SELF-CONFIDENCE AND SOCIAL INTERACTIONS[1]

Roland Bénabou[2] Jean Tirole[3]

First Version: June 1999
This Version: December 1999

[1]We are grateful for helpful comments and discussions to Isabelle Brocas, Robert Lane, Marek Pycia, Gérard Roland, Julio Rotemberg, and participants at the Franqui conference on "The Economics of Contracts" (Brussels, November 1999).

[2]Princeton University, NBER, CEPR and IRP.

[3]IDEI and GREMAQ (UMR 5604 CNRS), Toulouse, CERAS (URA 2036 CNRS), Paris, CEPR, and MIT.

Abstract

This paper studies the interactions between an individual's self–esteem and his social environment, whether in the workplace, at school, or in personal relationships. A person generally has only imperfect knowledge of his own ability (or long–term payoff) in pursuing a task, and will undertake it only if he has sufficient *self–confidence.* People who interact with him (parent, spouse, friend, teacher, manager, colleague, etc.) often have complementary information about his ability, but also a vested interest in his completing the task. This generates an incentive for such principals to distort their signals so as to manipulate the agent's self–confidence.

We first study situations where an informed principal chooses an incentive structure, such as offering payments or rewards, delegating a task, or simply giving encouragement. We show that *rewards may be weak reinforcers* in the short term and that, as stressed by psychologists, they may have *hidden costs* in that they become negative reinforcers once withdrawn. By offering a low–powered incentive scheme, the principal signals that she trusts the agent. Conversely, rewards (*extrinsic motivation*) have a limited impact on the agent's current performance, and reduce his *intrinsic motivation* to undertake similar tasks in the future. Similarly, empowering the agent is likely to increase his motivation and effort, while offers of help or assistance may create a dependance. More generally, we identify under which conditions the hidden costs of rewards are a myth or a reality.

We then consider the fact that people often criticize or downplay the achievements of their spouse, child, colleague, coauthor, subordinate or teammate. We formalize such situations of *ego–bashing,* and argue that they may reflect *battles for dominance.* By lowering the other's ego, an individual may gain (or regain) real authority within the relationship.

Finally, we turn to the case where it is the agent who has superior information, and may attempt to signal it through a variety of *self–presentation* strategies. In particular, people with low self–esteem often deprecate their own accomplishments in order to obtain leniency (a lowering of expectancies) or a "helping hand" on various obligations. Such strategies are costly: they are met with disapproval, and may backfire if the desired indulgence is denied. We analyze this signalling game, and characterize the levels of self–esteem that give rise to *self–deprecation.*

Keywords: self–confidence, self–presentation, motivation, rewards, incentives, standards, signalling, psychology and economics.

JEL Classification: A12, C70, D10, D60, J22, J24, J53.

1 Introduction

Should a child be rewarded for passing an exam? What impact do performance bonuses, monitoring or empowerment have on employees' morale and productivity? What are the costs and benefits of standards and expectancies? How can depressive subjects react negatively to information favorable to their self–esteem while simultaneously seeking help from others in costly fashions? These questions are approached in this paper from a unifying perspective which emphasizes the interaction between an individual's self–esteem, his self–presentation strategies, and his environment.

The first premise of our analysis is that people have imperfect knowledge of their own abilities in many of the tasks they face. We therefore study decision–making by an agent (child, student, employee, etc.; "he") who faces uncertainty about his payoffs from pursuing a ceratin course of action. The unknown variable could be a characteristic of the individual himself, such as his talent, 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. In presenting the model we shall generally emphasize the case of unknown ability, which corresponds most closely to the intuitive notion of self–confidence; but it will be clear from the analysis that all three are formally equivalent.

Second, we adopt a cognitive approach, assuming that the individual is an information processor who extracts from his environment signals that are relevant for his self–confidence.[1] 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 to account for the fact that they cannot systematically and repeatedly fool themselves, or others for that matter.

Third, self–knowledge is relevant to the extent that, in most tasks, ability and effort are complementary factors in the production of performance.[2] Thus, an agent undertakes an activity only if he has sufficient self–confidence in his ability to succeed.

