Credible Threats, Reputation and Private Monitoring.*

Olivier Compte†

First Version: June 2001
This Version: November 2003

Abstract

In principal-agent relationships, a termination threat is often thought to be an effective way to provide the agent with incentives to sustain high effort. We examine the credibility of termination threats in a setting where the principal only receives private signals concerning the other party's (i.e. the agent's) effort level. We also assume that with small probability, the principal is weak: termination is very costly for him. We show that the principal cannot avoid building a reputation from being a weak type, and as a result, there can be no equilibrium where the threat to terminate is credible, hence no equilibrium where the agent makes high effort.

1 Introduction

Termination is a stick that can be used in principal-agent relationships to discipline the agent. By terminating the relationship when the agent makes

*I thank seminar participants at the Cowles foundation conference on Repeated games with private monitoring, at which the results of this paper were first presented.

†CERAS-ENPC, e-mail compte@enpc.fr.

little effort, a principal can provide the agent with incentives to make high effort. Often though, effort cannot be observed directly, and only a signal (imperfectly) correlated with effort can be observed. In some cases, the signal can be the quantity of output, but sometimes the principal cares about dimensions that are more subjective or more difficult to quantify. In the first case, the signal is public, because typically, the agent knows the quantity of output produced. In the second case, the signal is private, because typically the agent does not know for sure how the quality of his work was perceived.

Still, whether signals are public or private, termination threats would seem to be effective because by conditioning termination on the signal observed, the principal can make termination more likely when effort is not made than when effort is made.

One important issue has to be addressed though. Is the threat credible? That is, does the principal have incentives to carry out termination when the circumstances under which he had planned to terminate are met? Alternatively, does the agent believes that the principal will indeed carry out termination when supposed to? We answer these questions in a setting where at the start of the relationship, the agent has some doubt (possibly very small) about the principal's willingness to carry out the threat. We find that whether the termination threat is credible or effective crucially depends on the nature of the signals. With public signals and so long as the agent does not face direct evidence that the principal is not willing to carry out the threat, the agent's initial doubt will persist but remain at the same very small level, and thus affecting only slightly the agent's incentive to make high effort. With private signals, the dynamic of beliefs is quite different: as we will show, the agent's initial doubt will necessarily inflate over time, and eventually make the threat non credible. Anticipating this, the relationship must breakdown immediately: the termination threat is completely ineffective.

Our model is a simple principal agent game where the party that threatens to terminate the relationship (i.e. the principal) only receives private signals concerning the other party's (i.e. the agent's) effort level. The principal decides when to terminate the relationship, as a function of the signals he receives; the agent decides when to suspend effort, as a function of the signals he receives. We also assume that with small probability, the principal is weak: termination is very costly for him. We show that the principal cannot help building a reputation from being a weak type, and as a result, there can be no equilibrium where the threat to terminate is credible, hence no equilibrium where the agent makes high effort.

Private monitoring plays a key role in our model. If the signals received by the principal were public, the agent would know (in equilibrium) the circumstances under which the principal is supposed to terminate the relationship; and when such circumstances arise, the principal would know that if he does not terminate, he will be thought to be a weak type, and no further effort will be made by the agent. Hence the principal, if not weak, truly has incentives to terminate the relationship, which thus implies that the threat of termination is credible. It also implies that as long as the circumstances above are not met, the principal's reputation does not change: he continues to be thought of as a strong type with high probability.

Under private monitoring, the logic above breaks down, because the agent never knows if and when the circumstances under which termination should be triggered are met. He only knows that they are likely to be met at some point, and seeing that the principal does not terminate then constitute evidence that the principal is likely to be weak. Over time, the principal cannot help building a reputation for being a weak type. As we shall see, this phenomenon is quite robust, and arises even if signals are almost public.

¿From a more general perspective, this paper opens interesting questions concerning reputation building in long term relationship. Our example is an instance in which a player would like to build *a reputation for using a* *particular decision rule* (say terminate if T consecutive bad signals arise). Our result illustrates that as soon as monitoring is not public, it is hard to build up such a reputation, because other players cannot observe the information on which the decision is based, hence they cannot observe or check whether the decision rule is being followed on not. And as a consequence, agents may have a hard time building the reputation they would like.

