"IMMEDIATE SELF-REPORTING OF MISTAKES"

by

Luc WATHIEU*

93/54/TM

* PhD. student, at INSEAD, Boulevard de Constance, Fontainebleau 77305 Cedex, France.

Printed at INSEAD, Fontainebleau, France

IMMEDIATE SELF-REPORTING OF MISTAKES*

LUC WATHIEU

Abstract

Individuals are naturally tempted to delay reporting of mistakes that occur under their responsibility. This results from ordinary features of intertemporal preferences. In many settings however, a principal (a superior authority) needs to be immediately notified, so that she can quickly react to minimize the consequences of the mistake. It then becomes of primary importance to obtain an immediate self-report from the responsible agent. In the incentives scheme that we propose, we give an unusual role to monitoring: the principal sets monitoring dates so as to manipulate the agent's temporal horizon, and this (together with an appropriate combination of real incentives) suffices to ensure timely self-reporting.

* Special thanks are due to Jerry Ross and Bernard Sinclair-Desgagné for their suggestions and encouragement. I also thank Albert Angehrn, Olivier Cadot, Xavier de Groote, Wesley A. Magat, Xavier Wauthy and fellows PhD students at INSEAD for their comments. All errors and shortcomings are mine. I gratefully acknowledge financial support from C.I.M. (Brussels).

I. INTRODUCTION

What incentives could we design to ensure that an agent will find in her best interest to immediately report her own mistakes? Because she fears painful consequences (such as punishment or loss of reputation), a responsible agent would have a natural tendency to keep her faults hidden, at least temporarily. In many settings however, a principal (a superior authority) needs to be promptly notified, in order to quickly react to minimize the consequences of the mistake. The purpose of the present paper is to resolve this type of conflict of interests. We focus on mistakes that could remain unobserved for some time, so that immediate information can only come from the deviant agent's voluntary self-report.

There is a wide variety of potential applications for a theory on how to obtain immediate self-reporting of mistakes. Important applications would concern loan officers in banks who have given credit to mismanaged firms, or (in a completely different domain) doctors involved in the transfusion of contaminated blood. What we will provide here may also apply to the case of a person who kills another person and wonders whether to immediately confess. Self-report, in this case, would save substantial inquiry costs to the police.

The problem under examination is of particular relevance in the context of environmental pollution. Indeed, when an environmental accident happens, it is generally crucial that emergency signals be emitted without any delay.[1] But it is tempting for a polluter to postpone signalling (together with its painful consequences), in the hope of either getting temporary relief or benefiting from a dilution of her responsibility. Such procrastination has notably been observed in the case of Chernobyl's nuclear accident. Also, inside firms, the problem of enforcing immediate self-reports by employees is

[1] See, e.g., "Transnational corporations and industrial hazards disclosure".

certainly of great concern for chief executive officers who are aware of the reputational impact of pollution disasters.

From a more theoretical point of view, our work sheds some light on the timing of communication in organizations. We see it as a combination of regular evaluations and emergency meetings. Emergency meetings are demanded by agents, who perceive in their best interest to do so, thanks to appropriate incentives. Marschak and Radner [1971] offered many insights about a similar type of dynamic coordination (which they called "management by exception") in the context of cooperative teams, assuming away conflicts of interests and corresponding incentive problems. But no existing agency model has considered the timing of communication as an object of conflict in need of resolution. On the other hand, this topic has received some attention from researchers in organizational behavior (see, e.g., Ross and Staw [1991] and the references therein).

This paper can also be seen as an essay in the theory of law enforcement. Traditionally (Becker [1968]), offences were assumed to be detected through (randomized) monitoring. Recently, Mookherjee and Png [1992] suggested that enforcement policies might rely on complaints by victims. Here, we add a new detection device: self-reporting of deviant behavior. And we analyze how incentives for immediate reporting of mistakes interfere with penalties for non-compliance. Malik [1993] and Kaplow and Shavell [1991] address the issue of self-reporting, but with a different emphasis: they examine the properties of monitoring procedures that rely on self-reported data.

We will assume that the principal can intervene at three levels, in order to obtain immediate self-reporting: (i) he can modify the painful consequences experienced by the agent for committing a mistake, (ii) he can design penalties for late report, and (iii) he can use the threat of monitoring.

