Review of Economic Studies (2003) **70**, 489–520 0034-6527/03/00190489$02.00

Intrinsic and Extrinsic Motivation

ROLAND BÉNABOU
Princeton University and Institute for Advanced Study

and

JEAN TIROLE
IDEI (Université de Toulouse I), CERAS and MIT

First version received February 2000; *final version accepted January* 2003 (*Eds.*)

A central tenet of economics is that individuals respond to incentives. For psychologists and sociologists, in contrast, rewards and punishments are often counterproductive, because they undermine "intrinsic motivation". We reconcile these two views, showing how performance incentives offered by an informed principal (manager, teacher, parent) can adversely impact an agent's (worker, child) perception of the task, or of his own abilities. Incentives are then only weak reinforcers in the short run, and negative reinforcers in the long run. We also study the effects of empowerment, help and excuses on motivation, as well as situations of ego bashing reflecting a battle for dominance within a relationship.

> Tom said to himself that it was not such a hollow world, after all. He had discovered a great law of human action, without knowing it—namely, that in order to make a man or a boy covet a thing, it is only necessary to make the thing difficult to attain. If he had been a great and wise philosopher, like the writer of this book, he would now have comprehended that Work consists of whatever a body is obliged to do, and that Play consists of whatever a body is not obliged to do.
>
> Mark Twain, *The Adventures of Tom Sawyer* (1876, Chapter 2).

INTRODUCTION

Should a child be rewarded for passing an exam, or paid to read a book? What impact do empowerment and monitoring have on employees' morale and productivity? Does receiving help boost or hurt self-esteem? Why do incentives work well in some contexts, but appear counterproductive in others? Why do people sometimes undermine the self-confidence of others on whose effort and initiative they depend?

These questions will be studied here from a unifying perspective, emphasizing the interplay between an individual's personal motivation and his social environment. We shall thus model the interactions between an agent with imperfect self-knowledge and an informed principal who chooses an incentive structure, such as offering rewards and threatening punishments, delegating a task, or simply giving encouragement, praise, or criticism.

It is a central theme of economics that incentives promote effort and performance, and there is a lot of evidence that they often do (*e.g.* Gibbons (1997), Lazear (2000)). In other words, contingent rewards serve as "positive reinforcers" for the desired behaviour. In psychology, their effect is much more controversial. A long-standing paradigm clash has opposed proponents of the economic view to the "dissonance theorists", who argue that rewards may actually impair performance, making them "negative reinforcers", especially in the long run (see, *e.g.* Kruglanski (1978) for an account of this debate, and Deci, Koestner and Ryan (1999) for a recent and comprehensive meta-analysis of experimental results).

Indeed, a substantial body of experimental and field evidence indicates that extrinsic motivation (contingent rewards) can sometimes conflict with intrinsic motivation (the individual's desire to perform the task for its own sake). In a now classical experiment (see Deci, 1975), college students were either paid or not paid to work for a certain time on an interesting puzzle. Those in the no-reward condition played with the puzzle significantly more in a later unrewarded "free-time" period than paid subjects, and also reported a greater interest in the task. This experiment has since been replicated many times, with numerous variations in design (*e.g.* Wilson, Hull and Johnson, 1981) and in types of subjects. For instance, similar effects were found for high-school students in tasks involving verbal skills (Kruglanski, Friedman and Zeevi, 1971), and for preschool children in activities involving drawing with new materials (Lepper, Greene and Nisbett, 1973). In daily life, parents are quite familiar with what we shall call the "forbidden fruit" effect: powerful or salient constraints employed by adults to enforce the prohibition of some activity often decrease the child's subsequent internalization of the adults' disapproval.[1] Kohn (1993) surveys the results from a variety of programmes aimed at getting people to lose weight, stop smoking, or wear seat belts, either offering or not offering rewards. Consistently, individuals in "reward" treatments showed better compliance at the beginning, but worse compliance in the long run than those in the "no-reward" or "untreated controls" groups. Taken together, these many findings indicate a limited impact of rewards on "engagement" (current activity) and a negative one on "re-engagement" (persistence).

A related body of work transposes these ideas from the educational setting to the workplace. In well-known contributions, Etzioni (1971) argues that workers find control of their behaviour via incentives "alienating" and "dehumanizing", and Deci and Ryan (1985) devote a chapter of their book to a criticism of the use of performance-contingent rewards in the work setting.[2] And, without condemning contingent compensation, Baron and Kreps (1999, p. 99) conclude that:

> There is no doubt that the benefits of [piece-rate systems or pay-for-performance incentive devices] can be considerably compromised when the systems undermine workers' intrinsic motivation.

Kreps (1997) reports his uneasiness when teaching human resources management and discussing the impact of incentive devices in a way that is somewhat foreign to standard economic theory. And indeed, recent experimental evidence on the use of performance-contingent wages or fines confirms that explicit incentives sometimes result in worse compliance than incomplete labour contracts (Fehr and Falk (1999), Fehr and Schmidt (2000), Gneezy and Rustichini (2000*a*)). Relatedly, Gneezy and Rustichini (2000*b*) find that offering monetary incentives to subjects for answering questions taken from an IQ test strictly decreases their performance, unless the "piece rate" is raised to a high enough level. In the policy domain, Frey and Oberholzer-Gee (1997) surveyed citizens in Swiss cantons where the government was considering locating a nuclear waste repository; they found that the fraction supporting siting of the facility in their community fell by half when public compensation was offered.

Our aim here will be twofold. First, we want to analyse the "hidden costs" of rewards and punishments from an economic and cognitive perspective, rather than just posit an aversive impact on motivation. Indeed, given that incentives work quite effectively in many instances, one needs to understand in what cases they should be used with caution. More generally, we seek to give a precise content to the loosely defined notions of intrinsic and extrinsic motivation, and to clarify when, in the terminology of Frey (1997), the latter should be expected to "crowd out"

1. See, *e.g.* Lepper and Greene (1978). Relatedly, Akerlof and Dickens (1982) suggest that imposing stiffer penalties for crimes might be counterproductive, if it undermines individuals' "internal justification" for obeying the law.
2. See also Lepper and Greene (1978), Kohn (1993) and Frey (1997).

or "crowd in" the former. This information-based, strategic analysis distinguishes our approach from Frey's reduced-form treatment of these issues.

We consider an individual (the agent, "he") who faces uncertainty about his payoff from taking a particular action. The unknown variable could be a characteristic of the person himself, such as raw ability, of the specific task at hand (long-run return, how difficult or enjoyable it is to complete, etc.), or of the match between the two. Naturally, the agent will undertake the task only if he has sufficient confidence in his own ability to succeed, and in the project's net return. As a result, people with a stake in his performance have strong incentives to manipulate signals relevant to his self-knowledge. Given that effort and ability are usually complements in the production of performance, they will want to boost his self-confidence, as well as his interest in the task. Thus, in much of this paper, a principal (parent, spouse, friend, teacher, boss, colleague, etc., "she") has a vested interest in (derives a benefit from) the agent's undertaking and succeeding in the activity.

In many circumstances, both parties have private information about the agent's suitability to the task. The agent usually has better knowledge of his previous performances and of the relevant circumstances (his effort intensity, the idiosyncratic factors that may have come into play). He will often also receive privately signals about the attractiveness or unpleasantness of the task, either from third parties (friends tell him that school is not fun, while cigarettes are cool), from having performed similar ones in the past, or simply from his own experience as he starts carrying out the current one. The principal, on the other hand, often has complementary private information about the task or the agent's prospects from it. For example, a teacher or manager is better able to judge the difficulty of the subject or assignment, which, together with the agent's ability, conditions the probability of success. The principal may know better than the agent whether the task is attractive, in terms of either being enjoyable to perform, or having a high payoff for the agent. Last, while having less direct information about the agent's previous performances, she may be better trained at interpreting it due to her having performed the task herself, or having seen many others attempt it. As we shall discuss later on, the observation that others may have private information relevant to an individual's self-view underlies several fields of research in education and management. It is this type of private information that will be our focus.[3]

In the first part of the paper we thus study the *attributions* made by an agent when a principal with private information makes a decision, such as selecting a reward, delegating a task or more simply encouraging the agent, that impacts the latter's willingness to perform the task. As was pointed out by Cooley (1902), the agent should then take the principal's perspective in order to learn about himself. The agent's attribution of ulterior motivation to the principal, or, in economics parlance, his attempt to infer her private information from her decision, is what Cooley termed the "looking-glass self". The influence of the principal's decision on the agent's behaviour is then twofold: direct, through its impact on the agent's payoff from accomplishing the task (keeping information constant), and indirect, through his inference process. In analysing intrinsic and extrinsic motivation we thus adopt a cognitive approach, assuming that the individual seeks to extract from the words and deeds of those around him signals about what they know that concerns him.[4]

3. Delfgauw and Dur (2002), in contrast, focus on the more standard case where workers have private information about their own (dis)utility from working on the task, which they may then want to signal to, or conceal from, the employer.

4. We in fact focus on the polar case where individuals are fully rational and Bayesian. Although people surely make mistakes in processing information, we want the model to reflect the fact that they cannot constantly fool themselves, or others.

We first show that rewards may be only *weak reinforcers* in the short term and that, as stressed by psychologists, they may have *hidden costs*, in that they become negative reinforcers once they are withdrawn. The idea is that by offering low-powered incentives, the principal signals that she trusts the agent. Conversely, rewards (extrinsic motivation) have a limited impact on current performance, and reduce the agent's motivation to undertake similar tasks in the future. We then use the same logic to show that empowering the agent is likely to increase his intrinsic motivation. Similarly, help offered by others may be detrimental to one's self-esteem and create a dependence.

More generally, we conclude that explicit incentives may, but need not, be negative reinforcers; our analysis actually suggests when rewards and punishments work, and when they backfire. The "crowding out" case first requires that the agent be less knowledgeable in some dimension than the principal; this asymmetry of information is likely to be more important in some settings (education, health, new occupations) than in others (relatively standardized jobs). Furthermore, a sorting condition must hold, in that the principal must be more inclined to offer a reward when the agent has limited ability or the task is unattractive. Otherwise, there will be "crowding in". Thus, when concerned about a potential negative impact of rewards, one should first check whether the reward provider has private information about the task or the agent's talent. One should then, as the agent does, think through the provider's ulterior motivation and how her payoff from giving a contingent reward is affected by her knowledge.

In addition to low-powered incentives, we also investigate how the principal can sometimes use non-contingent payments (similar to "burning money") to signal her confidence in the agent's ability. While their short-term incentive effects differ, reducing the slope of the compensation schedule (the piece rate) and increasing its base part (the fixed salary) are two related ways in which the principal's confidence-management motive will be reflected in equilibrium contracts. Each has its domain of applicability (as we show), but both have similar effects on intrinsic, or long run, motivation, as well as on wage inequality. Indeed, by weakening the link (elasticity) between performance and compensation, both signalling strategies reduce earnings inequality across workers, in a Lorenz sense.

While most of the social psychology and human resource management literatures emphasize the necessity of boosting and protecting the self-esteem of one's personal and professional partners, people often criticize or downplay the achievements of their spouse, child, colleague, coauthor, subordinate or teammate. In the second part of the paper we consider several reasons why this may be, and formalize in more detail what is perhaps the most common one. We argue that such "ego bashing" may reflect *battles for dominance:* by lowering the other's self-confidence, an individual may gain real authority within the relationship, enabling her to steer joint decisions or projects in a preferred direction. This generally comes at a cost, however, namely the risk of demotivating the partner from seeking good projects, or from exerting effort at the implementation stage. We study this tradeoff, distinguishing two related forms of ego bashing: one is "by omission", where the principal omits to report news favourable to the agent; the other is active "disparaging", in which she explicitly belittles the agent. While both strategies lower the agent's self-confidence, the first one is reversible (the news can always be revealed later on), whereas the second is not. This is shown to have interesting implications for the timing of strategic disclosures of information (ego bashing and ego boosting) in situations where both the agent's initiative and joint control rights are at stake.

The paper is organized as follows. Section 1 provides a general introduction to the "looking-glass self" mechanism. Section 2 analyses the interplay of intrinsic and extrinsic motivation, focusing in particular on the hidden cost of rewards. Section 3 shows how the main insights carry over to other confidence-management strategies such as delegation, help and coaching. Section 4 studies the costs and benefits of ego bashing. Section 5 offers concluding remarks.

1. THE LOOKING-GLASS SELF

We begin here with a relatively general and abstract framework, then specialize it in the rest of the paper. Readers more interested in specific psychological and economic applications than in a unified presentation of the underlying mechanisms may want to proceed directly to Section 2.

There are two players, an agent (he) and a principal (she). The agent selects a continuous action or effort level e that impacts both his and the principal's utilities. The principal knows a parameter β, such as the difficulty of the task or the agent's ability to perform it, that affects the agent's payoffs from e. Thus informed, she selects a policy p (belonging to $\mathbb{R}$, for expositional simplicity) prior to the agent's choice of action; this may be a wage or contingent reward, help, surveillance, delegation, disclosure of information, or any other "extrinsic motivator" that can affect, directly or indirectly, the agent's behaviour. The agent's and the principal's payoffs are denoted $U_A(\beta, e, p)$ and $U_P(\beta, e, p)$. Prior to his decision, the agent may also privately receive a signal σ that is informative about β. We shall assume for simplicity that this signal is redundant if one already knows the principal's information (β is a sufficient statistic for (β, σ)), but none of our main conclusions hinge on this assumption. What really matters is that the principal has information relevant to the agent's perception of himself or his task, and (for the specific "trust effect" discussed below) that the principal be uncertain about the agent's motivation. The timing of the game is as follows:

Stage 1: The principal learns the parameter β and selects a policy p.
Stage 2: After observing the policy chosen by the principal and learning σ, the agent chooses an action e.

Let us assume here (for notational simplicity) that the agent's optimal action e^* depends only on p and on his conditional expectation $\hat{\beta}(\sigma, p)$ of the unknown parameter.[5] The conditioning of $\hat{\beta}$ on p is the "looking-glass-self" phenomenon, whereby the agent tries to see through the principal's ulterior motives that led to p being selected. As long as the agent's participation in the relationship is not at stake, the principal's expected payoff from choosing a policy p when she has information β is thus

$$\mathsf{E}_\sigma[U_P(\beta, e^*(p, \hat{\beta}(\sigma, p)), p) \mid \beta].$$

Assuming differentiability (again for simplicity), the principal's choice of policy takes three effects into consideration:

$$\mathsf{E}_\sigma\left[\frac{\partial U_P}{\partial p} + \frac{\partial U_P}{\partial e}\cdot\frac{\partial e^*}{\partial p} + \frac{\partial U_P}{\partial e}\cdot\frac{\partial e^*}{\partial \hat{\beta}}\cdot\frac{\partial \hat{\beta}}{\partial p}\,\middle|\,\beta\right] = 0. \tag{1}$$

The first term on the L.H.S. of (1) is the direct effect of p on the principal's payoff. For example, if the policy is a wage or bonus, as in the next section, this term is the direct cost of this compensation, keeping the agent's behaviour constant. The second term corresponds to the direct impact of p on the agent's behaviour. Thus, *ceteris paribus*, a bonus increases the incentive to exert effort. These two effects have been investigated in detail in the agency literature.