Because of this complementarity, people interacting with this individual and having a stake in his action have an incentive to manipulate information relevant to his self–knowledge. Thus, in much of the paper, a principal (parent, spouse, friend, teacher,

[1] In section 3.3 we will compare this approach to a more behaviorist rendition of social interactions.

[2] There are also instances of substitutability, and we shall consider them as well. What is essential for our argument is nonseparablility.

boss, colleague, etc.; "she") has a vested interest in (derives a benefit from) the agent's undertaking and succeeding in the activity. In other words, success generates a spillover or positive externality on the principal.

In many circumstances, both the agent and the principal have private information about the agent's ability to perform the task. The agent usually has better access to memories of his previous performances and of the way these were accomplished (his effort intensity, the idiosyncratic random events facilitating or inhibiting performance). That is, the agent often has more factual information that is relevant to his self.

By contrast, the principal often has private complementary information about the task. For example, a teacher or manager has information about he difficulty of the subject or assignment which, together with the agent's ability, conditions the probability of success. She may also know better than the agent whether the task is attractive, in terms of either being enjoyable to perform, or of having a high payoff for the agent. Last, while having less information relevant to the agent's self than him, she may be better trained at interpreting it. This may result from her having performed the current task herself, having seen many others do it, and thus having more experience with inferring ability from outcomes. As we discuss below, the observation that others have private information about an individual's self underlies several fields of research in education and management.

The paper analyzes the consequences of both types of private information. Sections 2 and 3 look at attributions made by the 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 agent'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 the principal's information from the latter's decision, was called the "looking–glass self" by Cooley. The influence of the principal's decision on the agent's behavior 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. We first show that *rewards may be 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. By offering a low–powered incentive scheme, 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 the same (or a similar) task 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 dependance.

We conclude that rewards may, but need not, be negative reinforcers. Our analysis actually suggests when rewards work and when they backfire. The negative reinforcement effect requires that the individual be less knowledgeable in some dimensions than the principal; this asymmetry of information is important in some settings, and negligible in others. Furthermore, a "sorting condition" must hold, in that the principal must be more tempted to offer a reward when the agent has limited ability or the task is boring. Thus, 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.

While much of the social psychology and human resource management literatures emphasize the necessity of coaching and boosting 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. We consider several reasons why this may be so, and formalize in more detail the most important one. We argue that *ego–bashing* may reflect *battles for dominance.* By lowering the other's ego, an individual may gain real authority within the relationship.

Section 4, in contrast to sections 2 and 3, looks at the impact of the agent's having superior information about his self. The theme there, related to the literature on *self–presentation,* is that the agent may attempt to signal his self–relevant information to the principal. We are particularly interested in the strategies employed by individuals with low self–esteem. These often "blackmail" others for attention and try to get reassurance about their talent or worthiness (as colleagues, students, spouses, or children). They also deprecate their own accomplishments in order to obtain leniency (a lowering of expectancies) or a "helping hand" on various obligations. Such strategies are costly: they are met with disapproval, and may backfire if the response does not bring the desired reassurance or indulgence. We provide an equilibrium analysis of these phenomena, and characterize the levels of self–esteem that give rise to *self–deprecation.*

2 The looking–glass self

2.1 General confidence–management strategies

Let us begin with an abstract framework, then specialize it. There are two players, an agent and a principal. The agent selects an action or effort level e that impacts his and the principal's utilities. The principal knows a parameter θ (e.g., the agent's ability to perform the task) that affects the effectiveness of this action. Furthermore, the principal selects a policy p. Depending on the application, this policy may be a contingent reward, help, surveillance, delegation, disclosure of information, or any other policy that may affect, directly or indirectly, the agent's behavior. The agent's and the principal's payoff functions are $U_A(\theta, e, p)$ and $U_P(\theta, e, p)$. The agent may privately receive a signal σ that is relevant to his self–knowledge, i.e., be informative about θ. The timing goes as follows

Stage 1 : The principal selects a policy p.

Stage 2 : The agent, after observing the policy chosen by the principal and learning σ, chooses an action e.