Here, monitoring has a non-standard role: it defines an *artificial horizon* for the agent and thereby constitutes a manipulation of the agent's decision problem. This is very much consistent with Akerlof's [1991, p.2] suggestion that "an important function of

management may be to set schedules and deadlines and not simply to establish "appropriate" price-theoretic incentives schemes to motivate employees". However, it should be pointed out that manipulation may not always be necessary.

We find that the optimal structure of incentives implies a relatively low penalty for the agent's responsibility in the accident *per se*. This penalty is always well below the agent's liability constraint, under immediate self-report. In some cases, one could even have a compensation from the principal to the inadvertent agent in order to reduce the agent's loss. We also introduce the problem of optimal monitoring. Monitoring periodicity determines the fines that can be credibly imposed in case of (self-reported) mistake and, thereby, it also determines how prudent the agent will decide to be. The principal faces a trade-off between the possibility of imposing high fines in case of a mistake (which we find to be credible only under frequent monitoring) and the cost of monitoring.

The following section sets up the model and section III gives the analytical results. In section IV, we discuss how to integrate the problem of obtaining immediate self-reports in a more global agency problem. Section V concludes.

II. THE MODEL

A. Objective of the agent and liability constraint

The inadvertent agent a decides on when to report, according to the following initial objective:

$$\min_{d\in\{0,\ldots,T\}} \left[x_a(d) = \left(c_a(d) + f_1 + f_2(d)\right)\cdot \varphi_a(d)\right] \qquad (1)$$

where d is the delay in self-report, and $x_a(d)$ is the disutility experienced by the responsible agent, as a function of this delay.

Date 0 is the instant when the mistake occurs. At date T, called the agent's horizon, the accident becomes observable by the principal (e.g., thanks to monitoring, or because a victim reports).

$c_a(d) > 0$ represents the adverse consequences, in monetary terms, experienced by the agent as of the instant of report, *except* those that imply a transfer to the principal. It can be a loss of reputation, for instance. We write c_a as a function of d, because the agent might try to reduce the adverse consequences before reporting.[2] f_1 is the fine, in monetary terms, imposed by the principal to the responsible agent, as a punishment for inadvertence. We do not restrict it to a positive value. $f_2(.)$, a function with $f_2(0) = 0$, gives the fine, also in monetary terms, imposed by the principal as a punishment for delay in self-report.

$\varphi_a(.)$ is a generalized discount function that encompasses standard exponential discounting as well as more descriptively relevant discount factors à la Loewenstein and Prelec (1992). The only requirements are $\varphi_a(0) = 1$, $0 < \varphi_a(.) \leq 1$, and $\varphi_a(.)$ is non-increasing. The discount function in Akerlof (1991), for instance, would correspond to the special case with $\varphi_a(d) = \alpha$, $0 < \alpha < 1$, $\forall d \geq 1$.

As must be obvious by now, we assume time-separability and a linear utility function.

Now, the agent is also characterized by a *liability constraint* that needs to be fulfilled by the incentives scheme, for the fines to be acceptable:

$$c_a(d) + f_1 + f_2(d) \leq \bar{x}_a \,, \quad \forall d \geq 0 \qquad (2)$$

where $\bar{x}_a$ is strictly positive. (2) can be seen as a participation constraint, with $\bar{x}_a - c_a(d)$ being the amount that separates what the agent can get under the agency contract from her outside options.

---

[2] The idea of a dilution of responsibility, evoked in the introduction, might be captured in part by this function.

B. Problem of the principal and the need for an immediate report

For the very purpose of this paper, it would be enough to characterize the principal by a willingness to (and the power to) set up both monitoring and credible penalties, so as to ensure immediate self-reporting by the deviant agent. However, we find interesting to write down a more detailed objective for the principal p : this will allow us to clarify what determines the need for an immediate report.

The decision problem is as follows:

$$\min_{f_1, f_2(.)} \left[x_p(d) = \left(c_p(d) - f_1 - f_2(d) \right) \cdot \varphi_p(d) \right] \qquad (3)$$
$$\textit{subject to } (2)$$

where $x_p(d)$ is the disutility of the principal, as a function of the delay in the agent's report.