We shall be interested in the third, more novel effect, which corresponds to the principal's *confidence-management* motive. Whenever the principal's choice of policy is guided by private information, the agent will update his beliefs in reaction to the choice of p (term $\partial\hat{\beta}/\partial p$). The principal must then take into account how the agent's interpretation of her choice will affect his self-confidence—that is, his perceived prospects from undertaking the task. A particularly important issue is whether a higher level of self-confidence affects the agent's decision making in

5. More generally, it will depend on p and on the conditional distribution of β, given (σ, p).

a direction that the principal likes ($(\partial U_P/\partial e)(\partial e^*/\partial \hat{\beta}) > 0$) or dislikes ($(\partial U_P/\partial e)(\partial e^*/\partial \hat{\beta}) < 0$). Sections 2 and 3 will examine many common situations where the principal gains from boosting the agent's self-confidence. Section 4, on the other hand, will focus on cases where she may be reluctant to enhance the agent's self-confidence, or may even want to undermine it.

The confidence-management motive itself can itself arise through two channels, which we term the *profitability effect* and the *trust effect*. The former arises when the agent's type, on which the principal has private information, enters the *principal's objective function* in a way that would lead her to offer different policies to different agents, even if it did not affect anyone's effort level. This differential profitability of a given policy across types corresponds to a standard sorting condition; thus, for a one-dimensional policy it means that[6]

$$\frac{\partial}{\partial \beta}\left(\frac{\partial U_P/\partial p}{\partial U_P/\partial e}\right) \text{ has a constant sign.} \tag{2}$$

Suppose, for instance, that an employer's expected profits are (proportionally) more sensitive to the employee's ability when the latter is empowered to make decisions than when he is closely monitored. The principal will then, *ceteris paribus*, delegate more to employees she thinks more highly of, and delegation will be seen as good news by employees. In contrast, no such profitability effect exists when the principal's private knowledge concerns the cost of accomplishing the task, or other aspects of it that bear solely on the agent's utility, and not on her own payoff.

The trust (or distrust) effect, on the contrary, arises when the principal's private information concerns a parameter, such as the cost or pleasure of accomplishing the task, that directly enters only in the *agent's incentive problem*—as envisioned by the principal. The issue here is how confident the principal is as to the agent's intrinsic motivation—that is, how she thinks the agent perceives the task and his suitability to it. A principal who has bad news about the agent's parameter β will be pessimistic about the agent's own signal σ, and may consequently fear that he will not be motivated enough to exert effort in the absence of added incentives. Providing stronger incentives, however, will at least partially reveal the principal's damaging information (compounding the signal σ). Thus, once again, extrinsic motivation may "crowd out" intrinsic motivation, and the optimal contract will be shaped by this tradeoff.

It is worth noting also that for the pure trust effect to operate, there must be some uncertainty (σ) on the part of the principal about the exact incentives perceived by the agent. Otherwise, the latter's response $\hat{\beta}(p)$ to any policy p would be perfectly predictable, and the principal would simply maximize $U_P(\beta, e^*(p, \hat{\beta}(p)), p)$. It is easily verified that, absent a profitability effect (*i.e.* when the expression in (2) is zero, *e.g.* when U_P is independent of β), the optimal policy is then completely independent of, and hence uninformative about, the agent's type β. It is quite reasonable to assume, however, that the agent does receive a private signal, causing the principal to worry (be uncertain) about his resulting motivation. As mentioned earlier, this may come from past personal experience, or from friends and brethren. Another important source of information is the agent's own initial perception as he starts performing the task: how he feels after a few weeks or months on the job, at school or in a diet programme; after reading a few chapters of a challenging book, a few minutes of working on a puzzle or painting a fence, etc. In many such real-world cases, the effort decision e should be interpreted as a continuation or "perseverance" decision—that is, whether or or not the agent will carry the task to completion.

6. In Section 2.2.2 we shall actually consider two-dimensional policies, consisting of a lump-sum payment and a performance-contingent bonus. We shall then use the general "implementability condition" (see, *e.g.* Fudenberg and Tirole (1991, pp. 258–260)) that extends the Spence–Mirlees sorting condition to multidimensional policies $p = (p_1, \ldots, p_n) : \sum_{i=1}^{n} \frac{\partial}{\partial \beta}\left(\frac{\partial U_P/\partial p_i}{\partial U_P/\partial e}\right)\left(\frac{dp_i}{d\beta}\right)$ must have a constant sign.

To bring into sharper focus the trust and profitability effects, let us specialize our framework further. Many of our applications (*e.g.* bonuses, help, delegation) will share a common structure, where the principal's payoff function can be written as $U_P(\beta, e, p) = e\Lambda(\beta, p)$. In this formulation, the agent's equilibrium effort e is a zero–one decision of whether or not to undertake the task, and the function Λ is the principal's expected payoff when $e = 1$. The profitability effect is then governed by the sorting condition

$$\frac{\partial}{\partial \beta}\left(\frac{\partial U_P/\partial p}{\partial U_P/\partial e}\right) = \frac{\partial}{\partial \beta}\left(\frac{e\Lambda_p(\beta, p)}{\Lambda(\beta, p)}\right) = e \cdot \frac{\partial^2 \ln U_P}{\partial \beta \partial p} \text{ has a constant sign,} \tag{3}$$

which is a simple form of complementarity.

To capture the pure trust effect, assume now that U_P, hence also Λ, does not depend on β at all. Then, under mild conditions on $U_A(\beta, e, p)$ and the conditional distribution $G(\sigma \mid \beta)$, the agent will work only when he receives a signal σ better than some threshold $\sigma^*(p)$, which depends on the policy p due to the looking-glass phenomenon. The principal's expected profits can then be written as $[1 - G(\sigma^*(p) \mid \beta)]\Lambda(p)$. One may then, intuitively, treat the agent's action threshold σ^* as the effort variable that the principal is trying to influence through her policy p, and look at a sorting condition for the "reduced-form" objective function $[1 - G(\sigma^* \mid \beta)]\Lambda(p)$. This yields

$$\frac{\partial}{\partial \beta}\left(\frac{\partial U_P/\partial p}{\partial U_P/\partial \sigma^*}\right) = \frac{\partial}{\partial \beta}\left(\frac{1 - G(\sigma^* \mid \beta)}{g(\sigma^* \mid \beta)}\right) \times \frac{\Lambda'(p)}{\Lambda(p)}.$$

The sign of this expression corresponds to a monotone likelihood ratio property (MLRP), that will be seen to play a key role in the trust effect. Intuitively, a principal who observes a "bad" β is worried that the agent will receive (or has received) a bad signal σ, so she feels compelled to offer him a higher p. This, in turn, is bad news for the agent.

2. THE HIDDEN COSTS OF REWARDS

We shall now specialize the general framework to a more concrete model, where the interplay of intrinsic and extrinsic motivation can be more transparently and completely analysed. In demonstrating how the "looking-glass self" mechanism can make high-powered incentives schemes too costly for an optimizing principal to adopt, we shall first emphasize the trust effect, then show how the profitability effect can reinforce or counteract it. Finally, we shall relate the premises and results of our model to the relevant psychology literature, and argue that they accord rather well with it.

There may of course be still other sources for the hidden costs of rewards; let us mention here two fairly obvious ones. Concerning the engagement part, Condry and Chambers (1978, p. 66) suggest that "rewards often distract attention from the process of task activity to the product of getting a reward". As for the re-engagement part, these same authors argue that current rewards may decrease the individual's willingness to persist, because they orient activity toward performance rather than progress. In other words, Condry and Chambers offer what to economists is a familiar multitask interpretation: the individual is led by short-term rewards to sacrifice long-run payoffs.[7] Thus, subjects who are paid to solve problems typically choose easier ones than those who do not expect any payment. While this explanation is well taken, it does

7. For example, in Laffont and Tirole (1988) an agent exerts effort today both to reduce current operating cost and to increase future efficiency. Faced with a higher powered incentive scheme (a greater sensitivity of current reward to current cost level), the agent substitutes toward current cost reduction and sacrifices long-term investment. For a broader perspective on multitasking, see Holmström and Milgrom (1991). Condry and Chambers' argument follows a similar pattern, with the individual allocating his attention between the resolution of the current problem and a "deeper understanding" of the problem.

not apply uniformly. For instance, the individual may not be aware of future re-engagement opportunities, or may just not face any investment decision that crowds out current efficiency—as in the previously mentioned programmes involving weight loss, smoking, and seat belts. The multitask story can also not account for the evidence drawn from subjects' posterior reports about their intrinsic interest in the activity.

2.1. *Task attractiveness and the trust effect*

This section describes the interplay between intrinsic and extrinsic motivation in a situation with only a trust effect. It formalizes the idea emphasized in the psychology literature that the subject finds the task less attractive when offered a reward.

As before there are two players, an agent and a principal. The agent chooses whether to undertake an activity or task (exert effort) or not (exert no effort). His disutility or cost of effort is denoted $c \in [\underline{c}, \bar{c}]$. If the task is successful it yields direct payoffs $V > 0$ to the agent and $W > 0$ to the principal; if it fails, their gross payoffs are both equal to 0. Success requires effort, but effort is not sufficient for success: let $\theta \in (0, 1]$ denote the probability of success when the agent exerts effort.

Our focus in this paper is on the principal's superiority of information; in this section, the asymmetry concerns the cost that the agent will bear if he decides to undertake the task; that is, $\beta = c$.[8] With little loss of generality, we assume that the principal knows c perfectly. The agent knows that c is drawn from a cumulative distribution function $F(c)$ with a density $f(c)$ that has full support; he also learns a signal $\sigma \in [0, 1]$ with conditional distribution $G(\sigma \mid c)$ and positive density $g(\sigma \mid c)$. We assume that a higher σ is "good news", in the sense of the MLRP

$$\text{for all } \sigma_1 \text{ and } \sigma_2 \text{ with } \sigma_1 > \sigma_2, \ \frac{g(\sigma_1 \mid c)}{g(\sigma_2 \mid c)} \text{ is decreasing in } c. \tag{4}$$

In order to induce the agent to perform the task, the principal can offer a reward that is contingent on effort if she observes it, or on output if she does not. In the present context where the probability of success θ is common knowledge and both parties are risk-neutral, the former situation is equivalent to the latter, as is perhaps most easily seen for $\theta = 1$. We shall therefore focus the exposition on contracts where the principal selects a reward or *performance-based* "bonus" $b \leq W$, to be paid in case of success. In applications where the agent is paid just to carry out the task, successfully or unsuccessfully (*e.g.* paying a child to read a book, independently of whether it will turn out to be useful to him or his parents), one will remember that when e is observable a bonus scheme is equivalent to a *wage offer* of $w \equiv \theta b$, in exchange for the mere supply of effort. Note also that negative wages or bonuses are allowed (as in the Tom Sawyer illustration discussed below).[9]

We shall initially abstract from the agent's participation constraint, and normalize the non-contingent part of the contract to zero. As we show later on, this does not affect the results, since the agent's cost c has no effect on the principal's gain from inducing him to perform the task. The agent's net benefit in case of success is thus $V + b$ and the principal's is $W - b$, while both parties obtain 0 in case of failure. The stage-1 policy decision for the principal is thus the choice of a reward which we formalize as being a monetary one; but, in line with the psychology literature,

8. Equivalently it could be V, as long as it is uncorrelated with the principal's payoff W. As explained earlier, the pure trust effect concerns private information of the principal that directly enters only in the agents' incentive problem.

9. Alternatively, rewards could be constrained to be non-negative (an assumption that makes sense when the agent can sabotage his observable performance without destroying V, or when V is a private benefit or learning experience from undertaking the task). Proposition 1, in what follows, would then still hold, as the non-negativity constraint affects only the *extent* to which the principal can reveal her information.

b could with slight modifications be interpreted as working conditions, praise, friendliness or (minus) punishment.

Were the agent to know his cost c, he would choose to exert effort if and only if

$$\theta(V + b) \geq c.$$

Thus, when the agent has the same information as the principal, the reward is a positive reinforcer. In our model, however, only the principal observes c; the agent receives only a signal σ about c.

We shall now analyse the perfect Bayesian equilibria of this two-stage game. When offered a reward b, the agent updates his beliefs about c using the principal's equilibrium strategy. Let $\hat{c}(\sigma, b) \equiv \mathsf{E}[c \mid \sigma, b]$ denote the agent's (interim) assessment of the task's difficulty, that is, his expectation of the cost, conditional on his signal and the reward he is offered. This expectation is a weakly decreasing function of the signal σ. Letting $e \in \{0, 1\}$ denote the agent's effort, his utility is $U_A = [\theta(V + b) - \hat{c}(\sigma, b)]e$, and there exists a threshold signal $\sigma^*(b)$ in $[0, 1]$ such that:[10]

$$\hat{c}(\sigma, b) \leq \theta(V + b) \qquad \text{if and only if } \sigma \geq \sigma^*(b). \tag{5}$$

The principal's payoff if she offers the performance bonus b when her information is c is thus

$$\mathsf{E}_\sigma[U_P] = \theta[1 - G(\sigma^*(b) \mid c)][W - b], \tag{6}$$

which she maximizes over b.

Throughout the paper, we shall ignore "degenerate" equilibria where the principal receives zero regardless of her type because the agent exerts no effort whenever $b < W$. Such equilibria, when they exist, are supported by very pessimistic beliefs that a principal who offers any bonus below W must have really bad information, say, $c = \overline{c}$. Conversely, degenerate equilibria are ruled out when $\theta\,(V + W) > \overline{c}$: by offering a bonus slightly below W, the principal can ensure that the agent works.[11] Let us now denote by B the set of *equilibrium* bonuses; that is, $b \in B$ if and only if b is an equilibrium offer by the principal for some "type" c. Clearly, if b_1 and b_2 both belong to B, with $b_1 < b_2$, then

$$\sigma^*(b_1) > \sigma^*(b_2). \tag{7}$$

If this inequality did not hold the principal could, regardless of her information about c, (weakly) increase the likelihood of effort while offering the lower wage. Therefore, b_2 could not be an equilibrium offer.