Let us assume, for notational simplicity, that the agent's optimal action e^* depends on p and on the agent's conditional expectation $\widehat{\theta}(\sigma, p)$ of her ability. The conditioning of $\widehat{\theta}$ on p means that the agent tries to see through the principal's ulterior motivation in choosing p. The principal's expected payoff from choosing policy p when she has information θ is

$$\mathsf{E}_\sigma \left[U_P \left(\theta, e^* \left(p, \widehat{\theta}(\sigma, p) \right), p \right) \right].$$

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

$$\mathsf{E}_\sigma \left[\frac{\partial U_P}{\partial p} + \frac{\partial U_P}{\partial e} \frac{\partial e^*}{\partial p} + \frac{\partial U_P}{\partial e} \frac{\partial e^*}{\partial \widehat{\theta}} \frac{\partial \widehat{\theta}}{\partial p} \right] = 0. \tag{1}$$

The first term on the left–hand side of (1) is the direct effect on the principal's payoff. So, for example, if the policy is a bonus as in the next section, the first term is the direct compensation cost of this bonus, keeping the agent's behavior constant. The second term corresponds to the direct impact on the agent's behavior. For example, ceteris

paribus, a bonus increases the agent's incentive to exert effort. These two effects have been investigated in detail in the agency literature.

We shall be interested in the third effect, which represents "confidence management". To the extent that the choice of policy is guided by the principal's knowledge, the agent updates his beliefs in reaction to the policy choice (term $\partial\widehat{\theta}/\partial p$). The principal must account for the fact that her choice is interpreted by the agent, and thereby affects his self–confidence. A key aspect is then whether a higher self–confidence level influences the agent's decision–making in a direction that the principal likes $\left(\frac{\partial U_P}{\partial e}\frac{\partial e^*}{\partial\widehat{\theta}} > 0\right)$ or dislikes $\left(\frac{\partial U_P}{\partial e}\frac{\partial e^*}{\partial\widehat{\theta}} < 0\right)$.

Section 2 examines situations in which the principal gains from boosting the agent's self–confidence. Section 3.4, on the other hand, argues that in a variety of situations the principal may be reluctant to enhance the agent's self–confidence, or may even want to undermine it; section 3.5 provides a formal illustration.

2.2 The hidden cost of rewards: the debate

It is a common theme of economics that contingent rewards encourage effort and performance, and there is a good amount of evidence that they actually do (Gibbons 1997, Lazear 1996).[3] In other words, rewards serve as "positive reinforcers" for the desired behavior. In psychology, their effect is more controversial. A long–standing paradigm clash has opposed proponents of the economic view of a positive relation between rewards and attitude to the "dissonance theorists", who argue that rewards may actually impair performance, i.e., be "negative reinforcers".[4]

A different and substantial body of evidence has shown that extrinsic motivation (contingent rewards) may conflict with intrinsic motivation (the individual's drive to perform the task for its intrinsic qualities). For example, in an experiment run by Wilson et al. (1981), college students were either paid or not paid to work on a interesting puzzle. Subjects who were rewarded played with the puzzle significantly less in a later unrewarded

[3] At least along the dimension that is being targeted by the reward. Other tasks may be impaired by the reward given for a specific task, for instance through a crowding-out effect as in Holmström and Milgrom (1991).

[4] See, e.g., Kruglanski (1978) for an account of this debate.

"free–time" period than unrewarded subjects. Similarly, Kruglanski et al. (1971) induced high–school children to perform tasks involving verbal skills. Some were promised a reward, others not. Those in the no–reward condition later reported a greater interest in the task. In daily life, parents are quite familiar with what we may call the "forbidden fruit" effect: powerful or salient extrinsic constraints employed to induce a child to comply with an adult's prohibition of an activity decrease the likelihood of the child's subsequent internalization of the adult's preference over this activity.[5] Taken together, these two points hint at a limited impact of rewards on "engagement" (current activity) and a negative one on "re–engagement" (persistence).

Another 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 behavior 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.[6] 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 on 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.

Our goal here is twofold. First, for reasons explained in more detail in section 3.3, we want to analyze potential hidden effects of rewards from an economic angle instead of just positing an aversive impact of rewards on motivation. Second, and given that extrinsic incentives work effectively in many well–documented settings, we want to understand when they should be employed with caution.