$c_p(d) > 0$ represents the adverse consequences, in monetary terms, experienced by the principal because of the mistake. We do *not* restrict $c_p(d)$ to a non-decreasing function, since the agent's actions before reporting might somewhat correct the effects of her inadvertence. $\varphi_p(.)$ represents the time-preferences of the principal. We impose the same requirements as for $\varphi_a(.)$. Here again, we assume time-separability and a linear utility function.

Now, under which condition would an immediate report be necessary? We propose the following sufficient condition, that defines what we could call an *absolute need* for immediate self-reporting:

$$c_p(0) \leq x_p(d), \quad \forall d \in \{1,...,T\}$$

when announced fines are such that the agent's liability constraint is binding. That is, the present paper focuses on cases such that [by (2) and (3)]:

$$c_p(0) \leq \left(c_p(d) + c_a(d) - \bar{x}_a \right) \cdot \varphi_p(d), \quad \forall d \in \{1,...,T\}. \qquad (4)$$

Informally, this condition means that neither the imposed fines nor the corrective actions of the agent can possibly compensate the principal for the consequences of any delay in reporting. The absolute need for an immediate report is favored by a relatively low liability of the agent, or by relatively high principal's discount factors. Inequation (4) also indicates that we focus on mistakes that have little (or not too big) initial consequences for the principal, in comparison with subsequent consequences for both the principal and the agent.

III. INCENTIVES DESIGN

From expressions (1) and (2), it is apparent that immediate self-reporting is credibly enforced when, $\forall d \in \{0,...,T\}$, f_1 and $f_2(d)$ are such that:

$$\begin{cases} c_a(0) + f_1 \leq (c_a(d) + f_1 + f_2(d)) \cdot \varphi_a(d) \\ c_a(d) + f_1 + f_2(d) \leq \bar{x}_a \end{cases}$$

This system of conditions reduces to

$$\bar{x}_a \geq c_a(d) + f_1 + f_2(d) \geq (c_a(0) + f_1) \cdot \frac{1}{\varphi_a(d)}, \quad \forall d \in \{0,...,T\}. \qquad (5)$$

Now, immediate self-reporting implies that only fine f_1 is actually received by the principal. Therefore, f_1 should be set as high as possible. The need for consistent boundaries in condition (5), however, implies that

$$f_1 \leq \bar{x}_a \cdot \varphi_a(d) - c_a(0), \quad \forall d \in \{0,...,T\}. \qquad (6)$$

Given (6) and our assumptions on the shape of $\varphi_a(.)$, the principal should set

$$f_1^* = \bar{x}_a \cdot \varphi_a(T) - c_a(0). \qquad (7)$$

This equation reveals that two identical offences would be punished differently depending on how remote observability is. More precisely, *for a given type of mistake, the announced fine associated with responsibility should be a non-increasing function of the agent's horizon.* To understand this intuitively, it suffices to realize that the penalty associated with immediate reporting (f_1^*) needs to be lower than the discounted penalty associated with any possible delayed reporting. Since remote penalties are more heavily discounted, this constraint is more stringent when the horizon is large.

Because frequent monitoring reduces the expected length of T, *there is a direct positive relation between the pace of monitoring and the fine that the principal can expect to require in case of a mistake.* Indeed, there exists an underlying problem of jointly selecting the monitoring policy and f_1, knowing the cost of monitoring and equation (7), and taking into account that the agent has some power on the probability of a mistake occurring.[3]

We observe that f_1^* *is always below the agent's liability constraint.* [Since $f_2(0) = 0$ and immediate self-reporting is obtained, this constraint is $f_1 \leq \bar{x}_a - c_a(0)$].

Note also that f_1^* *can even be negative*: it then plays the role of a *premium* for immediate self-reporting. The presence of such a premium in the enforcement scheme is favored under the following characteristics of the agent: liability is relatively low, horizon is long, discount function is sharply decreasing with time, or the adverse (non transferable) consequences experienced at the instant of report are heavy.

