat{x} \) are equilibria with \( \hat{x}_n > x_n \) requires that \( \frac{\partial f_n}{\partial x_n} \) be nonnegative at \( \hat{x} \) and nonpositive at \( x \). Thus the left-hand side of (18) is positive. Yet, the integrant in (18) is strictly negative according to the dominant diagonal condition, a contradiction.

Under Alternative Assumption 3 (substitutes) the change of variable \( \hat{x}_1 = -x_1 \) makes sure that \( f_1, f_2, \) and \( f_3 \) again define a supermodular game, and the same arguments apply. \( \square \)

Proof of Lemma 2. \( v_h(\alpha_T) > v_l(\alpha_T) \) follows directly from \( \pi_h > \pi_l \). The proof of part 2 uses

\[
\begin{align*}
v_i(\alpha_T) = \tilde{v}_i(e^*_T(\alpha_T), f^*_i(\alpha_T)) = \max_{f_i} \tilde{v}_i(e^*_T(\alpha_T), f_i(\alpha_T)),
\end{align*}
\]

for \( i \in \{l, h\} \) and where \( \tilde{v}_i(\cdot, \cdot) \) is defined as \( \tilde{v}_i(e, f) = \pi_l B_2(e, f) + (1 - \pi_l) B_2(0, f) - f \). Hence, we have \( \frac{\partial \tilde{v}_i(e^*_T(\alpha_T), f^*_i(\alpha_T))}{\partial e} = 0 \) and therefore

\[
\begin{align*}
\frac{d v_i(\alpha_T)}{d \alpha_T} = \frac{d \tilde{v}_i(e^*_T(\alpha_T), f^*_i(\alpha_T))}{d \alpha_T} = \pi_l \frac{\partial B_2(e^*_T(\alpha_T), f^*_i(\alpha_T))}{\partial e} \frac{\partial e^*_T(\alpha_T)}{\partial \alpha_T} > 0,
\end{align*}
\]

\( \square \)

Notice that the equilibrium strategies are independent of the belief \( \alpha_n \) of the untrustworthy agent.

Let \( \mathbb{B}_2 \) be the supremum of \( B_2(\cdot, \cdot) \). Then all strategies of \( x_i \geq \mathbb{B}_2 \) are strictly dominated. I can therefore restrict the analysis to \( x_i \in [0, \mathbb{B}_2] \).
since all factors are positive. Finally, \( \frac{\partial b_2(e_2(\alpha_T), f_2^D(\alpha_T))}{\partial e} > \frac{\partial b_2(e_2(\alpha_T), f_2^D(\alpha_T))}{\partial e} \) follows from Eq. (20) together with the fact that \( \pi_h > \pi_l \) and \( \frac{\partial b_2(e_2(\alpha_T), f_2^D(\alpha_T))}{\partial e} \). The last inequality follows in the case of complements (under Assumption 3) because both types could potentially profit from a deviation. Thus, these perfect Bayesian equilibria are also intuitive equilibria.

**Proof of Proposition 1.** I go through all perfect Bayesian equilibria candidates and check whether they pass the intuitive criterion. Notice that the beliefs \( \alpha_{il} \) of the untrustworthy agent are irrelevant for behavior and payoffs.

**Separating equilibrium.** In any separating equilibrium the trusting principal writes no contract and the distrusting principal writes a contract. (The reverse situation cannot be an equilibrium since the trusting type would have a strict incentive to imitate the trusting type.) In such an equilibrium the agent’s beliefs have to be \( \alpha_T = 1 \) and \( \alpha_T = 0 \). It is then optimal for the trusting principal to offer the incomplete contract if and only if \( v_h(1) - v_h(0) \geq (1 - \pi_h)B_1 \). This coincides with Condition 1. For the distrusting principal, it is optimal to offer the complete contract if and only if \( v_l(1) - v_l(0) \leq (1 - \pi_l)B_1 \). The separating equilibrium passes always the intuitive criterion, since there are no off-equilibrium beliefs. Hence, under Condition 1 there exists a separating intuitive equilibrium if and only if \( B_1 \geq \frac{v_l(1) - v_l(0)}{1 - \pi_l} \).

**Pooling on writing no contract.** If both types of principal pool on no contract then the agent’s belief on the equilibrium path has to be \( \alpha_T = \sigma_T \). The principal’s incentive to deviate is smallest, if the off-equilibrium belief is \( \alpha_T = 0 \). Furthermore, the trusting principal always has a smaller incentive to deviate. Thus, there exist perfect Bayesian pooling equilibria on writing no-contract if and only if \( v_l(\sigma_T) - v_l(0) \geq (1 - \pi_l)B_1 \), or equivalently \( B_1 \leq \frac{v_l(\sigma_T) - v_l(0)}{1 - \pi_l} \). The intuitive criterion has no bite since both types could potentially profit from a deviation. Thus, these perfect Bayesian equilibria are also intuitive equilibria.

**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 \( \alpha_T = \sigma_T \) and the equilibrium is sustained by an off-equilibrium belief \( \alpha_T = \sigma_T - v_h(\sigma_T) + (1 - \pi_h)B_1 \). The intuitive criterion requires that \( \alpha_T = 1 \) if the deviation of writing no contract is potentially profitable for a trusting principal yet equilibrium dominated for an untrustworthy principal. This is the case for \( (\frac{v_l(1) - v_l(0)}{1 - \pi_l} - B_1 < v_h(\sigma_T)) \). Under this condition, there exists no intuitive pooling equilibrium on writing a contract, because with \( \alpha_T = 1 \) the trusting type would want to deviate.