Proposition 1. *In equilibrium:*

(i) *Rewards are positive short-term reinforcers: if $b_1 < b_2$, then $\sigma^*(b_1) > \sigma^*(b_2)$.*

(ii) *Rewards are bad news, in that a confident principal offers a lower wage or bonus: if b_1 is a reward offered when the principal knows the task's difficulty to be c_1, and b_2 is offered when she knows it to be $c_2 > c_1$, then $b_2 \geq b_1$.*

(iii) *Rewards undermine the agent's assessment of the task's attractiveness: for all (σ_1, σ_2) and all equilibrium rewards $b_1 < b_2$,*

$$\mathsf{E}[c \mid \sigma_1, b_1] < \mathsf{E}[c \mid \sigma_2, b_2].$$

10. If $E(c \mid 1, b) \geq \theta(V + b)$, one can define $\sigma^*(b) = 1$; if $E[c \mid 0, b] \leq \theta(V + b)$, one can define $\sigma^*(b) = 0$.

11. Degenerate equilibria would also disappear if we assumed that the principal's information is an almost, but not totally, sufficient statistic for the agent's true cost c, in the sense that there is always a very small but positive probability that the agent's signal σ is so favourable that he undertakes the task regardless of the inference drawn from the bonus offer.

Future assessments of task attractiveness are also always reduced by an increase in the reward: the expectation of c conditional on σ, b, the action and the outcome is decreasing in b regardless of σ, the action and the outcome.

Proof. Part (i) has already been established. The proof of part (ii) rests on a standard revealed preference argument. Suppose that b_i is an optimal bonus when the principal has information c_i, $i = 1, 2$, and denote $\sigma_i = \sigma^*(b_i)$. Since b_i is optimal given c_i, it must be that

$$\theta[1 - G(\sigma_i \mid c_i)][W - b_i] \geq \theta[1 - G(\sigma_j \mid c_i)][W - b_j],$$

hence

$$\frac{1 - G(\sigma_1 \mid c_1)}{1 - G(\sigma_2 \mid c_1)} \geq \frac{W - b_2}{W - b_1} \geq \frac{1 - G(\sigma_1 \mid c_2)}{1 - G(\sigma_2 \mid c_2)}. \tag{8}$$

Since $c_2 > c_1$, the MLRP requires that $\sigma_1 \geq \sigma_2$. Hence $b_1 \leq b_2$ since $\sigma^*(\cdot)$ is decreasing. This establishes part (ii) which, in turn, implies that pooling occurs only over intervals.[12] Therefore, if the principal offers b_1 to types $[\underline{c}_1, \bar{c}_1]$ and $b_2 > b_1$ to types $[\underline{c}_2, \bar{c}_2]$, it must be that $\bar{c}_1 \leq \underline{c}_2$. This establishes part (iii) of the proposition. ‖

Proposition 1 demonstrates the main idea of how the trust effect—the principal's expectation of what views the agent is likely to hold—gives rise to a conflict between extrinsic and intrinsic motivation, which in turn shapes the optimal contract.[13] It also has a number of interesting additional implications and extensions.

- *Forbidden fruits*. A higher reward is, in equilibrium, associated with a less attractive task; therefore, bonuses (or higher wages when effort is observable) reduce intrinsic motivation. Conversely, "forbidden fruits" are the most appealing.[14] Indeed, the optimal bonus could well be zero, perhaps even negative. A famous (literary) case is that of Tom Sawyer demanding bribes from other boys to let them paint a fence in his place:

> There was no lack of material; boys happened along every little while; they came to jeer, but remained to whitewash.... And when the middle of the afternoon came, from being a poor poverty-stricken boy in the morning, Tom was literally rolling in wealth. He had a nice, good, idle time all the while—plenty of company—and the fence had three coats of whitewash on it! If he hadn't run out of whitewash he would have bankrupted every boy in the village.[15]

12. There is in fact no pure-strategy separating equilibrium. In such an equilibrium, the agent's behaviour would not depend on his signal. The principal's preference over bonuses that induce compliance with probability 1 will then be the same for all c (choose the lowest one), and so some pooling must necessarily occur.

13. Regarding part (iii), it might be objected that the undermining of the agent's view of the task's attractiveness has no consequence, as he will learn the cost c by doing the task anyway. Note first, however, that the agent may be discouraged from undertaking the task, and therefore not learn. Second, our parameter c may stand for the expected utility cost, while the realized cost also depends on a period-specific shock; the choice of bonus is then informative even when the agent exerts effort. Finally, there may be no learning by doing when the bonus signals the probability of success θ, or the long-run payoff V attached to it, and success is linked to a sequence of investments whose ultimate outcome is observed only with a delay.

14. A simple relabelling of actions extends our results to situations of conflicting interests, where the agent's success is the principal's failure. Suppose for instance that a parent threatens a child with an expected punishment $p \geq 0$ if he smokes. The punishment is (equally) costly to both. Let $W > 0$ and $V \gtrless 0$ denote the parent's and the child's payoffs from his *not* smoking (savings, long-term health benefits) and c the pleasure to be expected from smoking. Confronted with a signal σ positively correlated with c according to the MLRP, the child refrains from smoking if $\mathsf{E}(c \mid \sigma, p) \leq V + p$, which occurs for $\sigma \leq \sigma^*(p)$; the parent chooses p to minimize $G(\sigma^*(p) \mid c))(W + p)$. The same reasoning as earlier shows that a stronger punishment signals a greater concern that smoking is likely to be attractive.

15. Mark Twain, *The Adventures of Tom Sawyer* (1876, Chapter 2).

- *Improper causal attributions*. While our analysis shows that the short-term incentive effect of rewards is reduced by their informational content, it also demonstrates how an outside observer might actually underestimate the power of these incentives. The probability of effort, $1 - G(\sigma^*(b) \mid c)$, and the probability of success, $\theta[1 - G(\sigma^*(b) \mid c)]$, are both decreasing in c, which is known only to the principal. Because c covaries positively with b in equilibrium, the observer who simply correlates b with outcomes may conclude that rewards are negative reinforcers *even in the short run*. The reason is that such unconditional correlations or regressions fail to take into account the fact that a principal seeking to induce compliance offers the highest incentives to the agents who would otherwise be the least likely to work.[16]
- *Robustness*. We have so far assumed that the bilateral relationship was not at stake, and have therefore ignored the agent's participation constraint. Let $\overline{U}$ denote his outside reservation utility, which we assume to be independent of the attractiveness of the task at hand. If $\theta(V + b) - E(c \mid \sigma^*(b), b) \geq \overline{U}$, the participation constraint is not binding. Otherwise, $\sigma^*(b)$ must be replaced by $\max\{\sigma^*(b), \sigma^{**}(b)\}$, where $\sigma^{**}(b)$ is defined by $\theta(V + b) - E(c \mid \sigma^{**}(b), b) = \overline{U}$. The proof of Proposition 1 is otherwise unaltered, and the conclusions therefore unchanged.
- *Immediate re-engagement effects*. The re-engagement effect may occur even if the agent does not undertake the same task repeatedly. First, the information conveyed by incentives on one task (say, math homework) spills over to correlated tasks (physics homework). Second, and more interestingly, rewards may have an *immediate* negative impact when performance measurement is state-contingent. Consider the same model as above, where the principal can initially threaten to punish the agent in case of poor performance or bad behaviour, but let the effectiveness of her monitoring technology now fluctuate randomly. The agent learns, before making his decision, whether he is likely to be caught if he misbehaves, or to escape detection. The threat of punishment then has a positive (short term) reinforcement effect in instances when the agent knows that monitoring is effective, but only a negative one (the analogue of a re-engagement effect) when he thinks that he can "get away with it". A familiar case is a teenager's heightened temptation to violate his (her) parents' strict prohibition on smoking, in situations where they cannot catch him (her).

2.2. *Self-confidence: trust and profitability effects*

When the principal has private information about the agent's *ability* θ rather than the cost of implementing the task, a new effect may enter into the agent's inference process. As we shall see, this profitability effect, when it is present (this will depend on the type of contract allowed), works here in the same direction as the trust effect.

We now assume that c and V are common knowledge, whereas both parties are differentially informed about the agent's probability of success θ, which is drawn from a distribution $F(\theta)$ with density $f(\theta)$ on $[\underline{\theta}, \overline{\theta}]$. The principal observes θ exactly, whereas the agent only receives an imperfect signal $\sigma \in [0, 1]$, with conditional distribution $G(\sigma \mid \theta)$ and density $g(\sigma \mid \theta)$ satisfying the MLRP—just as in (4), but with a higher σ now signalling a higher θ. The agent's effort is unobservable to the principal, who therefore conditions the bonus b on a successful performance.

16. In the two-type example developed in what follows (Proposition 3) for instance, the observer will see the agent working with positive probability (perhaps even a high probability) even when no reward is offered. From this he might be led to infer that rewards do not make much of a difference, and could thus perhaps be reduced or done away with. This would be a mistake, because in situations where the reward is actually given, it does have a significant impact on performance.

2.2.1. No lump-sum payment. As in the previous section, we first restrict attention to the case where contracts offered by the principal do not involve any non-contingent transfers (lump-sum payments) in either direction. There is then still no profitability effect: in the principal's objective function $U_P = \theta e(W - b)$, the marginal rate of substitution between b and e is independent of θ. As a result, Proposition 1 carries over with a mere change of notation:[17]

Proposition 2. *All the results in Proposition* 1 *apply* (*with the appropriate changes in notation and terminology*) *when the principal's private information and the agent's noisy signal bear on the agent's probability of success* θ *rather than on the task's difficulty* c.

Let us now further specialize the model by assuming that θ can take only two values, θ_H and $\theta_L < \theta_H$, with associated conditional densities for the agent's signal $g_H(\sigma)$ and $g_L(\sigma)$. The MLRP means that g_H/g_L is increasing. We also assume that rewards cannot be negative, $b \geq 0$,[18] and that if we denote by b_k^*, $k \in \{L, H\}$, the minimum feasible bonus that induces effort when the agent is fully informed about his ability,

$$b_k^* = \max\left\{0, \frac{c}{\theta_k} - V\right\}, \tag{9}$$

then $0 = b_H^* < b_L^* < W$. The agent's reservation utility will, without loss of generality, be normalized to zero. We shall refer to this combination of assumptions as the "two-type case". It allows for a more explicit version of Proposition 2, both characterizing all perfect Bayesian equilibria and identifying a unique refined equilibrium; the refinement used here is Cho and Kreps' (1987) version of "Never a Weak Best Response" (NWBR).[19]

Proposition 3. *In the two-type case, where* g_H/g_L *has full support* $(0, +\infty)$*:*

(i) *In any equilibrium, the principal offers a low bonus* $b < b_L^*$ *to a more able agent* $(\theta = \theta_H)$*, and randomizes between the bonuses* b *and* b_L^* *when dealing with a less able agent* $(\theta = \theta_L)$.

(ii) *There is a unique NWBR-refined equilibrium, and it is such that* $b = 0$*. The probability of pooling* (*offering* $b = 0$ *when* $\theta = \theta_L$)*,* $x^* > 0$*, and the unconditional probability of no bonus,* $f_H + f_L x^*$*, both increase with the agent's initial self-confidence,* f_H*. The trust effect thus forces the principal to adopt low-powered incentives, and the more so the more self-confident the agent is.*

Proof. See the Appendix. ‖

2.2.2. Lump-sum payments. We have so far ruled out lump-sum payments. In some applications, these might indeed not be feasible—*e.g.* when one of the two parties has no cash, or is protected by limited liability. Note also that any equilibrium outcome in the absence of lump-sum payments is still an equilibrium when they are allowed (sustained by out-of-equilibrium beliefs that such transfers convey no information).

17. As before, "degenerate" equilibria, in which the principal receives a zero payoff regardless of her type, are ruled out—for instance by assuming that $\underline{\theta}(V + W) > c$.

18. Here again, imposing $b \geq 0$ only reduces the scope for signalling in equilibrium, but leaves all the results qualitatively unaffected. See also footnote 9 for reasons why $b < 0$ may just not be feasible.

19. See, *e.g.* Fudenberg and Tirole (1991, p. 454) for a formal definition. As explained there, NWBR is somewhat stronger than Cho and Kreps' "intuitive criterion".

In many important cases, however, a more general menu of contracts is feasible, allowing the principal to offer a (positive or negative) up-front wage a, together with a bonus b in case of success.[20] Under symmetric information, the lump-sum transfer only enables the principal to tax the (high-ability) agent for the rent he derives from the activity. In a situation of private information, in contrast, the principal can use it to signal that she knows the agent to have a high probability of success. The intuition is akin to "burning money", in an amount that would wipe out any profits to be expected by inducing a low-ability agent to work, but would still leave the principal with a positive surplus if a high-ability agent undertook the task. As we shall see below, this is a form of what we termed the profitability effect.

There are, however, many equilibria in the multidimensional signalling game where a contract is a pair (a, b). We shall not attempt a complete analysis of the (potentially very large) equilibrium set, but rather limit ourselves to the two-type framework, which yields the key insights. Suppose therefore, as above, that the agent's ability can be either high or low, $\theta \in \{\theta_H, \theta_L\}$, and denote respectively by $G_H(\sigma)$ and $G_L(\sigma)$ the cumulative distribution functions of his signal σ in each case. We shall assume that

$$\theta_H G_H(\sigma) > \theta_L G_L(\sigma), \qquad \text{for all } \sigma > 0. \tag{10}$$

This is essentially a limited informativeness condition, requiring that the signal's distribution does not vary too much with the underlying state. Its differentiable version, $-\partial \ln G(\sigma \mid \theta)/\partial \ln \theta < 1$, states that the elasticity of non-participation with respect to θ of an agent using any given cutoff rule σ must be less than one.[21]

Proposition 4. *In the two-type case, with limited informativeness, there is a unique separating perfect Bayesian equilibrium that satisfies the NWBR criterion. A principal who observes the agent to be of ability θ_k, $k \in \{L, H\}$, offers the contract (a_k, b_k), with $b_k = b_k^*$ and $a_L = 0 < c - \theta_L V = a_H$. The principal's expected utility is $U_P^L \equiv \theta_L(V + W) - c$ when $\theta = \theta_L$, and $U_P^H \equiv \theta_L V + \theta_H W - c$ when $\theta = \theta_H$. The agent's expected utility in each case is, respectively, 0 and $(\theta_H - \theta_L)V$.*

Proof. See the Appendix. ‖

Thus, even with this more general class of contracts, it is still the case that a more high-powered incentive scheme—that is, a higher b and a lower a:

(i) is a positive reinforcer in the short-run, since it leads (low θ) agents to exert effort who otherwise would not have done so;
(ii) is bad news for the agent and permanently damages his self-confidence, no matter what the task's outcome turns out to be.