This paper also gives some insight on the effectiveness of a very common form of threat: the ultimatum, where one of the parties threatens to take dramatic action unless certain conditions are met by the other party. When the conditions under which the ultimatum should be triggered are not public, and when agents have some doubt about the costs of carrying out the ultimatum, repeated use of this threat will contribute to exacerbate these doubts, making the threat ineffective.

The paper is organized as follows. Section 2 presents the basic model. In Section 3, we present the main result. We then consider the case where signals are public, and the case where signals are almost public. Finally, the results and the related literature are discussed in Section 4.

2 The Model

We consider a simple principal agent relationship where (i) the principal decides upon when to terminate the relationship, and where (ii) the agent decides upon a date to suspend effort.

Formally, at any date t, the principal decides whether to continue or terminate the relationship. Then the agent decides whether to continue to make an effort one more period, or to suspend it for ever. So each player has one irreversible action that he may take: termination for the principal, suspension of effort for the agent.[1]

[1]The fact that suspension of effort is irreversible is not essential to our result, but it greatly simplifies the argument.

Payoffs: In case of termination, we assume that the principal and the agent each obtain a continuation payoff normalized to 0. At any date t where the principal decides to continue the relationship, the parties' expected payoff depend on whether the agents exert effort. Denote by e the agent's effort: $e \in \{N, E\}$. Expected payoffs for the agent and the principal are gathered in the following bi-matrix:

$$\begin{array}{cc} E & N \\ 1,1 & -\ell, 1+g \end{array}$$

where ℓ and g are positive scalars.

That is, both players benefit from the relationship if the agent makes an effort. Otherwise only the agent benefits from the relationship. We assume that players discount future payoffs with a common discount factor $\delta < 1$, and maximize their average discounted expected payoff. Formally, let τ_1 denotes the date at which the principal terminates the relationship, and let τ_2 denote the date at which the agent suspends effort. The average discounted expected payoff of player i associated with the pair (τ_1, τ_2) is denoted $u_i(\tau_1, \tau_2)$. Let $\tau = \min\{\tau_1, \tau_2\}$. During the first $\tau - 1$ periods, both parties get an expected payoff equal to 1. Then during $\tau_1 - \tau$ periods, the principal gets $-\ell$ while the agent gets $(1 + g)$. So we have:

$$\begin{aligned} u_1(\tau_1, \tau_2) &= (1 - \delta^\tau) - (\delta^\tau - \delta^{\tau_1})\ell \\ u_2(\tau_1, \tau_2) &= (1 - \delta^\tau) + (\delta^\tau - \delta^{\tau_1})(1 + g) \end{aligned}$$

Information: We assume that at any date where the relationship continues, the principal receives a signal $y \in Y$ imperfectly correlated with the agent's effort $e \in \{N, E\}$. We assume that $Y = \{\underline{y}, \bar{y}\}$ and let

$$p = \Pr\{\bar{y} \mid E\} \text{ and } q = \Pr\{\bar{y} \mid N\}$$

We assume that

$$1 > p > q > 0$$

That is, the signal $\bar{y}$ is good news for the principal: this signal is more likely when the agent makes an effort than when he does not. In the basic version of our model, we assume that the agent receives no signal.[2]

Strategies: A strategy for player 2 is a (random) date at which he chooses to suspend effort. A strategy for player 1 specifies at each date t whether to terminate the relationship, as a function of the sequence of signals $h_1^t = (y^1, ..., y^{t-1})$ she received so far.

Finally, with small probability $\nu > 0$, the principal will never terminate the relationship (for example because with small probability $l < 0$). We are interested in the behavior of the principal who may find it optimal to terminate the relationship. We look the perfect Bayesian equilibria of the game.

3 Main result

We show the following:

Proposition 1 *For any $\nu > 0$, the game has a unique perfect Bayesian equilibrium, in which the principal terminates the relationship right away.*

Proposition 1 implies that even if the probability ν that the principal is weak is very small, the principal that is not weak does not start the relationship.