Section 3.1 will relate our approach and conclusions to some of the relevant psychology literature, in particular Deci and Ryan (1985) and Lepper et al. (1973). Let us just here mention alternative interpretations of the hidden cost of rewards, that are unrelated to

[5] See, e.g., Lepper and Greene (1978).

[6] See also Kohn (1993) and several chapters in Lepper and Greene (1978).

our modeling. As for the engagement part, Condry and Chambers (1978, 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, Condry and Chambers argue that current rewards may decrease the individual's willingness to persist because they orient activity toward performance rather than progress; that is, Condry and Chambers offer what to economists is a familiar multitask interpretation: the individual is led by higher short–term rewards to sacrifice long–run payoffs.[7] For example, a stylized fact is that subjects who are paid to solve problems choose easier ones than those who do not expect a reward. While this explanation is well–taken, it does not apply uniformly. For instance, the individual may not be aware of future re–engagement or may not face an investment decision that crowds out current efficiency. Furthermore, the multitask story cannot account for the evidence on reported posterior intrinsic interest in the activity.

2.3 Interaction between extrinsic and intrinsic motivation

As in the general framework of section 2.1, 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 undertaking the task is denoted $c > 0$. If the task is successful it yields direct payoff $V > 0$ to the agent and $W > 0$ to the principal; if it fails, both get 0.

Success requires effort; yet effort is not sufficient for success. Let $\theta \in [0, 1]$ denote the probability of success when the agent works. This section focuses on the principal's superiority of information; without loss of generality, we shall assume that she knows θ perfectly. The agent knows that θ is drawn from a cumulative distribution function $F(\theta)$ with density $f(\theta)$, and learns a signal $\sigma \in [0, 1]$ with conditional cumulative distribution $G(\sigma \mid \theta)$ and positive conditional density $g(\sigma \mid \theta)$. We assume that a higher σ is "good news", in the sense that the expectation $\mathsf{E}[\theta \mid \sigma, I]$ is weakly increasing in σ for any information I the agent may have besides σ. We further assume the monotone likelihood

[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 (see Holmström and Milgrom 1991 for a broader perspective on multitasking). 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.

ratio property (MLRP):

$$\text{for all } \sigma_1 \text{ and } \sigma_2 \text{ with } \sigma_1 > \sigma_2, \ \frac{g(\sigma_1|\theta)}{g(\sigma_2|\theta)} \text{ is increasing in } \theta. \tag{2}$$

We take the principal–agent relationship as a given, and so the non–contingent part of the reward has no impact and is normalized to zero. By contrast, we allow the principal to select a reward or bonus $b < W$, in case of success. Thus, the agent's (respectively, the principal's) total benefit in case of success is $V+b$ (respectively, $W-b$), while both parties obtain 0 in case of failure (the agent's payoff, as noted earlier, is defined gross of the effort cost). The stage–1 policy decision for the principal is therefore the choice of a reward. We formalize the reward as being a monetary one, but, in line with the psychology literature, b could with slight modifications be interpreted as working conditions, praise, friendliness or (minus) punishment.

Note that, were the agent to know his ability θ, he would choose to exert effort if and only if

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

That is, when the agent has the same information as the principal, the reward is a positive reinforcer.

In our model, however, only the principal observes θ; the agent receives only a signal σ about θ.[8] We analyze perfect Bayesian equilibria of the two–stage game. Observing reward b, the agent updates his beliefs about θ, using the principal's equilibrium strategy. Let $\widehat{\theta}(\sigma, b) \equiv \mathsf{E}[\theta|\sigma, b]$ denote the agent's (interim) self–confidence, that is, the expectation of ability conditional on his signal and the reward he is offered. This expectation is a weakly increasing function of the signal σ. Letting $e \in \{0, 1\}$ denote the agent's effort, his utility is $U_A = \left[\widehat{\theta}(\sigma, b)(V+b) - c\right] e$. There exists a threshold signal $\sigma^*(b)$ in $[0, 1]$ such that[9]

$$\mathsf{E}[\theta|\sigma, b] \geq \frac{c}{V+b} \text{ if and only if } \sigma \geq \sigma^*(b). \tag{3}$$

[8] That the principal be uncertain about what the agent exactly knows is needed in order for her choice of a bonus to be informative. If the principal did not face some uncertainty, the probability that a bonus b elicits effort would be known in advance. The optimal bonus, which maximizes the principal's payoff $\Pr[e = 1|b] \times \theta[W - b]$, would then be independent of θ.