Souvorov (2003) shows that similar conclusions hold for all NWBR equilibria: under (10), the high type offers bonus $b_H^* = 0$ and the low type weakly mixes between b_H^* and b_L^*. Overall, these results highlight the workings of the profitability effect that comes into play when lump-sum payments (as opposed to conditional bonuses) are allowed: whereas $-(\partial U_P/\partial b)/(\partial U_P/\partial e) = e/(W - b)$ is independent of θ due to the multiplicative form of expected output, $-(\partial U_P/\partial a)/(\partial U_P/\partial e) = 1/(\theta(W - b))$ is decreasing in θ, meaning that a lump-sum transfer is an investment (in signalling) that has a higher rate of return when the agent

20. We thank an anonymous referee for prompting us to analyse the lump-sum payment case, and to more carefully discriminate between the trust and profitability effects.

21. It is, for example, satisfied by $G(\sigma \mid \theta) = 1 - e^{-\frac{\sigma}{\theta(1-\sigma)}}$ for all $\sigma \in [0, 1]$.

is talented.[22] The complete specification of the contract will depend on the particular equilibrium that is played (*e.g.* contrast those of Propositions 3 and 4), but under the standard refinement used here, the basic result of a less performance-dependent compensation scheme for more able agents remains verified.

There are also some differences with the no-lump-sum-transfer case. First, bonuses are the same as under symmetric information, so in that sense, the informed-principal game leads to no distortion of incentives. Second, because of the fixed wage, the (high-ability) agent's utility is now higher than under symmetric information. Thus, the general weakening of performance-based compensation which is the model's main insight now takes the form of a lower *share* of contingent compensation in total compensation. In particular, it still implies that the distribution of rewards across a given population of agents will be *more equal*, in a Lorenz sense, due to the confidence-management problem.

It is worth commenting here on the *uses and limits of money-burning strategies*. Proposition 4 unveils an important new dimension of confidence management, and indeed, one often observes principals burning resources in such an effort–spending for instance considerable time trying to "convince" agents to attempt challenging tasks by giving them pep talks, encouragements, and similar unverifiable, soft information. At the same time, the money- or time-burning strategy can only be part of the overall story, for several reasons. First, the profitability effect to which it gives rise when the principal's private information is about θ does not occur when it is about the task's attractiveness (the cost c, or the ultimate reward V). As a result, one can show that the no-lump-sum equilibrium (the analogue to Proposition 3 for unknown c or V) remains a NWBR equilibrium when such transfers are allowed, provided the cutoff σ^* is not too high. Second, burning money is a non-contingent reward. In contrast, the psychology literature on intrinsic motivation mostly emphasizes the effects of contingent rewards, and the leading experimental case is that of a task of unknown attractiveness.[23] Yet another limitation of lump-sum transfers in a broader framework is that they tend to attract "undesirable types". This would be the case for example in our model if a fraction of the population were lazy—had a large cost of effort c—and if the principal was unable to distinguish between lazy and diligent types.[24]

In summary, our analysis distinguishing between the trust and profitability effects makes clear that low-powered incentives and burning money are two ways in which the principal's confidence-management motive will be reflected in equilibrium contracts—each with its own domain of applicability, but with similar effect on wage inequality and long-run motivation.

2.3. *Back to the debate*

(a) *Relation to the psychology literature*

Let us now return to the hidden cost of rewards. Our approach, in the tradition of economics and cognitive psychology, focuses on the individual's beliefs and motivation. An alternative

22. Correspondingly, the implementability condition (see footnote 6) for the two-dimensional policy $p = (a, b)$ reduces to its component in a, which requires that $da/d\theta \geq 0$ in equilibrium. Another simple way to see that the role of lump-sum transfers here is to create a profitability effect is to observe that Proposition 4 applies unchanged when the agent does not receive any private signal ($G_H \equiv G_L$), in which case we saw in Section 1 that there can be no pure trust effect.

23. The experimental literature has also examined the effects of unconditional rewards on intrinsic motivation. Typically, none is found (*e.g.* Deci *et al.*, 1999). In these setups, however, the experimenter does not selectively give the lump sum to some subjects and not to others (a randomization is used, and subjects may not even be aware of differences in treatment). It is also generally not one where the experimenter can be thought to have a stake in the subject's performance, so he would again have no incentive to "burn money" to boost his motivation.

24. Another potential limitation on the use of lump-sum transfers (at least large ones) is that when agents are risk-averse, giving them a substantial lump-sum payment reduces their marginal utility of income, and can thus actually make it harder to motivate them to work.

viewpoint, along the lines of the behaviourist school (Hull (1943), Skinner (1953)) would shun the inner process and posit a direct link from stimulus to response. The agent would then just be assumed to exhibit an instinctive, aversive reaction to being offered a contingent reward, or threatened with a punishment. Not surprisingly, we are not inclined to adopt such a "reduced form" approach. While individuals do not really compute perfect Bayesian equilibria when interpreting signals from their environment, there is a lot of evidence that they are quite sophisticated at (and definitely intent on) getting to the motives behind the words and deeds of the people with whom they interact.[25] Moreover, in the present context the information-based approach delivers two important benefits. First, it helps understand *why* the response to the stimulus is what it is. Second, it generates testable predictions as to *when* rewards may indeed have real costs, and when this view is likely to be a myth.

Thus, we have identified a class of interactions between a privately informed principal and an agent who attempts to infer her motives from the type of contracts offered to him, that have the following implications:

- Rewards impact intrinsic motivation. Whereas under symmetric information the intrinsic ($\theta V - c$) and extrinsic (θb) motivations can be cleanly separated, under asymmetric information they cannot. When the agent is unsure about his ability, the intrinsic motivation $\hat{\theta}(\sigma, b)V - c$ decreases with the level of the bonus. Similarly, when he does not know how costly or exciting the task is, his perception of it, $\hat{c}(\sigma, b)$, is affected by the level of the wage or reward.
- A reward is a positive reinforcer in the short term, but always decreases future motivation.

These conclusions, as well as the underlying mechanisms, are well in line with an important branch of the social psychology literature. Indeed, the standard references on the hidden costs of rewards (Lepper *et al.* (1973), Deci (1975), Deci and Ryan (1985)) are based on self-perception and attribution theories, according to which individuals constantly reassess the reasons for their and others' behaviour. Both approaches emphasize the informational impact of rewards. As Deci (1975, p. 42) argues:

> Every reward (including feedback) has two aspects, a controlling aspect and an informational aspect which provides the recipient with information about his competence and self-determination.

Both views also stress the re-engagement effects of rewards. Thus Schwartz (1990), commenting on Lepper *et al.* (1973), argues:

> Reinforcement has two effects. First, predictably it gains control of [an] activity, increasing its frequency. Second, ... when reinforcement is later withdrawn, people engage in the activity even less than they did before reinforcement was introduced.

The tension between the short-term and long-term effects on motivation of offering a reward also suggests the following idea: once a reward is offered, it will be required—and "expected"—every time the task has to be performed again—perhaps even in increasing amounts.[26] In other words, through their effect on self-confidence, rewards have a *"ratchet effect"*. This irreversibility may explain people's (*e.g.* parent's) reluctance to offer them, even on occasions where they

25. With, naturally, some variations in the population. For example, adults usually have more experience in interpreting social signals than children, and the latter themselves exhibit different speeds of learning these skills.

26. The same task, or related tasks: see Frey, Fehr and Benz (2000) for experimental evidence showing that the crowding out of motivation may extend beyond the area of intervention.

would seem like a small price to pay to get the current job done. Souvorov (2003) analyses this question using a two-period extension of the present model; an additional effect arises, namely that the agent now has a strategic incentive to appear demotivated (having a low σ), in order to be given a higher bonus in the future. Souvorov establishes three monotonicity properties. First, in each period a low type is offered a (weakly) higher bonus, meaning that Propositions 1–3 hold in a dynamic context as well. Second, for a given type, the bonus is (weakly) increasing over time, validating our earlier conjecture. Finally, for each given type, the initial bonus is lower when the principal is the same in both periods than if there are two different principals; the reason is that a long-lasting principal internalizes the fact that rewards are habit-forming.

Our results are also consistent with Etzioni's (1971) claim that workers find control of their behaviour via incentives "alienating" and "dehumanizing", with Kohn's (1993) argument that incentive schemes make people less enthusiastic about their behaviour, and with Deci and Ryan's 1985 view that rewards change the locus of causality from internal to external, making employees bored, alienated and reactive rather than proactive.

(b) *Promised vs.* ex post *rewards*

Our analysis can also help clarify the difference between what we would label "promised" or "*ex ante*" contingent rewards, and "discretionary" or "*ex post*" rewards. Our model is about the *control* of behaviour through rewards: the principal selects a reward for a well-defined effort or performance *before* the agent's decision. The agent then rationally interprets the reward scheme as a signal of distrust or of a boring task.

By contrast, rewards that are discretionary (not contracted for) may well boost the agent's self-esteem or intrinsic motivation, because of a different learning effect: the worker or child learns from the reward that the task was considered difficult (and therefore that he is talented), or that the supervisor or parent is appreciative of, proud of, or cares about his performance—and therefore that it is worth repeating it. Giving *ex post* a bicycle to a hard-working child, or a special pay rise or early promotion to a productive assistant professor will not lead him to infer that his behaviour was controlled, because the principal was under no obligation (no commitment) to reward any particular outcome. And receiving the reward is good news, because the agent initially did not know how to interpret his performance. The reward then provides the agent with an indirect measure of his performance.[27]

(c) *When does extrinsic motivation undermine intrinsic motivation?*

The next point is alluded to in Deci (1975, p. 41):

> If a person's feelings of competence and self-determination are enhanced, his intrinsic motivation will increase. If his feelings of competence and self-determination are diminished, his intrinsic motivation will decrease . . . . We are suggesting that some rewards or feedback will increase intrinsic motivation through this process and others will decrease it, either through this process or through the change in perceived locus of causality process.

27. Consider an infinite-horizon extension of our model where, at the end of each period, agent and principal both observe whether the agent's effort was successful (which happens with a known probability $\theta \leq 1$), but only the principal knows whether the long-run payoffs that will ultimately be reaped from this success are high or low (*i.e.* $(W, V) \in \{(W_H, V_H), (W_L, V_L)\}$, with $W_H > W_L$ and $V_H > V_L$; this positive correlation may, but need not, reflect altruism). One can think of a parent and a child whose real payoffs to success on exams or test will come much later in life, in the form of a better career. There is then a natural (Bayesian perfect) reputational equilibrium where the principal gives a reward b^* in case of success *only* to high-payoff agents (hence rewards are good news); conversely, the agent works only if all previous successes were rewarded. The reward b^* is determined by the principal's incentive compatibility condition, namely $b^* = \delta(W_L - b^*)/(1-\delta) < \delta(W_H - b^*)/(1-\delta)$, where δ denotes her discount rate. The agent's hypothesized behaviour is optimal provided $V_L < c < V_H + b$.

Our economic analysis indeed unveils important necessary conditions for rewards to have a negative impact on self-confidence. The first is that the principal has *information* about the agent or the task that the agent does not. This may explain why the existence of hidden costs of rewards is less controversial in educational settings than in the workplace. Children have particularly imperfect knowledge of their selves and of their aptitudes in the quickly changing tasks which they face as they grow up (curriculum, sports, social interactions, etc.). In contrast, the structure of rewards in the workplace is often more anonymous: in most sectors, it is the same for all workers with the same "job description". The terms of this (contingent) contract still reflect information about the nature of the job, but much of it may already be publicly known.

The second key condition is the *sorting condition:* for rewards to signal a low ability or a boring task, it must be that the principal is comparatively more tempted to offer performance incentives under those circumstances. Conversely, consider the case of a manager who is promoted from a fixed-salary job and given the leadership of a new project or division, together with a pay-for-performance scheme. In this example (which is related to Section 3.1 below), the sorting condition works in the opposite direction: the contingent reward is associated with a *high* level of trust from the principal, demonstrated by a large "empowerment" effect, and should therefore boost the manager's self-confidence.

Another example in which the sorting condition works in the opposite direction, and rewards are long-term positive reinforcers, is when a task is subject to learning by doing and learning is more effective for a talented agent. By offering a reward, the principal is then really saying: "I know that you are talented. Encouraging you to try would make no sense if you were unable to learn by doing". That the sorting condition is needed in order for the principal to boost the agent's self-confidence is also demonstrated by the standard observations that the use of compliments to ingratiate oneself with a person may backfire, that parents often have a hard time motivating their children to work at school by telling them about their ability (θ), the rewards from education (V), and the pleasure of learning (c); and that depressed individuals often attribute ulterior motivation to those who try and comfort them.[28]

To sum up, before worrying about the negative impact of rewards, one should first check that the reward provider has private information about the task or the agent's talent (including as we have noted, a greater ability to interpret the agent's track record). One should then, as the agent does, think through the provider's ulterior motivation and how her payoff from giving a contingent reward is affected by her knowledge.

(d) *Retrospective justification and self-perception*

The same reasoning that held for inferences about the task's difficulty c obviously applies to the agent's payoff V from succeeding in it. Combined with imperfect memory, this result has an interesting implication for situations where currently available information provides only insufficient justification for a certain course of action.[29] Suppose that, at some later date, the agent again faces the choice of whether to undertake the same or a similar task; and that, come that time, he remembers only that he chose to engage in it, and the extrinsic incentives that were then offered,

28. It would be interesting to assess in this light the evidence on the role of expectations. For example, teachers with initially over-optimistic expectations about their students lead to changes in the performances of the students which tend to confirm the expectations (Rosenthal and Jacobson (1968); see also Merton (1948) for a discussion of self-fulfilling prophecies). It seems, however, that while the students' behaviour changes, their self-confidence is unaffected (Darley and Fazio, 1980).

29. See Bénabou and Tirole (2002*a*) for a model of endogenously selective memory or awareness. The present argument requires only that memory be imperfect, especially with regard to one's past feelings and emotions (hedonic payoffs). For a framework where agents with imperfect recall make retrospective inferences about their own preferences from their past choices, see Bénabou and Tirole (2002*b*).

but not his intrinsic interest in the task (and the later observation of V). For instance, an individual engaged in a long-term project—writing a book, proving a theorem, running a marathon—may, at times, be seized by doubt as to whether the intellectual and ego-gratification benefits which successful completion is likely to bring will, ultimately, justify the required efforts. ("Why am I doing this?") He may then reflect that since he chose to embark on this project once again in spite of low financial and career incentives, the personal satisfaction enjoyed from previous completions (and which, at this later and perhaps somewhat stressful stage, he cannot quite recall) must have been significant. Hence it is worth persevering on the chosen path. The result that $\mathsf{E}[V \mid \sigma, b] > \mathsf{E}[V \mid b']$ for $b < b'$ can thus provide a formal explanation for this kind of *ex post* rationalization of one's *own* choices (Festinger and Carlsmith (1959), Bem (1967), Staw (1977)).