Consider a perfect Bayesian equilibrium $\sigma = (\sigma^P, \sigma^A)$, and assume (by contradiction) that the (rational) principal does not terminate right away.

We first observe that there cannot exist a date t at which the rational principal would terminate with probability one. Because otherwise, the agent would make no effort at $t-1$, hence the rational principal would want

[2]We will later examine cases where the agent receives also receives a signal. In one case, we will assume that y is public. In another case, we will assume that y is almost public, that is, the agent receives a signal almost perfectly correlated with y.

to terminate at $t-1$, and so on backward. Similarly, there cannot exists a date at which the agent makes no further effort with probability 1.

Now consider a date t at which the agent has not suspended effort yet. We denote by $y^{t,T}$ a sequence of signals $(y^t, ..., y^{t+T})$ received by the principal during $\{t, ..., t+T\}$. We also denote by $\kappa^{t,T}$ the probability (induced by σ^P) that the principal terminates during $\{t, ..., t+T\}$ given that he has not terminated before t, when the agent makes an effort until $t+T$ at least; and by $\kappa^{t,T}_{susp}$ that same probability when the agent instead chooses to suspend effort at t. For convenience, for any event z, $\kappa^{t,T}(z)$ will denote the probability of termination conditional on event z. We also let $\xi = \min(p, 1-p)$.

Step 1: We first derive a lower bound on the agent's continuation gain, as viewed from date t, if he suspends effort at t.

Whether the agent suspends effort at t or continues making an effort until $t+T$ has an effect on the distribution over the sequences of signals $y^{t,T}$ received by the principal, hence over the expected probability that the principal terminates during $\{t, ..., t+T\}$. To provide the agent with incentives to prolong effort, one would hope that the probability of termination is large when the agent suspends effort (i.e. $\kappa^{t,T}_{susp}$ large); and to ensure that the relationship will not be too short, one would also hope that the probability of termination is small when the agent does not suspend effort (i.e. $\kappa^{t,T}$ small). We observe below that if $\kappa^{t,T}$ is very small, then $\kappa^{t,T}_{susp}$ cannot be large.

More precisely, we have:

$$\kappa^{t,T}_{susp} \leq \max_{y^{t,T}} \kappa^{t,T}(y^{t,T}) \leq \frac{\kappa^{t,T}}{\xi^T}.$$

Since, in the event where termination does not occur, the agent obtains a payoff equal to $1+g$ until date $t+T$ at least, we obtain a lower bound on the expected payoff obtained by the agent when he suspends effort at t:

$$G_t^{agent}(\sigma^{E,t}, \sigma^P) > (1 - \frac{\kappa^{t,T}}{\xi^T})(1 - \delta^T)(1 + g).$$

where the index in G_t^{agent} indicates that payoffs are viewed from date t.

Step 2: We now establish a lower bound on $\kappa^{t,T}$. Intuitively, if $\kappa^{t,T}$ is very small, then suspending effort is a very attractive strategy, so attractive that not suspending effort may become a dominated strategy, contradiction our initial observation that the agent should sustain effort with positive probability at every date.

Formally, we fix α^* small, and choose T large enough so that

$$(1 - \delta^T)(1 + g) > (1 - \alpha^*)(1 + g) + \alpha^*$$

Having fixed T, we now choose κ^*so that

$$(1 - \frac{\kappa^*}{\xi^T})(1 - \delta^T)(1 + g) > (1 - \alpha^*)(1 + g) + \alpha^*.$$

The right hand side is an upper bound on the gain that the agent can obtain if he sustain effort until at least date $t + \alpha^* T$. By step one, it follows that if $\kappa^{t,T} < \kappa^*$, then the agent would prefer to suspend effort at t rather than wait until date $t + \alpha^* T$ or a later date to do so. This would imply that the agent terminates before $t + \alpha^* T$ with probability 1, contradicting our initial observation. It therefore follows that

$$\kappa^{t,T} \geq \kappa^*.$$

Final step: We may now apply an idea which, since Fudenberg Levine (1989), is common in the reputation literature. Since, in any given lapse of time of length T, the rational principal is supposed to terminate the relationship with positive probability κ^* at least, on any path where he does not terminate, he must acquire a reputation for being a weak type in a bounded number of periods, say $\bar{T}$. This implies that in the equilibrium considered, the agent suspends effort with probability one at or before $\bar{T}$, contradicting our initial observation.