The penalty for delay in reporting, $f_2^*(d)$ can be obtained by substituting (7) in (5). One gets, $\forall d \in \{0,...,T\}$:

$$\bar{x}_a(1-\varphi_a(T)) + c_a(0) \geq f_2^*(d) + c_a(d) \geq \bar{x}_a \cdot \varphi_a(T) \cdot \frac{1-\varphi_a(d)}{\varphi_a(d)} + c_a(0). \qquad (8)$$

[3] See the discussion in the next section.

[This formula gives boundaries for the total consequences experienced by the agent for a delay of length d.]

Interestingly enough, the announced fine for late report $f_2^*(d)$ *can be negative.* Thus, a *premium for late report* would be offered, for any delay such that [using the upper bound of (8)]:

$$c_a(d) \geq \bar{x}_a \cdot (1 - \varphi_a(T)) + c_a(0). \qquad (9)$$

This paradoxical result comes from the necessity to respect the agent's liability constraint, while imposing f_1^*, when the adverse consequences experienced by the procrastinating agent are already high. Such high adverse consequences constitute, by themselves, a threat that largely suffices to induce immediate reporting, but that might jeopardize participation of the agent if no compensation is offered.

If, for some delay d, $c_a(d)$ is as in (9), and if $c_a(0) \geq \bar{x}_a \cdot \varphi_a(T)$, *then both* f_1^* *and* $f_2^*(d)$ *should be premiums.* Such instances are far from absurd, as the example of health care workers who catch AIDS (and strongly fear unemployment) may suggest.

Upper bound in (8) decreases as the agent's horizon reduces. Therefore, *we would generally expect low penalties for delays in self-report under frequent monitoring.*

Now, we note that information problems about d should be taken into account in the design of incentives. In particular, suppose that d *cannot be assessed precisely*, but the principal can tell whether there *was* a delay in self-report. Penalty f_2 must then be of the form

$$\begin{cases} f_2(0) = 0 \\ f_2(d \mid d > 0) = f_2^{\oplus} \end{cases}$$

where $f_2^{\oplus}$ is a constant. We thus loose some flexibility in design. Also, condition (5) becomes

$$\begin{cases} \bar{x}_a - c_a(0) \geq f_1 \\ \bar{x}_a \geq c_a(d) + f_1 + f_2^{\oplus} \geq (c_a(0) + f_1)\dfrac{1}{\varphi_a(d)}, \quad \forall d \in \{1,\ldots,\Delta\} \end{cases} \tag{10}$$

where Δ is the length of time between the latest monitoring date and the announced next one. Δ is the lowest integer that the principal can guarantee to be greater than or equal to T.

Condition (10) is equivalent to

$$\begin{cases} \bar{x}_a - c_a(0) \geq f_1 \\ \bar{x}_a - \max\limits_{d\in\{1,\ldots,\Delta\}} c_a(d) \geq f_1 + f_2^{\oplus} \geq (c_a(0) + f_1)\dfrac{1}{\varphi_a(\Delta)} - \min\limits_{d\in\{1,\ldots,\Delta\}} c_a(d) \end{cases}$$

Consistency of the boundaries in the above constraint and optimization of f_1 yields

$$f_1^{\oplus} = \bar{x}_a \cdot \varphi_a(\Delta) - c_a(0) - \left(\max_{d\in\{1,\ldots,\Delta\}} c_a(d) - \min_{d\in\{1,\ldots,\Delta\}} c_a(d) \right) \cdot \varphi_a(\Delta). \tag{11}$$

As we can see from (11), *the principal's ignorance of d leads to an even lower penalty for the mistake per se.* Also, the direct relation between frequency of monitoring and expected revenue from fines becomes *conditioned by*

$$\bar{x}_a \geq \max_{d\in\{1,\ldots,\Delta\}} c_a(d) - \min_{d\in\{1,\ldots,\Delta\}} c_a(d). \tag{12}$$

$f_2^{\oplus}$ will be negative whenever $f_1^{\oplus}$ threatens respect of the agent's liability constraint (2), for *any* $d \in \{1,\ldots,\Delta\}$.

IV. COMPATIBILITY WITH CAREFULNESS

In the above developments, we show how the need for an immediate report generates a constraint on the penalty that can be credibly imposed on the inadvertent agent. Then, a

question naturally arises: is the announced penalty still high enough to induce effort (carefulness) in the first place?