(e) *Paternalism: altruism towards a time-inconsistent agent*

Another interesting class of situations for which our framework is relevant arises when an agent (child or adult) has time-inconsistent preferences, generating a divergence between *his own* short- and long-run interests. As a result of this "salience of the present", he may for instance shirk on homework or professional duties, fail to stick to a necessary diet or exercise regimen, or remain addicted to tobacco, drugs or alcohol. A well intentioned principal—parent or close friend—who takes the *long run* view of the agent's welfare will then have the exact same incentives as those we analyse here to manipulate the agent's perceptions of himself and of the tasks he faces—"for his own good".[30]

3. OTHER CONFIDENCE-ENHANCEMENT STRATEGIES

3.1. *Empowerment and motivation*

Section 2 showed that the principal may signal the agent's ability (high θ), the attractiveness of the task (low c) or its long-term payoff (high V) through the use of a low-powered incentive scheme. In the same spirit, we now investigate the use of delegation or empowerment to induce an agent to carry out the objectives of the principal (Miles, 1965). Intuitively, the principal demonstrates her confidence in the agent's ability (or, more generally, his intrinsic motivation) by delegating control of the task to him. This, in turn, makes it more likely that the agent exerts effort. The delegation vs. supervision problem is also interesting because it provides an example where the profitability and trust effects work in opposite directions, in contrast to the previous section where they reinforced each other.[31]

We shall abstract here from the explicit rewards that were our earlier focus ($a = b = 0$).[32] Let $\mathcal{W}_1(\theta)$ and $\mathcal{W}_0(\theta)$ denote the principal's expected payoff when she delegates ($d = 1$) and does not delegate ($d = 0$) to an agent with ability θ, and the agent exerts effort. The principal

30. Formally, W in this case is equal to V/β, where $\beta < 1$ is the agent's quasi-hyperbolic discount factor; $1/\beta$ measures the salience of the effort cost c for the agent, at the time when he must incur it.

31. The analysis given here is not based on the initiative effect studied in Aghion and Tirole (1997). There, an agent invests more in the acquisition of information about potential projects if he knows that the principal will not interfere too much with his suggestions. In Dessein (2002), the principal delegates so as to ensure that the decision taken will better reflect the agent's information than when the latter communicates it strategically and the principal decides. By delegating to the agent, the principal signals a greater *congruence* of their objectives. Salancik (1977) proposes yet another viewpoint, namely the "co-optation of personal satisfaction", related to our earlier discussion of retrospective justification: *"By having a person choose to do something, you create a situation that makes it more difficult for him to say that he didn't want to do it. And the ironic thing is that the more freedom you give him to make the decision, the more constraining you make his subsequent situation."*

32. In footnote 35, we will sketch how the results extend to contracts with lump sum or fixed wage payments $a \neq 0$. We shall continue to abstract from contingent rewards, however, as these were the main focus of the previous section. The constraint $b = 0$ could reflect, for instance, the fact that performance is not publicly verifiable.

receives 0 if the agent does not try. As shown below, these reduced forms can be derived from a situation where the principal decides to either relinquish some control rights to the agent, or put in place a supervisor or monitoring technology. As earlier, we assume that the principal knows the agent's probability of success θ, while the latter receives a signal σ drawn from a cumulative distribution $G(\sigma \mid \theta)$, with density $g(\sigma \mid \theta)$ satisfying the MLRP. The agent's utility is, as usual, $\theta V - c_d$ for $d \in \{0, 1\}$ if he exerts effort, and 0 otherwise. We assume that $c_1 \leq c_0$: *ceteris paribus*, the agent prefers delegation. The timing is as follows. At stage 1, the principal selects $d \in \{0, 1\}$. At stage 2, the agent decides whether to undertake the task; the principal's payoff is $\mathcal{W}_d(\theta)$ if he does, and 0 otherwise.

Assumption 1. For all $\theta \in [0, 1]$,

$$\frac{d}{d\theta}\left(\frac{\mathcal{W}_1(\theta)}{\mathcal{W}_0(\theta)}\right) > 0; \qquad \text{moreover,} \qquad \frac{\mathcal{W}_1(0)}{\mathcal{W}_0(0)} < 1 < \frac{\mathcal{W}_1(1)}{\mathcal{W}_0(1)}.$$

In words, an empowered agent is less likely to create damage to the principal when he is talented than when he is not. Furthermore, the principal does not want (*ceteris paribus*) to delegate the task to an inept agent ($\theta = 0$), and prefers to delegate the task to a very talented one ($\theta = 1$). This implies that there exists a θ^* in (0, 1) such that, under symmetric information, it is efficient to delegate if $\theta > \theta^*$, and to monitor if $\theta < \theta^*$.

- *Example*: Suppose that the agent, when paying a cost $c_1 = c_0 = c$, comes up with a project. The project, in its initial form, will succeed if the agent is "good", and fail if he is "bad". The principal knows the probability θ that the agent is good. If the project is unmodified by the principal and is successful, the agent receives new job offers, with value V to him; he receives no such offer if the project fails, is modified, or if he does not even try. Success also yields a monetary payoff W to the principal. "Delegation" means transferring the control right to the agent, who will then implement his project without modification, resulting in expected payoffs $\theta V - c$ for himself and θW for the principal (both get 0 if the agent does not try). Alternatively, the principal may keep the control rights and pay a fixed monitoring cost C to supervise the agent's project. This enables her to discover along the way (and with some probability) if it is headed for failure, and to then modify it so as to make it successful. We assume that the principal is able, with probability x, to turn a failing project into a successful one, where $xW > C$. Payoffs are then still $\theta V - c$ for the agent (he gets no credit for a project modified by the principal since it would have failed otherwise), but now $(\theta + (1 - \theta)x)W - C$ for the principal. Thus

$$\mathcal{W}_d = [\theta + (1 - \theta)x(1 - d)]W - C(1 - d)$$

for $d \in \{0, 1\}$, and therefore Assumption 1 is satisfied, since

$$\frac{\mathcal{W}_1(\theta)}{\mathcal{W}_0(\theta)} = \frac{1}{1 - x + \frac{x - C/W}{\theta}}.$$

Assumption 1 corresponds to what we termed a profitability effect, which here pushes the principal towards giving greater autonomy to more able agents. The trust effect, on the other hand, works in the opposite direction: when σ is highly correlated with θ, the principal is very concerned about boosting the motivation of low-ability agents (offsetting the bad signals which they are likely to receive). Thus, if delegation is thought to be reserved for high-ability agents, she may want to in fact give it to some very low ability individuals as well. Moreover, when $c_1 < c_0$ delegation also involves an implicit "reward", in that it makes the task more

pleasant to perform. Of course, in equilibrium the looking-glass-self principle will operate, and agents will (on average) correctly infer the principal's motivations in delegating or not delegating to them (*e.g.* rewards are bad news, as we showed earlier). Due to the conflict between the profitability and trust effects, however, Assumption 1 is generally not sufficient to make the informational content of delegation unambiguous, even though its effect on the agent's *effort* will be unequivocal. This additional result will require a second assumption:

Assumption 2. For all (σ_0, σ_1) with $\sigma_0 > \sigma_1$, the elasticity of the odds ratio $(1 - G(\sigma_0 \mid \theta))/(1 - G(\sigma_1 \mid \theta))$ with respect to θ is less than that of $W_1(\theta)/W_0(\theta)$.

This limited-informativeness condition imposes an upper bound on the trust effect arising from the correlation between the signals θ and σ received by the principal and the agent.[33] It implies that the profitability effect dominates in the agent's inference problem.

Proposition 5. *In equilibrium, under Assumption* 1*:*

(i) *Empowerment always increases the probability that the agent will exert effort (no matter what his type θ is).*
(ii) *There is more empowerment than under symmetric information: there exists a $\theta^{**} < \theta^*$ such that the principal delegates whenever $\theta > \theta^{**}$.*
(iii) *Suppose that Assumption* 2 *holds as well. Empowerment is then always good news for the agent about his ability, and permanently changes his attitude towards the task: for any signal σ, $E[\theta \mid \sigma, d = 1] > E[\theta \mid \sigma, d = 0]$.*

Proof. See the Appendix. ‖

Proposition 5 and its premises are consistent with Pfeffer's (1994) observation that:

> When employees are subjected to close external monitoring or surveillance, they may draw the psychological inference that they are not trusted and thus not trustworthy, acting in ways that reinforce this perception.[34]

Note, finally, that while we have abstracted here from the use of fixed payments by the principal to signal her trust in the agent ("burning money"), the main insights and results conveyed by Proposition 5 are, once again, robust to allowing for lump-sum transfers.[35]

3.2. *Help*

Still assuming that the agent is unsure about his ability, suppose that the principal offers to contribute a level of help h (at private cost h) in case the agent decides to undertake the

33. For example, with the conditional distribution $G(\sigma \mid \theta) = 1 - e^{(\frac{\sigma}{\sigma-1})/(\theta+k)}$ for $\sigma \in [0, 1]$, Assumption 2 amounts to imposing a lower bound on k.

34. Cited in Baron and Kreps (1999), who provide an illustration at Hewlett-Packard.

35. Suppose that there are two types, θ_H and θ_L, with frequencies f_H and f_L respectively, such that $W_1(\theta_L)/W_0(\theta_L) < 1 < W_1(\theta_H)/W_0(\theta_H)$ and $\theta_H V - c > 0 > \theta_L V - c$. Performance is non-verifiable, so the contract can only specify a fixed payment $a \geq 0$. If f_H is high enough that $(f_H\theta_H + f_L\theta_L)V - c > 0$, then $\{a = 0, d = 1\}$ constitutes a pooling equilibrium, in which the agent exerts effort. Since both types of principal then receive their maximum possible payoff, no intuitive-criterion type of reasoning could upset this equilibrium (we believe that it is in fact the only Cho–Kreps robust equilibrium in this case). When $(f_H\,\theta_H + f_L\theta_L)V - c < 0$, on the other hand, the Cho–Kreps robust equilibrium will involve separation or semi-separation via money burning, provided that $W_1(\theta_L) \leq W_1(\theta_H)$, so that one can find an $a_H > 0 = a_L$ such that $W_1(\theta_L) - a_H \leq W_1(\theta_H) - a_H$.

task. This help improves the probability of success, which is thus a function $P(\theta, h)$ with $P_\theta > 0$ and $P_h > 0$. The agent then undertakes the project if and only if $\sigma \geq \sigma^*(h)$, where $\mathsf{E}[P(\theta, h) \mid \sigma^*(h), h]V = c$, and $\sigma^*(h)$ is a decreasing function. Ignoring rewards, the principal's payoff is

$$U_P = [1 - G(\sigma^*(h) \mid \theta)][P(\theta, h)W - h]. \tag{11}$$

The term in the second bracket is her expected payoff conditional on the agent's undertaking the task. Let us assume that the percentage increase in that payoff achieved by a higher level of help (the expected rate of return on investing in help) is smaller when the agent is talented than when he is untalented:

Assumption 3. For all θ and h, $\partial^2 \ln(P(\theta, h)W - h)/\partial\theta\partial h < 0$.

In other words, help makes more of a difference for weak agents than for strong ones.[36] We shall also use a limited-informativeness condition, similar to Assumption 2 earlier.

Assumption 4. For all (σ_0, σ_1) with $\sigma_1 > \sigma_0$ and (h_0, h_1) with $h_1 > h_0$, the elasticity of the odds ratio $(1 - G(\sigma_0 \mid \theta))/(1 - G(\sigma_1 \mid \theta))$ with respect to θ is less than that of $(P(\theta, h_1)W - h_1)/(P(\theta, h_0)W - h_0)$.

Following the steps in the proof of Proposition 5, one easily shows:

Proposition 6. *In equilibrium:*

(i) *Under Assumption* 3, *giving more help always decreases the probability that the agent exerts effort* (*his action threshold* $\sigma^*(h)$ *is increasing in* h).
(ii) *When Assumption* 4 *also holds, a high level of help is always bad news for the agent, permanently weakening his self-confidence with respect to his ability for the task.*

Proposition 6 may explain why help, like rewards or lack of delegation, can be detrimental to self-confidence. For example, depression, a recognized disorder of self-esteem (Bibring (1953)), is relatively common among individuals with "dependent" personality patterns—that is, individuals with backgrounds characterized by pampering and overprotection (Snyder, Higgins and Stucky, 1983, p. 233). Similarly, Gilbert and Silvera (1996) observe that a parent who finds dependence of his or her child gratifying may provide unnecessary assistance.

A sorting condition like the one assumed above seems quite appropriate when task performance is of a *zero–one* nature: graduating high school or passing an exam, getting a job or keeping it, etc. In other situations the sorting condition may be reversed, so that receiving help is a positive signal. This is likely to occur when the principal's payoff in case of success rises with the agents' ability, or with the level of help which was provided (a more helping principal gets more "credit").[37] One can think of situations such as joining a start-up firm, or contributing time and money to a political party or candidate. The two types of sorting conditions can be illustrated by the contrast between the case of a professor helping a student write a term paper or getting his or her thesis done (the professor's payoff is largely independent of the margin of success with which the student passes the hurdle), and that where the same professor coauthors a research

36. This is again a profitability effect: since $-(\partial U_P/\partial h)/(\partial U_P/\partial e) = -e(\partial \ln(PW - h)/\partial h)$, Assumption 3 implies the standard sorting condition.

37. Formally, replacing W by $W(\theta)$ or $W(h)$ in the expected payoff $P(\theta, h)W - h$ (with $W' > 0$) tends to reverse the sorting condition, by generating a complementarity between θ and h.

paper with the student or with a younger faculty member (helping is then more attractive, the better the prospects for the paper's success due to the coauthor's talent).

3.3. *Coaching*

The use of encouragement, praise, strategies to minimize the effect of failures and the like, is a central theme in human resource management and education.[38] Successful coaches are viewed as those who build up others' confidence (Kinlaw, 1997).

(a) *Encouragement*

The usual complementarity between effort and talent makes it clear why even a selfish coach may gain by building up the agent's self-esteem. (Section 4 will nonetheless identify settings where coaches may have the reverse incentives, and bash agents' egos.) Formally, the principal's policy p here is the disclosure (or absence of disclosure) to the agent of hard private information about his ability. The release of a signal covarying positively (negatively) with θ boosts (lowers) the agent's self-confidence.