The case of public signals. In order to better understand to the role played by private monitoring in our model, we now examine the case where signals are also observed by the agent, and show that the principal may both (i) induce high effort and (ii) avoid getting a reputation for being a weak type.

Proposition 2 *Assume that signals are also observed by the agent. Then if the discount factor δ is not too small, and if signals are sufficiently informative, there exists an equilibrium that yields a positive payoff to both the principal and the agent.*

The proof is standard. Consider the strategy profile σ^T defined as follows: for the principal, terminate the relationship as soon as T consecutive bad signals are observed; for the agent, suspend effort as soon as T consecutive signals are observed.

If the principal does not terminate after T consecutive bad signals, he is believed to be weak with probability one, and it is thus optimal for the agent to suspend effort. And given that the agent suspends effort, it is also optimal for the principal to terminate the relationship.

We just have to check that for an appropriate choice of T the agent has incentives to make an effort after any other realizations of the signals. Intuitively, for T not too large, suspending effort is likely to induce a sequence of T consecutive bad signals in a not too distant future, and the agent therefore has incentive to continue making an effort. Computations are standard and relegated to the Appendix.

The case of almost public signals. In Proposition 1 and 2, we have examined two extreme cases. One in which the agent receives a signal perfectly correlated with that of the principal (Proposition 2), and one in which the agent receives no signal. What outcome should we expect in the intermediate case where the agent receives a signal that is imperfectly correlated with that of the Principal ? We will show below that even if signals are almost perfectly correlated, it is the logic of Proposition 1 that applies, and effort by the agent cannot be supported in equilibrium.

Formally, we shall consider the case where at each date, the agent receives a signal $y_a \in \{\underline{y}, \bar{y}\}$ imperfectly correlated with y. Specifically, we assume that

$$\Pr\{y_a = \bar{y} \mid y = \bar{y}\} = 1 - \varepsilon = \Pr\{y_a = \underline{y} \mid y = \underline{y}\},$$

where ε is a small positive scalar. Note that compared to the set up analyzed in earlier, a strategy for the agent now specifies a (possibly stochastic) date of suspension as a function of the history of signals received. We have the following Proposition:

Proposition 3 *For any $\epsilon > 0$, $\nu > 0$, the game has a unique perfect Bayesian equilibrium, in which the principal (if not weak) terminates the relationship right away.*

The intuition is very similar to that of Proposition 1. The trick is to consider a history of signals received by the agent for which, at any date, the agent continues to make an effort with positive probability. Such a history of signals necessarily exists because otherwise, the agent would suspend effort with probability 1 before some fixed date, and thus the principal would necessarily terminate with probability 1 before that date. The proof consists in showing that along that history of signals (and as long as termination does not occur), the principal cannot help building a reputation for being a weak type.

Formally, let h_a be that history of signal, and h_a^t the truncation of that history on the t first dates. By an argument analogous to Step 1 and 2 above, one can find a lower bound κ^* on the expected probability of termination $\kappa^{t,T}(h_a^t)$. In contrast to the case analyzed before, one cannot conclude directly that upon seeing no termination at date $t+T$, the agent will update his belief on the principal being weak by a factor larger than 1, because it could well be that $\kappa^{t,T}(h_a^t)$ is positive, and yet $\kappa^{t,T}(h_a^t, y_a^{t,T}) = 0$. However, for any $y_a^{t,T}$,

$$\kappa^{t,T}(h_a^t, y_a^{t,T}) \geq \varepsilon^T \max \kappa^{t,T}(h_a^t, y^{t,T}) \geq \varepsilon^T \kappa^{t,T}(h_a^t),$$

so a lower bound on $\kappa^{t,T}(h_a^t)$ implies a lower bound $\kappa^{t,T}(h_a^t, y_a^{t,T})$, hence an increase in the principal's reputation for being a weak type by a factor larger than 1.