**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 randomizes between both contractual choices, if and only if \( v_l(\sigma_T) < v_l(0) + (1 - \pi_l)B_1 < v_l(1) \). This is an intuitive equilibrium. The trustworthy agent’s beliefs are \( \alpha_T = 0 \) and \( \alpha_T = v_l^{-1}(v_l(0) + (1 - \pi_l)B_1) \) in such an equilibrium. The distrusting principal chooses no-contract with probability \( q = \frac{\pi_l}{1 - \pi_l} \).

I now derive this result. In such a hybrid equilibrium the agent understands that he is facing a distrusting principal if he observes a contract. Hence, \( \alpha_T = 0 \). If he observes no contract he updates his beliefs by Bayes-rule \( \alpha_T = \sigma_T - (1 - \pi_l)B_1 \). Notice that \( \sigma_T < \alpha_T < 1 \) for every \( q \in (0, 1) \). Vice versa, for any \( \alpha_T \) with \( \sigma_T < \alpha_T < 1 \) there exists a \( q \in (0, 1) \) that leads to this \( \alpha_T \), namely \( q = \frac{\sigma_T}{1 - \sigma_T} \). These beliefs and contractual choices form a perfect Bayesian equilibrium if and only if

\[
\{C_5\} \quad v_h(\alpha_T) - v_h(0) \geq (1 - \pi_h)B_1, \tag{21}
\]

\[
\{C_6\} \quad v_l(\alpha_T) - v_l(0) = (1 - \pi_l)B_1. \tag{22}
\]

(9C) implies (IC) by Lemma 2. So there exists a perfect hybrid Bayesian equilibrium in which the distrusting principal randomizes, if and only if there exists a \( q \in (0, 1) \) such that \( v_l(\alpha_T) = v_l(0) + (1 - \pi_l)B_1 \). This is equivalent to demanding that there exists an \( \alpha_T \) with \( \sigma_T < \alpha_T < 1 \) such that \( v_l(\alpha_T) = v_l(0) + (1 - \pi_l)B_1 \) (see above). Since \( v_l(\cdot) \) is continuous and strictly increasing, this is the case (see above) if and only if

\[
v_l(\sigma_T) < v_l(0) + (1 - \pi_l)B_1 < v_l(1). \tag{23}
\]

Here, a perfect Bayesian hybrid equilibrium is also an intuitive equilibrium since there are no off-equilibrium beliefs.

**Hybrid equilibria with random contract choice by 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 randomizes between “contract” and “no contract” if and only if \( B_1 > \frac{v_h(1) - v_h(\sigma_T)}{1 - \pi_h} \). This is also an intuitive equilibrium. In this equilibrium the trustworthy agent’s beliefs are \( \alpha_T = \sigma_T \) and \( \alpha_T = 1 \). The trusting principal writes no-contract with probability \( q' = \frac{1}{1 - \sigma_T} \).
To derive this, note that in this equilibrium the agent knows, when observing "no contract," that the principal is of the trusting type, i.e., $\alpha^T_T = 1$. In case he observes a contract he updates his beliefs by Bayes-rule: $\alpha^C_T = \frac{(1-q')\sigma_T}{1-(1-q')\sigma_T}$. Notice that $0 < \alpha^C_T < \sigma_T$ for every $q' \in (0, 1)$. In reverse, for any $0 < \alpha^C_T < \sigma_T$ there is a $q' \in (0, 1)$ that leads to belief $\alpha^C_T$, namely $q' = 1 - \frac{1-\sigma_T}{1-\alpha^C_T}$.

These beliefs and contractual choices form an equilibrium if and only if

$$
(\text{IC}_h) \quad v_h(1) - v_h(\alpha^C_T) = (1 - \pi_h)\bar{B}_1.
$$

$$
(\text{IC}_c) \quad v_l(1) - v_l(\alpha^C_T) \leq (1 - \pi_l)\bar{B}_1.
$$

$(\text{IC}_h)$ implies already $(\text{IC}_l)$ by Lemma 2. Thus, there exists a perfect Bayesian hybrid equilibrium in which the trusting principal randomizes if and only if there exists a $q' \in (0, 1)$ such that $v_h(\alpha^C_T) = v_h(1) - (1 - \pi_h)\bar{B}_1$. Equivalently, we can demand that there is an $\alpha^C_T \in (0, \sigma_T)$ with $v_h(\alpha^C_T) = v_h(1) - (1 - \pi_h)\bar{B}_1$. By continuity and strict monotonicity of $v_h(\cdot)$ this is the case if and only if

$$
v_h(0) < v_h(1) - (1 - \pi_h)\bar{B}_1 < v_h(\sigma_T).
$$

The first inequality is guaranteed already by Condition 1. The second part of (26) violates the condition in Proposition 1(b), so under this condition, such an equilibrium cannot exist. Again, a hybrid perfect Bayesian equilibrium is also an intuitive equilibrium since there are no off-equilibrium beliefs.

The proof of Proposition 1(a) follows directly from these derivations and Remark 1. Proposition 1(b) follows from the restrictions derived in the paragraph on the pooling equilibrium on contract and the hybrid equilibrium in which the trusting principal plays a mixed action. □

References

Friebe, G., Schneider, W., 2007. Team governance: Empowerment or hierarchical control. IZA working paper.
Holmstrom, B., Milgrom, P.R., 1992. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. J. Law, Econ., Organ. 7 (Special Issue), 24–52.
Spier, K.E., 1992. Incomplete contracts and signalling. RAND J. Econ. 23 (3), 432–443.