(b) *Praise, criticism and excuses*

Let us now turn to the coach's *ex post* assessment of the agent's performance. This assessment exercise is of course still forward looking, in that it is meant to improve the agent's future performances. Taking it for granted that the principal wants to boost the agent's self-esteem, it is interesting to note that, in some circumstances, reassurance can nonetheless have ambiguous consequences. Suppose that the agent failed. The principal may then try to convince him that the link from talent and effort to performance is rather random ("the jury was incompetent") or, relatedly, that the agent was discriminated against (*e.g.* racial prejudice by employers or educators, methodological bias by referees). Offering such excuses may sometimes prove self-defeating.[39] Indeed, if the noise affecting the past performance is recurrent (*e.g.* the agent is likely to be discriminated again in the future if he has been in the past), then the excuse may discourage rather than encourage him. To illustrate this idea, suppose that there are two periods, $t = 1, 2$. The agent's payoff in period t is

$$(\varepsilon_t \theta V - c_t) e_t, \tag{12}$$

where $e_t \in \{0, 1\}$ is the date-t effort, and $\varepsilon_t \in \{0, 1\}$ is the date-t noise, which is serially correlated: $\mu \equiv \Pr(\varepsilon_2 = \varepsilon_1) \geq 1/2$. At date t, the agent's current cost c_t of undertaking the task is drawn from a random distribution with conditional support $[0, \infty)$ (this is just to avoid possible indifferences in the principal's disclosure decision). Finally, to simplify computations we assume that $\theta = \theta_H \equiv 1$ with probability f_H, while $\theta = \theta_L < 1$ with probability $f_L = 1 - f_H$. The agent's expected talent will be denoted $\theta^e \equiv f_H \theta_H + f_L \theta_L$.

When the agent undertakes the task at date 1 (which he does when c_1 is low enough) and succeeds, this of course reveals that $\varepsilon_1 = 1$. When he fails, however, only the principal learns the realization of ε_1. More specifically, assume that if the agent faced a handicap ($\varepsilon_1 = 0$) the principal receives hard evidence of it, whereas if he faced no handicap she learns nothing. When the principal has evidence of a handicap, she decides whether or not to reveal it. For simplicity, ε_1 is the only piece of private information that the principal may hold; in particular, she has no information about θ.

38. For example, Korman (1970) emphasizes the positive role of one's self-image in the determination of work attitude and effort, and argues that managers should attempt to improve the employee's self-image.

39. See Snyder *et al.* (1983) for a broad discussion of excuses.

At the end of period 1, the principal seeks to maximize the likelihood that the agent will undertake the date-2 task. Suppose that he has failed, but that the principal knows that he faced a handicap ($\varepsilon_1 = 0$). Will full disclosure occur in equilibrium? The agent's expected gross benefit from undertaking the date-2 task is $(1 - \mu)\theta^e$ if he is informed that $\varepsilon_1 = 0$, and $\mu\theta_L$ if he receives no such information (he then infers that his talent is low). Truthful disclosure is thus an equilibrium behaviour if and only if

$$\frac{\mu}{1-\mu} \leq \frac{\theta^e}{\theta_L}. \tag{13}$$

The principal trades off the benefit of boosting the agent's self-esteem by offering an excuse against the risk that this excuse may itself demotivate the agent. If external circumstances exhibit little serial correlation (μ is close to 1/2), or if self-esteem is badly affected by failure in the absence of an excuse (θ_L is very low), then in the unique equilibrium, the principal will want to disclose the excuse. When condition (13) does not hold, by contrast, the principal either does not provide any excuse, or plays a mixed strategy.[40]

Straightforward variations on this model allow us to identify the costs and benefits of praise ("you succeeded even though ε was low") and criticism ("you failed even though ε was high"). Praise boosts the agent's self-esteem but makes him doubt his environment; criticism lowers it, and yet may not discourage him. For instance, minority parents who feel that their child is being discriminated against at school, or themselves at work, are often reluctant to convey these views to their child, for fear that he or she would lose faith in the school system and the returns to educational investment.

4. UNDERMINING THE OTHER'S EGO

4.1. *Possible rationales*

Our premise until now has been—as in much of the human resources and education literatures—that a person generally benefits from a higher self-esteem of her spouse, child, colleague, coauthor, subordinate, or teammate. Yet while boosting others' self-confidence is a pervasive aspect of social interactions, people also often criticize or downplay the achievements of their colleagues and relatives. The study in Sections 2 and 3 must therefore be part of a broader construct, in which the principal sometimes wants to repress the agent's ego. We first consider some simple potential motivations for such behaviours, then turn to a more interesting one.

(a) *Direct competition*

A rather trivial reason arises when the two individuals are in direct competition (for a job, a mate, a discovery, a title, and so forth). The former is then directly hurt when the latter succeeds; in the context of our model, W is negative.[41]

(b) *The risk of "coasting"*

A basic premise in social psychology (and, consequently, our starting point in Section 2) is that the marginal payoff to an individual's effort is generally increasing in his ability. In certain situations, however, effort and ability are substitutes rather than complements, creating the risk

40. More precisely, let $\tau \equiv \Pr(\varepsilon_1 = 0)$ and $\hat{\theta} \equiv [\tau/(\tau + (1-\tau)f_L)]\theta^e + [(1-\tau)f_L/(\tau + (1-\tau)f_L)]\theta_L$. If $\mu/(1-\mu) \geq \hat{\theta}/\theta_L$, the unique equilibrium involves no disclosure. With intermediate degrees of correlation, $\hat{\theta}/\theta_L < \mu/(1-\mu) < \theta^e/\theta_L$, it involves a randomization between disclosing and not disclosing.

41. See footnote 14 for the application of our basic model to punishments and other costly disincentives.

that the agent may reduce effort when feeling more self-confident ("resting on his laurels"). This case arises in particular when the agent's private payoff from performance is of a "pass–fail" nature. For example, a pupil whose only ambition is to pass an exam may study less if he feels talented. Similarly, an individual who aims at little more than keeping his spouse and takes her for granted will not put much effort into remaining attractive to her.[42] The teacher or parent may then want to downplay the pupil's achievements, and the spouse may tell her partner that he is not so great after all.

In these examples, a high self-confidence reduces effort. In other examples, it may induce the wrong type of effort. For example, the agent may demonstrate excess initiative, selecting a new and risky path that he feels will pay off due to his talent, while the principal would have preferred a more conservative approach. There are probably many situations in which the principal's payoff as a function of the agent's self-confidence $\hat{\theta}$ is hill-shaped, as opposed to constantly increasing as posited in Section 2: an increase in the agent's self-esteem helps up to a point, beyond which it becomes hubris and starts hurting the principal.

(c) *Shadow cost of reputation*

A teacher or a manager who makes very complimentary comments to every pupil or employee may lose her credibility. As we already noted, when disclosing *soft* information to several agents the principal must realize that they will see through her ulterior motivation, and believe her only if she builds a reputation for not exaggerating claims. Refraining from boosting some agents' self-esteem may help her make more credible statements to the others.[43] A related tradeoff appears in Fang and Moscarini (2002), where a firm has private information about the ability of each of a continuum of workers.[44]

We now turn to, and analyse in more detail, what is probably the most common reason for restraining another person's ego.

4.2. *Ego bashing and battles for dominance*

Many circumstances in private life or in the workplace are characterized by power relationships. Egos clash as individuals try to establish dominance over each other along some dimension (intellectual, physical). What matters in such situations is one's relative standing in the group, rather than any absolute standing. Shattering the other's self-confidence in the relevant dimension may then increase one's power in the relationship.

42. A simple formalization of these two examples goes as follows. Suppose the agent aims at performance y_0 and gets no extra utility from $y > y_0$. Consider a deterministic technology $y = \theta e$ where θ is talent and e effort. Then $e = y_0/\theta$, and so self-confidence reduces effort.

43. This can be modelled either in a repeated-action setting or, more simply, by having the principal make simultaneous announcements to a large number of agents who each have ability θ_H or θ_L, with probabilities f_H and f_L. Whereas the principal could not credibly convey any information to a single agent (as long as $f_H\theta_H + f_L\theta_L < c/V < \theta_H$, she would always want to announce θ_H), the law of large numbers makes truth-telling an equilibrium where each compliment has a clear opportunity cost. (The equilibrium is supported by off-the-equilibrium path beliefs that a principal who declares more than a fraction f_H of agents to be talented must be babbling.) Because high-ability agents are the most productive, it is indeed optimal for the principal to reserve his praise for those types.

44. Workers observe all wage contracts (lump-sum plus bonus). Abilities are independently distributed, so by the law of large numbers the firm must convey bad news to some workers if it is to convey good news to others. Whereas under symmetric information it would give higher wages to the more able, it may then choose not to differentiate if the (motivation) costs of bad news exceeds the benefits of good ones. Fang and Moscarini show that this occurs in particular if most workers are overoptimistic about their ability.

4.2.1. Private benefits from dominance. Consider a pair of individuals, 1 and 2, who must make a joint decision (they share the "formal control right" over it). Each comes with an idea or project, but only one of these can be selected. Individual i's idea yields, in expectation, $\theta_i V + B$ to i and $\theta_i V$ to j, where θ_i is individual i's talent and $B > 0$ is a private benefit accruing to individual i when his point of view prevails. The existence of a private benefit is natural, since individuals are more likely to search for (or reveal) ideas that favour them; B could thus arise from the fact that i's favoured project is easier for him to carry out, has positive spillovers on to his other activities, or will bring him outside credit for having had the idea.

Let us assume for simplicity that θ_1 is common knowledge, whereas θ_2 can take either one of two values, θ_2^L and $\theta_2^H > \theta_2^L$, such that

$$\theta_2^H V + B > \theta_1 V > \theta_2^L V + B \qquad \text{and} \qquad \theta_1 V + B > \theta_2^H V. \tag{14}$$

We consider situations where individual 1 may have hard information about θ_2 that individual 2 himself does not have, for example some third-party's feedback about individual 2's earlier performances. In our terminology, individual 1 can thus be viewed as the principal and individual 2 as the agent, even though there is no hierarchy in terms of *a priori* control rights. For simplicity, we assume that the principal either has no information regarding θ_2, or else knows its true value. In the latter case, she can either disclose the information or conceal it. Formally, the principal receives a signal $s \in \{\varnothing, \theta_2\}$, where $\varnothing$ is uncorrelated with θ_2, and can report $r \in \{\varnothing, s\}$ to the agent. To fix ideas, suppose that two coauthors with different research styles, tastes or installed bases of contributions must make modelling choices, or decide what to emphasize. Individual 1 has private information on the popularity of individual 2's research agenda or preferred approach, about which she may have heard or read comments. She may then disclose the (good or bad) news to individual 2, or conceal them from him.

We rule out monetary transfers, again for simplicity. The timing is as follows:

Stage 1: The principal learns either nothing or learns θ_2. In the latter case, she chooses whether or not to disclose the information.

Stage 2: Both individuals come up with one idea each for a joint undertaking.

Stage 3: With probability 1/2 each, one of them is selected to make a take-it-or-leave-it offer, *i.e.* gets to choose which project will be implemented.

It is easy to see that when the principal learns that $\theta_2 = \theta_2^L$, she wants to convey this bad news to the agent, because in doing so she establishes dominance: by (14), even if the agent gets to propose the course of action he will then defer to the principal, which he would not do if he were more self-confident. By lowering the other's ego, individual 1 enjoys *real authority* despite sharing *formal authority* over decisions with individual 2. Similar ideas hold when monetary transfers between the two individuals are feasible, as long as talent-contingent outside options are available; again, the basic point is that the principal's bargaining power is enhanced when the agent's self-confidence is damaged.

The situation described above may still be viewed as a relatively tame and efficient version of the "battle of the egos", since the principal's lowering of the agent's self-confidence by revealing that $\theta_2 = \theta_2^L$ is Pareto-improving (introducing monetary transfers would thus not affect anyone's decision in this case). When this information is brought to him, individual 2 may feel disappointed, but should recognize that he is being saved from making a costly mistake.

The other state in which the principal is informed ($\theta_2 = \theta_2^H$) can, however, yield a much less harmonious and efficient outcome. Let $\overline{\theta}_2$ denote the agent's self-confidence conditional on $s \in \{\varnothing, \theta_2^H\}$, which is his information set when the principal reports no signal (since she always

passes on any bad news). By (14), when the principal learns that $\theta_2 = \theta_2^H$ she cannot lose by concealing this favourable information. It actually pays to do so if

$$\overline{\theta}_2 V + B < \theta_1 V, \tag{15}$$

as a report of "no news" will then induce the agent to submit to her authority. The principal will thus censor positive signals about the agent's ability, and would even be willing to spend resources in order to prevent them from reaching him.[45] In contrast to the earlier case, her undermining of the agent's self-confidence (by omission) is now detrimental to the latter, and may even result in a lower total surplus (if $\theta_2^H > \theta_1$). This case corresponds well to that of a mediocre and insecure manager, who abstains from passing on to his subordinates positive feedback about their performance from higher-ups or customers, for fear that they may then challenge his authority and diminish his ability to shape decisions (an extreme case being going after his job).[46]

Proposition 7. *In the game described above, individual* 1 *bashes individual* 2*'s ego in order to establish dominance and acquire real authority. She does so both by disclosing bad news to individual* 2 *and by feigning ignorance when learning good news.*

One could further enrich the analysis to capture escalating "arguments" by allowing agent 2, in response to an attack on his ego, to seek costly counter-evidence, as well perhaps as information that reflects negatively on agent 1's ability.

4.2.2. The sequencing of ego bashing and ego boosting. Ego bashing may also have costs for the principal. As shown earlier, the agent's lack of self-confidence can reduce his initiative in coming up with good projects initially, as well as his motivation for putting effort into the joint endeavour later on. The resulting tradeoff has interesting implications for situations where the principal can choose the timing of information revelation. Suppose that, as shown in Figure 1, she can release information either before or after the agent searches for a project. The principal thus first learns either the agent's true type (θ_2), or nothing ($\varnothing$); if she has information she may then disclose it. Next, the agent decides whether to search for a project, and the principal learns whether she has one of her own. The agent finds a viable project if and only if he exerts effort, which involves a private cost c. As to the principal, for simplicity we assume that she has a relevant project of her own with exogenous probability $x < 1$. The principal's uncertainty about the existence of a good project is what may make it costly for her to lower the agent's ego early on. The remainder of the game is otherwise unchanged: the principal has a second chance to disclose her information if she has not done so yet, then both jointly select a project.