4 Discussion and Related Literature.

In this discussion, we start by explaining the logic of our result, address some robustness issues, and relate our result to recent work on reputation. We then move on to explaining the connection of our work with the literature on repeated games with private monitoring.

Our model has two critical ingredients: One concerns the presence of strategic uncertainty at the start of the relationship, the other concerns the information structure (there is partial private monitoring). Although our model is quite structured and puts strong constraints on the set of strategies available, we believe the logic of our analysis and the way the two ingredients above combine to generate novel insights extend to more general settings, as we explain below.

The principal's strategy can be viewed as a behavior rule that maps signals to decisions. The first ingredient implies that the agent will try to make inferences about which rule the principal is using. The second ingredient implies that the agent will try to make inferences concerning the principal's behavior rule based on limited information: the agent observes the decisions made by the principal, but he cannot observe the signals on which decisions are based.

Although the agent's information is limited, she can compare the realized sequence of decisions to the average decision she can expect from each possible behavior rule that the principal may be following, and thereby make some inferences about the type of principal she is facing. The problem is that given a particular behavior rule followed by the principal, for some realizations of the signals, the realized sequence of decisions taken by the principal will look like the average decision expected from a quite different type of principal, and the agent will make incorrect inferences. As a consequence, although the principal might wish to convince the agent that he follows a particular decision rule, for some (unlikely) sequence of signals, following this rule will actually convince the agent that he is most likely following another one.

To be more specific, if the rational principal was supposed to terminate the relationship after T consecutive bad signals, then the agent would expect the rational principal to terminate the relationship with probability close to 1 after a sufficiently long lapse of time. It could well be however that T consecutive bad signals do not occur even after such a very long lapse of time. In this case, and even if the principal did indeed conform to the proposed decision rule, the most reasonable inference that the agent will draw is that he is facing a weak type.

Admittedly, our model put a lot of constraints on what the principal can do: if we allowed the principal to temporarily terminate the relationship, he would have an easy to distinguish himself from a weak type that never terminates. Complicating the model however would not alleviate the problem that we have identified, because types could be more complex than the one we considered.

To see why, consider a model where the principal may only trigger temporary termination (of fixed duration), and consider two types of principal. The first one triggers (temporary) termination after T consecutive bad signals, with T not too large. The second one triggers (temporary) termination after $\bar{T}$ consecutive bad signals, with $\bar{T}$ very large. Even when facing the first type of principal, it is quite possible that termination is not triggered for a very long time (because the event T consecutive bad signals did not occur). Under such an event, the agent will get convinced that it is more likely that he is facing the second type of principal rather than the first type of principal. Hence, mimicking the first type of principal is no guarantee that he will acquire a reputation for being that first type.

A similar phenomenon is at work in the political correctness example of Morris (2001) and in the bad reputation example of Ely Valimaki (2003).[3] In Morris, an adviser reports information to a decision-maker. There is a small chance that the adviser is biased (racist), and the adviser would like to avoid getting a reputation for being biased. Honest reporting of information is no guarantee that the adviser will avoid getting a reputation for being biased. The rational response is then to distort reporting (thereby reducing the informational value of the message). In Ely Valimaki, a mechanics would like to avoid getting a reputation for being bad (i.e. always performing an engine repair). However, even if the mechanics is following an honest behavioral rule (i.e. performing a tune up or an engine repair depending on which is in the best interest of the agent), mimicking the honest mechanics does not guarantee that he will acquire a reputation for honesty. In the event where the engine repair is required for many periods, a mechanics behaving honestly will end up getting a reputation for being a dishonest mechanics that always perform the engine repair.[4]

[3]For a generalization of Ely Valimaki (2003), see Ely Fudenberg and Levine (2003).