There is no way to impose more than f_1^* if one needs an immediate self-report in case of a mistake. However, (7) constrains the penalty *for a given horizon T*. Therefore, by setting a pace of monitoring, the principal can actually determine the sequence of fines f_1^*. Thereby, she can control the average carefulness of an effort-averse agent who might otherwise prefer to take too much risk of being penalized (rather than being cautious).

By playing with both monitoring and penalties, it is possible to induce both immediate self-reporting of mistakes and a desirable average degree of carefulness.

V. CONCLUSION

This paper introduced incentive schemes that ensure immediate self-reporting of mistakes. The fine imposed on the agent for having deviated is a decreasing function of the agent's horizon. This fine is always well below the agent's liability constraint, and strongly so when the agent's discount function is sharply decreasing, and/or when the principal can not precisely estimate the delays in self-reporting. Actually, the fine might even become a premium. The average fine increases as the pace of monitoring accelerates. Even the announced penalties for delays in self-report might also become premiums. On average, these penalties are low under frequent monitoring.

A premium for disclosing mistakes is probably an underutilized device in the real world: since we don't want to reward undesirable behavior, we end up actually losing social utility by encouraging people to cloak their condition. In our framework, however (as in many real-life situations), even when the principal offers a premium as an incentive for self-reporting, the overall payoff for committing a mistake is always negative.

We should emphasize the important role played by the agent's limited liability. Low liability increases both the need for immediate self-reporting and the chances that the optimal fine is actually a premium.

Our results show that efficient enforcement of early avowals does not require an intensive and costly control activity by the principal. This may help understanding some puzzling stylized facts about the enforcement of laws. In the case of environmental compliance, empirical studies [as synthesized in Harrington (1988)] typically show that (i) the frequency of surveillance is low, (ii) low fines are imposed when a violation is discovered and (iii) sources are nonetheless in compliance a large part of the time.[4] As Harrington puts it: "if enforcement activity is carried on at such a low level, and if violations are rarely punished, why would any firm bother to comply?". We can offer an explanation for these facts, by assuming that the enforcement schemes partly aim at encouraging self-report of pollutions. Fact (iii) simply shows that the enforcement schemes are effective. Facts (i) and (ii) respectively correspond, in the terms of our model, to large average T and relatively low fine f_1. This can very well be seen as an illustration of the propositions that we deduced from equation (7).

Finally, we would like to point out that our work provides some support for a well-known piece of French conventional wisdom, which says: "Faute avouée est à moitié pardonnée" ("A sin confessed is a sin half-pardoned"). Mercy is perhaps a subject worthy of further economic investigation.

THE EUROPEAN INSTITUTE OF BUSINESS ADMINISTRATION (INSEAD)

[4] Harrington offers game-theoretic interpretations of these stylized facts.

REFERENCES

Akerlof, George A., "Procrastination and Obedience", American Economic Review, 81 (1991), 1-19.

Becker, Gary, "Crime and Punishment: An Economic Approach", Journal of Political Economy, 76 (1968), 169-217.

Harrington, Winston, "Enforcement Leverage when Penalties are Restricted", Journal of Public Economics, 37 (1988), 29-53.

Kaplow, Louis and Steven Shavell, "Optimal Enforcement with Self-Reporting of Behavior", NBER Working Paper #3822 (1991).

Loewenstein, George and Drazen Prelec, "Anomalies in Intertemporal Choice: Evidence and an Interpretation", Quarterly Journal of Economics, 107 (1992), 573-97.

Malik, Arun S., "Self-Reporting and the Design of Policies for Regulating Stochastic Pollution", Journal of Environmental Economics and Management, 24 (1993), 241-57.

Marschak, Jacob and Roy Radner, "Economic Theory of Teams" (Yale University Press, New Haven, 1971).

Mookherjee, Dilip and I.P.L. Png, "Monitoring vis-à-vis Investigation in Enforcement of Law", American Economic Review, 82 (1992), 556-65.

Ross, Jerry and Barry M. Staw, "Managing Escalation Processes in Organizations", Journal of Managerial Issues, 3 (1991), 15-30.

United Nations Centre on Transnational Corporations, "Transnational Corporations and Industrial Hazards Disclosure" (United Nations, New York, 1991).