To focus on the more interesting parameter configurations, we assume that (14) and (15) are satisfied, as well as

$$(1-x)(\theta_2^L V + B) < c < (1-x)(\theta_2^H V + B) + \frac{x}{2}(\theta_2^H V + B - \theta_1 V). \tag{16}$$

45. When $\overline{\theta}_2 V + B > \theta_1 V$, there are multiple equilibrium disclosure behaviours, but a unique equilibrium payoff outcome: the agent always chooses his own project. The principal's ability to conceal information is then irrelevant in the state where $\theta_2 = \theta_2^H$.

46. There are also other reasons why individual 2 may resent having his ego undercut. First, he may be suffering from a general self-motivation problem (perhaps most relevant for later, more important tasks) due to time-inconsistent preferences, which results in his attaching negative value to information about his ego (Carrillo and Mariotti (2000), Bénabou and Tirole (2002*a*)). Second, the two agents may be involved in bargaining over how to share the surplus created by their joint project, and the revelation that $\theta_2 = \theta_2^L$ may hurt individual 2's bargaining position more than it helps him by making sure that the efficient project is selected.

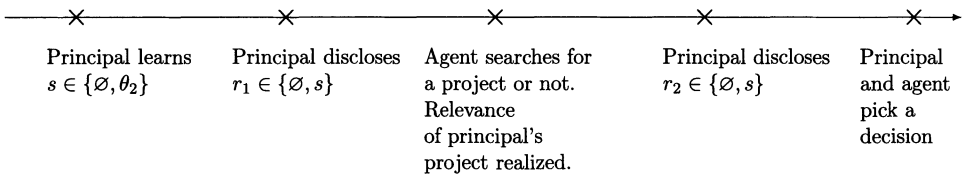

FIGURE 1
Timing of signals and reports

The first inequality in (16) states that bad news destroys incentives for effort: the project of an agent who knows he has a low talent will matter only if the principal has no viable alternative, *i.e.* only with probability $1 - x$. The second inequality, by contrast, means that a self-confident agent finds it optimal to conceive a project of his own. Having such a project makes a difference if either the principal has none (probability $1 - x$), or if she has one but the agent ends up with the bargaining power (probability $x/2$).

Definition 1. Consider stage $t = 1$ (*ex ante*) or 2 (*ex post*), and denote by $r_t(\theta_2) \in \{\varnothing, \theta_2\}$ the principal's disclosure at date t when she knows the agent's type θ_2. We say that the principal: (i) engages in ego bashing if her strategy is (or is payoff-equivalent to) $r_t(\theta_2^L) = \theta_2^L$ and $r_t(\theta_2^H) = \varnothing$; (ii) engages in ego boosting if it is (or is payoff-equivalent to) $r_t(\theta_2^L) = \varnothing$ and $r_t(\theta_2^H) = \theta_2^H$; (iii) conceals her information if her strategy is (or is payoff-equivalent to) $r_t(\theta_2) = \varnothing$ for $\theta_2 \in \{\theta_2^L, \theta_2^H\}$.

Proposition 8. *The principal engages in ego bashing* ex post (*provided she has not yet disclosed the information*). Ex ante, *she either engages in ego boosting, or refrains from disclosing information. Thus, good news tends to be revealed earlier than bad news; put differently, the relationship becomes more antagonistic over time.*

Proof. See the Appendix. ‖

To illustrate these results, one may think of a parent faced with his or her child's choice between different careers, or different sports. The parent may have information about the child's ability or value in some of these alternative activities—*e.g.* the family business. At the same time, there is always a chance (probability $1 - x$) that the parent's preferred choice will turn out not to be suited to the child anyway. The parent will then at first attempt to appear broad-minded (refrain from belittling the child's judgment or tastes), or even actively encourage him to explore alternatives. When coming close to a decision, however, the "father-knows-best" attitude will resurface, as long as the parent still has a credible alternative to the activity favoured by the child.

In the framework described by Figure 1, effort is related to the search for a project and comes before decision making. It is then quite natural that the early stage of the relationship be biased toward ego boosting, and the late stage toward ego bashing. Suppose, on the other hand, that the agent's effort relates to the implementation of a project rather than to its conception. With the choice of project coming first and the agent's effort second, it is clear that the timing of ego bashing and ego boosting is reversed: the principal tends to "beat up" on the agent at first, only to provide reassurance later on (if feasible), once the project has been selected and the principal wants the agent to be motivated.

We have emphasized here the obvious cost of ego bashing in terms of demotivation. Another cost may stem from individual 2's drawing more complex inferences about individual 1's preferences. Suppose that agent 2 cares not only about the project, but also about individual 1's altruism, friendship, or love towards him, over which he has incomplete information as well. Ego bashing may then be interpreted as individual 1 caring little about individual 2, and backfire.

Despite these costs, people may often be willing to belittle and criticize others in order to establish dominance. Because this can result in very inefficient outcomes, an interesting avenue for future research is how individuals and organizations try to limit the scope for such ego clashes. Let us, for now, content ourselves with a few thoughts in this regard. One possible strategy suggested by the model is to allocate formal control to individual 1. An example may be giving decision rights to parents until the children have reached a certain age. Another arrangement sometimes observed is the acceptance by individual 2 of individual 1's dominance (presumably because individual 2 also has private information about himself). Individual 2's "puppy dog" strategy may enable him to avoid ego clashes with individual 1. Another promising topic is the study of *institutional structures* and *personnel management strategies* designed to prevent excessive rivalry and ego clashes within organizations, and promote instead a cooperative interpersonal atmosphere.

5. CONCLUDING REMARKS

Psychologists, experts in human resource management and sociologists have long emphasized the central role played by intrinsic motivation in many social and economic interactions. In particular, they have called attention to the fact that explicit incentive schemes may sometimes backfire, especially in the long run, by undermining agents' confidence in their own abilities or in the value of the rewarded task. This side of social psychology has been largely neglected by economists.[47] The present paper has shown that these phenomena are often quite rational, and provided a formal analysis that helps reconcile the economic and the psychological views.

Several avenues of further research seem particularly interesting. The first one would combine the looking-glass self with the reverse form of signalling, namely self-presentation, in which the agent tries to signal his information to the principal.[48] The second avenue concerns the dynamics of motivation and its "management" in long-term relationships. Third, while our model already accommodates the possibility of altruism, friendship or love, it ought to be extended to allow for asymmetric information about such feelings. As noted earlier, each party would then draw from the other's behaviour subtle inferences not only about abilities and task characteristics, but also about how much the other cares about him or her. Finally, the analysis should be extended to groups. One hears frequent complaints about workplaces where egos loom large and clash too much to allow a pleasant and cooperative environment. More generally, the interactions between intrapersonal confidence-maintenance strategies, the looking-glass self, and self-presentation raise a fascinating set of questions (*e.g.* whether these strategies are mutually reinforcing), as well as issues of institutional design related to the optimal organization of educational and work environments.

APPENDIX

Proofs of Propositions 3 *and* 4. We derive here equilibrium behaviour in the two-type case, first without and then with lump-sum payments. The agent's ability can be high, θ_H (probability f_H), or low, θ_L (probability f_L).

47. With the previously noted exceptions of Frey (1997) and Kreps (1997).

48. For example, in the literatures on depression and on excuses, when the agent tries to lower the principal's expectations the latter's course of action is often a choice of whether to accept the stated reason and offer comfort, or to resist it. In so doing, the principal reveals information to the agent, that impacts his subsequent behaviour.

The conditional densities of his signal are denoted $g_H(\sigma)$ and $g_L(\sigma)$, and the MLRP simply means that g_H/g_L is increasing.

No lump-sum payment (*Proposition* 3)

Consider an arbitrary perfect Bayesian equilibrium. From Propositions 1 and 2, the equilibrium is monotonic: if B_H and B_L denote the sets of bonuses that, in equilibrium, are offered with positive probability by a principal of type θ_H and θ_L respectively (*i.e.* one who knows that the agent has ability θ_H, or θ_L) then for any $(b, b') \in B_H \times B_L$, one must have $b \leq b'$. This, in turn, implies that: (i) $B_H \subseteq B_L$, or else the θ_L type could profitably deviate to some b that was strictly less than any b' in B_L, thus leading the agent to think that he had high ability with probability one, and thereby (weakly) improving motivation while saving on incentive costs; (ii) B_H has at most one element, or else, since $B_H \subseteq B_L$, monotonicity would be violated; (iii) $B_L \setminus B_H$ also has a unique element, because it cannot be optimal to offer one bonus that signals θ_L if a lower one has the same informational content and also induces effort.

In summary, there are at most two equilibrium bonuses: a pooling bonus b and (possibly) a higher bonus signalling type θ_L. The latter is then necessarily equal to b_L^*, since the low-type principal can always guarantee herself her complete information payoff by offering b_L^*, and no bonus that signals θ_L for sure will induce the agent to work if it is less than b_L^*.

An equilibrium is thus defined by three parameters: the pooling bonus $b < b_L^*$, the probability x^* that type θ_L selects b, and

$$\left(\frac{f_H g_H(\sigma^*)}{f_H g_H(\sigma^*) + f_L g_L(\sigma^*) x^*} \theta_H + \frac{f_L g_L(\sigma^*) x^*}{f_H g_H(\sigma^*) + f_L g_L(\sigma^*) x^*} \theta_L \right) V = c, \tag{A.1}$$

which defines the agent's cutoff σ^*, with $\sigma^* \epsilon (0, 1)$ by the full-support assumption.

Let us next show that there exists a unique equilibrium satisfying the NWBR criterion, and that it is such that $b = 0$. Consider, by contradiction, a pooling bonus $b > 0$, with associated cutoff σ. We will show that, for any $\hat{b} < b$, if a deviation to $\hat{b}$ elicits from the agent a response, described by a cutoff $\hat{\sigma}$, that results in weakly higher profits for a principal of type θ_L,

$$\theta_L [1 - G(\hat{\sigma} \mid \theta_L)](W - \hat{b}) \geq \theta_L [1 - G(\sigma \mid \theta_L)](W - b), \tag{A.2}$$

then this same deviation will be strictly preferred by a principal of type θ_H. This is obvious when $\hat{\sigma} \leqslant \sigma$, since a deviation to $\hat{b}$ then yields higher effort by the agent, at a lower cost for the principal (of any type). Suppose now that $\hat{\sigma} > \sigma$; we then have

$$\frac{W - b}{W - \hat{b}} \leq \frac{1 - G(\hat{\sigma} \mid \theta_L)}{1 - G(\sigma \mid \theta_L)} < \frac{1 - G(\hat{\sigma} \mid \theta_H)}{1 - G(\sigma \mid \theta_H)}, \tag{A.3}$$

where the first inequality obtains from (A.2) and the second follows from the MLRP, since $\theta_L < \theta_H$ and $\sigma < \hat{\sigma}$. Thus, if a deviation to $\hat{b}$ is profitable for type θ_L, then *a fortiori* it is strictly profitable for type θ_H. Hence, according to the NWBR criterion, when the agent is offered a $\hat{b} < b$ (say, $\hat{b} = b - \varepsilon$), he should believe that it emanates from type θ_H. Therefore, $b > 0$ cannot be part of a pooling equilibrium, and so it must be that $B = \{0, b_L^*\}$.

Next, note that when $\theta = \theta_L$ the principal must randomize between $b = 0$ and $b = b_L^*$: if she only played b_L^*, then $b = 0$ would induce compliance with probability 1, and would therefore be preferred to b_L^*. The equilibrium is thus described by two parameters: $x^* \in (0, 1]$, the probability that a principal observing $\theta = \theta_L$ selects bonus 0 (pools); and σ^*, the agent's cutoff signal for working when he is offered a zero bonus. These are given by

$$\theta_L (W - b_L^*) = \theta_L [1 - G_L(\sigma^*)] W, \tag{A.4}$$

expressing the principal's willingness to randomize when $\theta = \theta_L$, and equation (A.1), which can be rewritten as

$$\frac{g_H(\sigma^*)}{g_L(\sigma^*)} = x^* \left(\frac{f_L}{f_H} \right) \left(\frac{c/V - \theta_L}{\theta_H - c/V} \right). \tag{A.5}$$

By our assumptions on the likelihood ratio, for any $x^* > 0$ there exists a unique solution $\sigma^* = s(x^*)$ to (A.1), with $\sigma^* > 0$. Substituting into (A.4), the principal's net incentive to offer bonus 0 when $\theta = \theta_L$ and the agent expects her to randomize with probability x^* is

$$\theta_L [1 - G_L(s(x^*))] W - \theta_L (W - b_L^*). \tag{A.6}$$

This function is increasing in x^*, and negative at $x^* = 0^+$. So either it has a unique zero on (0, 1), which then defines the principal's mixing strategy; or else it is non-positive on all of (0, 1], in which case the principal's equilibrium strategy is $x^* = 1$, meaning that *no bonus is ever offered.* In both cases the equilibrium is unique. Note, finally, that the agent works only with probability $1 - G_H(s(x^*))$ when $\theta = \theta_H$, and with probability $1 - x^* G_L(s(x^*))$ when $\theta = \theta_L$. Thus, in either state of the world, he works less than under symmetric information (where $e = 1$ with probability one). ||

B. Lump-sum payments (Proposition 4)

In any separating equilibrium the agent always works; when $\theta = \theta_L$ he thus gets bonus b_L^* and no lump-sum payment. Denoting by (a, b) the contract he gets when $\theta = \theta_H$, let us show that when $\theta = \theta_L$ the principal must be indifferent between offering $(0, b_L^*)$ and (a, b). Clearly the θ_L-type principal cannot strictly prefer (a, b), and were she to strictly prefer $(0, b_L^*)$, the θ_H-type could profitably deviate to $(a - \varepsilon, b)$, for small $\varepsilon > 0$. According to NWBR this would still convince the agent that $\theta = \theta_H$, since for ε small enough the θ_L-type still prefers $(0, b_L^*)$ to $(a - \varepsilon, b)$ even when the agent works with probability 1 after being offered the latter contract. Next, as in the proof of Proposition 3, it is easy to show that $b = 0$: were $b > 0$, the principal could offer $b - \varepsilon$ and the agent would interpret this (according to NWBR) as a sure sign that $\theta = \theta_H$.