[4]What Ely and Valimaki show is that even though the mechanics may avoid getting a bad reputation by performing a tune up, the agent, anticipating that the mechanic will

This paper is also related to the literature on repeated game with private monitoring. One issue addressed in this literature has been whether private monitoring could perform as well as public monitoring in providing agents with incentives to say, cooperate, or sustain effort. Our model is a clear-cut example in which public monitoring dominates private monitoring. Our key departure from the literature is the adjunction of reputation concerns. Of course, private monitoring alone may make the provision of incentives more difficult, even when there are no reputation concerns, as we have shown in earlier work.[5] Nevertheless positive results can be obtained even when monitoring is private.[6] Our work suggests that throwing in reputation concerns in otherwise standard repeated games with private monitoring is likely to harden the task of finding positive results. First because, as in the simple model we proposed, it may be difficult for players to avoid acquiring a reputation they do not want. Second, because once reputation concerns are thrown in, beliefs typically depend on the whole history of play and signals, hence the type of construction used in the work by Piccione (2002) or Ely Valimaki (2002) cannot be performed; nor does the construction of Mailath Morris (2002), which relies on the use of finite memory strategies.

References

[1] Compte, (2002a): "On failing to cooperate when Monitoring is Private", *Journal of Economic Theory*, **102**, 151-188,

do a tune up (even if the engine repair is required) will not want to bring the car.

[5] See Compte (2002a).

[6] In our model, and in the absence of reputation concerns, positive results could be obtained if reputation concerns were not present (as in Compte (2002b) which analyzes a model with a similar structure). Obtaining positive result would be even simpler in a slightly enriched version of our model, where the (rational) principal would have the option to replace at no cost the current agent.

[2] Compte (2002b): "On sustaining cooperation without Public Observations", *Journal of Economic Theory*, **102**, 106-150

[3] Ely, J. and Valimaki, J. (2002) : "A Robust Folk-Theorem for the Repeated Prisoner's Dilemma" *Journal of Economic Theory*, **102**, 84-105.

[4] Ely, J., Fudenberg, D and Levine, D (2003): "When is Reputation Bad?" mimeo

[5] Ely, J. and Valimaki, J. (2003) : "Bad reputation", *Quarterly Journal of Economics*, Forthcoming

[6] Fudenberg, D. and Levine, D. (1989): "Reputation and Equilibrium selection in Games with a Single Patient Player", *Econometrica*, **57**, 251–68.

[7] Mailath, G. and Morris, S. (2002): "Repeated Games with Almost-Public Monitoring", *Journal of Economic Theory*, **102**, 189-228

[8] Morris, S. (2001): "Political correctness", *Journal of Political Economy*, **109**, 231–65

[9] Piccione, M. (2002): "The Repeated Prisoner's Dilemma with Imperfect Private Monitoring", *Journal of Economic Theory*, **102**,

Proof of Proposition 2: We check that for an appropriate choice of T the agent has incentives to make an effort after any other realizations of the signals. Incentives are more difficult to provide right after a good realization. Let V be the expected value for the agent under the strategy σ^T in the continuation game following a good signal. We have:

$$V = (1-\delta)[1 + \delta p V + \delta^2 p(1-p)V + \cdots + \delta^T p(1-p)^{T-1} V]$$

Let

$$A = \frac{\delta(1-p)\delta^T}{(1-\delta)}. \tag{1}$$

We may rewrite V as

$$V = \frac{1}{1+A}.$$

If the agent deviates and suspend effort, he obtains an expected payoff V^d that satisfies

$$V^d = (1-\delta)[1 + g + \delta q V + \delta^2 q(1-q)V + \cdots + \delta^T q(1-q)^{T-1}V].$$

Let $\lambda = \frac{1-q}{1-p}$. We have

$$V^d = \frac{1+g}{1+\lambda A}.$$

The agent does not wish to suspend effort when $V^d < V$, or equivalently, when

$$A(\ell - 1 - g) > g \qquad (2)$$

If $\ell - 1 > g$, and if δ is sufficiently close to 1, it is possible to choose A that satisfies both (1) and (2), hence it is thus possible to construct an equilibrium in which the relationship does not terminate right away.[7]

[7]This is not the most efficient equilibrium one can construct, but this is sufficient for our purpose.