The NWBR separating equilibrium, if it exists, is thus necessarily defined by the contracts $(0, b_L^*)$ and $(a = c - \theta_L V, 0)$. Moreover, the limited-informativeness condition ensures that the NWBR criterion is indeed satisfied. Indeed, suppose that the principal deviates to a contract $(\hat{a}, \hat{b})$, and let $\hat{\sigma}$ define the agent's reaction to this offer. If this is weakly profitable for the θ_H type,

$$\theta_H(1 - G_H(\hat{\sigma}))(W - \hat{b}) - \hat{a} \geq \theta_H W - (c - \theta_L V),$$

then (10) implies that it is strictly profitable for the θ_L type,

$$\theta_L(1 - G_L(\hat{\sigma}))(W - \hat{b}) - \hat{a} > \theta_L W - (c - \theta_L V),$$

as long as either $b > 0$ or $\hat{\sigma} > 0$; when $\hat{b} = \hat{\sigma} = 0$ both types gain, equally. According to NWBR the agent should then interpret any such deviation as indicative of $\theta = \theta_L$, and shirk, so it cannot be profitable. ‖

Proof of Proposition 5

For a given delegation policy $d \in \{0, 1\}$, the agent undertakes the task if and only if $\mathsf{E}[\theta \mid \sigma, d]V \geq c_d$. Therefore, there is a cutoff σ_d^* such that he exerts effort if and only if $\sigma \geq \sigma_d^*$. The principal then chooses $d \in \{0, 1\}$ so as to maximize $[1 - G(\sigma_d^* \mid \theta)]\mathcal{W}_d(\theta)$. Suppose first that $\sigma_1^* \geq \sigma_0^*$. Then, from the MLRP and Assumption 1:

$$\frac{d}{d\theta}\left[\left(\frac{1 - G(\sigma_1^* \mid \theta)}{1 - G(\sigma_0^* \mid \theta)}\right)\left(\frac{\mathcal{W}_1(\theta)}{\mathcal{W}_0(\theta)}\right)\right] > 0, \tag{A.7}$$

and so the principal delegates if and only if $\theta \geq \tilde{\theta}$ for some $\tilde{\theta}$. This implies that delegation is good news: $E[\theta \mid \sigma, 1] > E[\theta \mid \sigma, 0]$ for all σ; in particular, $c_1 = E[\theta \mid \sigma_1^*, 1] > E[\theta \mid \sigma_0^*, 0] = c_0$, a contradiction. So it must be that $\sigma_1^* < \sigma_0^*$, meaning that (for given θ) the agent works more under delegation; hence part (i) of the proposition. As a result, for all $\theta > \theta^*$ we have $[1 - G(\sigma_1^* \mid \theta)]W_1(\theta) > [1 - G(\sigma_0^* \mid \theta)]W_0(\theta)$, so the principal strictly prefers to delegate. By continuity, this remains true on some interval $[\theta^{**}, \theta^*]$ with $\theta^{**} < \theta^*$. This proves part (ii). This property, however, does not necessarily make θ^{**} a cutoff such that delegation occurs if *and only if* $\theta > \theta^{**}$: indeed with $\sigma_1^* < \sigma_0^*$, the term in square brackets in (A.7) need no longer be increasing in θ (although it cannot be everywhere decreasing, otherwise the principal would delegate if and only if θ was above a cutoff $\hat{\theta}$, contradicting (ii)). When Assumption 2 holds, however, (A.7) again applies, implying that delegation occurs if and only if $\theta \geq \theta^{**}$, where $\theta^{**} < \theta^*$. This, in turn, ensures part (iii) of the proposition, and more generally implies that delegation raises the agent's posterior distribution about his ability, independently of his signal and of the task's outcome. ‖

Proof of Proposition 8

Let us work by backward induction.

(a) Suppose that at $t = 2$ the principal knows the agent's type, but has not yet disclosed it. Let $y(\theta_2^L)$, $y(\varnothing)$ and $y(\theta_2^H)$ denote the equilibrium probabilities that the agent will stick to his own project when: (a) the principal also has a relevant project, but it is the agent who gets to make the selection; and (b) the agent was previously told $r_2 = \theta_2^L$, $\varnothing$ or θ_2^H respectively. From (14) and (15), we have

$$y(\theta_2^L) = 0 \leq y(\varnothing) \leq y(\theta_2^H) = 1. \tag{A.8}$$

If $y(\varnothing) > 0$, it is clearly strictly optimal for the principal to disclose $\theta_2 = \theta_2^L$ *ex post*. In the absence of any news the agent will thus infer that $s \in \{\varnothing, \theta_2^H\}$, and condition (15) then implies that $y(\varnothing) = 0$ after all, a contradiction. The fact that $y(\varnothing) = 0 < 1 = y(\theta_2^H)$, in turn, makes it strictly optimal for the principal not to disclose that $\theta_2 = \theta_2^H$, when she has such news. Finally, when the principal knows that $\theta_2 = \theta_2^L$ she is indifferent between disclosing it and saying

nothing, since $y(\theta_2^L) = y(\varnothing) = 0$ (this is why we allow for payoff-equivalence in the definition preceding Proposition 8). We may thus, without loss of generality, assume that she chooses to reveal the bad news.

(b) Let us now look at the principal's *ex ante* behaviour ($t = 1$) and the agent's subsequent effort decision. Let $z(\theta_2^L)$, $z(\varnothing)$ and $z(\theta_2^H)$ denote the probabilities that the agent searches for a project after hearing reports $r_1 = \theta_2^L$, $\varnothing$ and θ_2^H respectively. Equation (16) implies that

$$z(\theta_2^L) = 0 \leq z(\varnothing) \leq z(\theta_2^H) = 1. \tag{A.9}$$

Either $z(\varnothing) = 0$, and then we might as well assume that the principal reports $r_1(\theta_2^L) = \varnothing$ (if not, change her behaviour to $r_1(\theta_2^L) = \varnothing$ for sure; then $z(\varnothing) = 0$ *a fortiori*). Or else $z(\varnothing) > 0$, and then it is strictly optimal for the principal not to disclose θ_2^L *ex ante*, given that she can (and will) do it *ex post*.

Last, when $\theta_2 = \theta_2^H$ the principal faces a tradeoff if $z(\varnothing) < z(\theta_2^H)$. Disclosing the good news raises incentives for effort, but emboldens the agent in the *ex post* negotiation. It can be shown that depending on the values of the parameters, the principal may or may not disclose $\theta_2 = \theta_2^H$ in equilibrium. ‖

Acknowledgements. The first version of this paper was titled "Self-Confidence and Social Interactions". We are grateful for helpful comments and discussions to Philippe Aghion, Mark Armstrong, Isabelle Brocas, Daniel Gilbert, Robert Lane, Marek Pycia, Gérard Roland, Julio Rotemberg, Ilya Segal, Anton Souvorov, participants at the Franqui conference on "The Economics of Contracts" (Brussels, 1999), participants at seminars at Harvard and Paris, and at the ISNIE 2002 Congress (Boston) and three anonymous referees. Bénabou gratefully acknowledges financial support from the National Science Foundation (SES-0096431) and the MacArthur Foundation, as well as the hospitality of the Institute for Advanced Study over the academic year 2002–2003.

REFERENCES

AGHION, P. and TIROLE, J. (1997), "Formal and Real Authority in Organizations", *Journal of Political Economy*, **105**, 1–29.

AKERLOF, G. and DICKENS, W. (1982), "The Economic Consequences of Cognitive Dissonance", *American Economic Review*, **72** (3), 307–331.

BARON, J. and KREPS, D. (1999) *Strategic Human Resources* (New York: John Wiley).

BEM, D. J. (1967), "Self-Perception: An Alternative Interpretation of Cognitive Dissonance Phenomena", *Psychological Review*, **74**, 183–200.

BÉNABOU, R. and TIROLE, J. (2002*a*), "Self-Confidence and Personal Motivation", *Quarterly Journal of Economics*, **117** (3), 871–915.

BÉNABOU, R. and TIROLE, J. (2002*b*), "Willpower and Personal Rules" (CEPR Discussion Paper 3143).

BIBRING, E. (1953), "The Mechanism of Depression", in P. Greenacre (ed.) *Affective Disorders* (New York: International University Press).

CARRILLO, J. and MARIOTTI, T. (2000), "Strategic Ignorance as a Self-Disciplining Device", *Review of Economic Studies*, **66**, 529–544.

CHO, I. K. and KREPS, D. (1987), "Signaling Games and Stable Equilibria", *Quarterly Journal of Economics*, **102**, 179–221.

CONDRY, J. and CHAMBERS, J. (1978), "Intrinsic Motivation and the Process of Learning", in M. Lepper and D. Greene (eds.) *The Hidden Cost of Reward: New Perspectives on the Psychology of Human Motivation* (New York: John Wiley).

COOLEY, C. (1902) *Human Nature and the Social Order* (New York: Scribner's).

DARLEY, J. and FAZIO, R. (1980), "Expectancy Confirmation Processes Arising in the Social Interaction Sequence", *American Psychologist*, **35**, 867–881.

DECI, E. (1975) *Intrinsic Motivation* (New York: Plenum Press).

DECI, E., KOESTNER, R. and RYAN, R. (1999), "A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation", *Psychological Bulletin*, **125** (6), 627–668.

DECI, E. and RYAN, R. (1985) *Intrinsic Motivation and Self-Determination in Human Behavior* (New York: Plenum Press).

DELFGAUW, J. and DUR, R. (2002), "Signaling and Screening of Workers' Motivation" (Tinbergen Institute DP 2002-050/3).

DESSEIN, W. (2002), "Authority and Communication in Organizations", *Review of Economic Studies*, **69** (4), 811–838.

ETZIONI, A. (1971) *Modern Organizations* (Englewood Cliffs, NJ: Prentice-Hall).

FANG, H. and MOSCARINI, G. (2002), "Overconfidence, Morale and Wage-Setting Policies" (Mimeo, Yale University).

FEHR, E. and FALK, A. (1999), "Wage Rigidity in a Competitive Incomplete Contract Market", *Journal of Political Economy*, **107** (1), 106–134.

FEHR, E. and SCHMIDT, K. (2000), "Fairness, Incentives, and Contractual Choices", *European Economic Review*, **44** (4–6), 1057–1068.

FESTINGER, L. and CARLSMITH, J. (1959), "Cognitive Consequences of Forced Compliance", *Journal of Abnormal and Social Psychology*, **58**, 203–210.

FREY, B. (1997) *Not Just for the Money—An Economic Theory of Personal Motivation* (Cheltenham, U.K.: Edward Elgar).
FREY, B., FEHR, E. and BENZ, M. (2000), "Does Motivation Crowding Out Spread Beyond the Area of Intervention? Experimental Evidence" (Mimeo, University of Zurich).
FREY, B. and OBERHOLZER-GEE, F. (1997), "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out", *American Economic Review*, **87** (4), 746–755.
FUDENBERG, D. and TIROLE, J. (1991) *Game Theory* (Cambridge, MA: MIT Press).
GIBBONS, R. (1997), "Incentives and Careers in Organizations", in D. Kreps and K. Wallis (eds.) *Advances in Economic Theory and Econometrics*, Vol. II (Cambridge, U.K.: Cambridge University Press).
GILBERT, D. and SILVERA, D. (1996), "Overhelping", *Journal of Personality and Social Psychology*, **70**, 678–690.
GNEEZY, U. and RUSTICHINI, A. (2000*a*), "A Fine is a Price", *Journal of Legal Studies*, **29** (1) (part 1), 1–17.
GNEEZY, U. and RUSTICHINI, A. (2000*b*), "Pay Enough or Don't Pay at All", *Quarterly Journal of Economics*, **115** (3), 791–810.
HOLMSTRÖM, B. and MILGROM, P. (1991), "Multi-Task Principal–Agent Analyzes: Incentive Contracts, Asset Ownership, and Job Design", *Journal of Law, Economics and Organization*, **7**, 24–52 (special issue).
HULL, C. L. (1943) *Principles of Behavior* (New York: Appleton-Century-Crofts).
KINLAW, D. (1997) *Coaching: Winning Strategies for Individuals and Teams* (U.K.: Gower Publishing).
KOHN, A. (1993) *Punished by Rewards* (New York: Plenum Press).
KORMAN, A. K. (1970), "Toward an Hypothesis of Work Behavior", *Journal of Applied Psychology*, **54**, 31–41.
KREPS, D. (1997), "Intrinsic Motivation and Extrinsic Incentives", *American Economic Review*, **87** (2), 359–364.
KRUGLANSKI, A. (1978), "Issues in Cognitive Social Psychology", in *The Hidden Cost of Reward: New Perspectives on the Psychology of Human Motivation* (New York: John Wiley).
KRUGLANSKI, A., FRIEDMAN, I. and ZEEVI, G. (1971), "The Effect of Extrinsic Incentives on Some Qualitative Aspects of Task Performance", *Journal of Personality*, **39**, 608–617.
LAFFONT, J.-J. and TIROLE, J. (1988), "Repeated Auctions of Incentive Contracts, Investment and Bidding Parity, With an Application to Takeovers", *Rand Journal of Economics*, **19**, 516–537.
LAZEAR, E. (2000), "Performance, Pay and Productivity", *American Economic Review*, **90** (5), 1346–1361.
LEPPER, M. and GREENE, D. (1978), "Overjustification Research and Beyond: Toward a Means-Ends Analysis of Intrinsic and Extrinsic Motivation", in *The Hidden Cost of Reward: New Perspectives on the Psychology of Human Motivation* (New York: John Wiley).
LEPPER, M., GREENE, D. and NISBETT, R. (1973), "Undermining Children's Interest with Extrinsic Rewards: A Test of the 'Overjustification Hypothesis'", *Journal of Personality and Social Psychology*, **28**, 129–137.
MERTON, R. (1948), "The Self-Fulfilling Prophecies", *Antioch Review*, **8**, 193–210.
MILES, R. (1965), "Human Relations and Human Resources", *Harvard Business Review*.
PFEFFER, J. (1994) *Competitive Advantage Through People: Problems and Prospects for Change*, Chapter 4 (Boston: Harvard Business School Press).
ROSENTHAL, R. and JACOBSON, L. (1968) *Pygmalion in the Classroom* (Holt-Rinehart-Winston).
SALANCIK, G. (1977), "Commitment and the Control of Organizational Behavior and Beliefs", in B. Staw and G. Salancik (eds.) *New Directions in Organizational Behavior* (Chicago: St. Clair Press).
SCHWARTZ, B. (1990), "The Creation and Destruction of Value", *American Psychologist*, **45**, 7–15.
SKINNER, G. F. (1953) *Science and Human Behavior* (New York: Macmillan).
SNYDER, C., HIGGINS, R. and STUCKY, R. (1983) *Excuses: Masquerades in Search of Grace* (New York: John Wiley).
SOUVOROV, A. (2003), "Addiction to Rewards" (Mimeo, GREMAQ, Toulouse).
STAW, B. (1977), "Motivation in Organizations: Toward Synthesis and Redirection", in B. Staw and G. Salancik (eds.) *New Directions in Organizational Behavior* (Chicago: St. Clair Press).
WILSON, T., HULL, J. and JOHNSON, J. (1981), "Awareness and Self-Perception: Verbal Reports on Internal States", *Journal of Personality and Social Psychology*, **40**, 53–71.