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

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

$$\mathsf{E}_\sigma [U_P] = \theta [1 - G(\sigma^*(b)|\theta)] [W - b], \tag{4}$$

which she maximizes over b. Let B denote 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" θ. Clearly, if b_1 and b_2 both belong to B, with $b_1 < b_2$, then

$$\sigma^*(b_1) > \sigma^*(b_2). \tag{5}$$

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

Proposition 1 *In an 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 trusting principal offers a lower bonus: if b_1 and b_2 are offered when the principal assesses the probability of success to be θ_1 and $\theta_2 < \theta_1$ respectively, then $b_1 \leq b_2$.

(iii) Rewards undermine the agent's self–confidence: for all (σ_1, σ_2) and equilibrium rewards $b_1 < b_2$,

$$\mathsf{E}[\theta|\sigma_1, b_1] > \mathsf{E}[\theta|\sigma_2, b_2].$$

Future self–confidence is also always reduced by an increase in the reward; that is, the expectation of θ conditional on σ, b, the action and the outcome is decreasing in b regardless of σ, the action and the outcome.

Proof. (i) has already been proved. 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 θ_i, $i = 1, 2$, and denote $\sigma_i = \sigma^*(b_i)$. Since b_i is optimal given θ_i, it must be that:

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

hence:

$$\frac{1 - G(\sigma_1 \mid \theta_1)}{1 - G(\sigma_2 \mid \theta_1)} \geq \frac{W - b_2}{W - b_1} \geq \frac{1 - G(\sigma_1 \mid \theta_2)}{1 - G(\sigma_2 \mid \theta_2)}.$$

Since $\theta_1 > \theta_2$, the MLRP requires that $\sigma_2 \leq \sigma_1$.[10] Thus $\sigma^*(b_2) \leq \sigma^*(b_1)$, 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.[11] Therefore, if the principal offers b_1 to types $\left[\underline{\theta}_1, \overline{\theta}_1\right]$ and $b_2 > b_1$ to types $\left[\underline{\theta}_2, \overline{\theta}_2\right]$, it must be that $\overline{\theta}_2 \leq \underline{\theta}_1$. This establishes part (iii) of the proposition. ■

Appendix 1 illustrates the computation of equilibrium in the case where θ takes only two values, θ_L and θ_H. In equilibrium, the principal offers no bonus ($b = 0$) to a more able agent ($\theta = \theta_H$), and randomizes between no bonus and a positive bonus when dealing with a less able agent ($\theta = \theta_L$).

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, $1 - G(\sigma^*(b) \mid \theta)$, and the probability of success, $\theta[1 - G(\sigma^*(b) \mid \theta)][W - b]$, are both increasing in θ, which is known only to the principal. Because θ varies negatively with b in equilibrium, the observer who simply correlates b with outcomes may conclude that rewards are not very effective. The reason is that such unconditional correlations or regressions fail to take into account the fact that the highest incentives are given to those who would otherwise be the least likely to work.[12]

- *Impact of rewards on intrinsic motivation.*

The literature on intrinsic versus extrinsic motivation refers to an apparently different argument: the subject finds the task less *attractive* when offered a reward. However, Proposition 1 holds unchanged when the principal *has private information about the attractiveness of the task*, rather than the probability of success. Let us assume that θ is symmetric information. By contrast, the principal knows from previous experience the cost c of undertaking the task, while the agent only has signal γ distributed according to

[10] The MLRP implies that $(1 - G(\sigma|\theta_1)) / (1 - G(\sigma|\theta_2))$ is increasing in σ, for all $\theta_1 > \theta_2$.

[11] There exists no pure strategy separating equilibrium. In such an equilibrium, the agent's behavior would not depend on his signal. The principal's preference over bonuses that induce compliance with probability 1 is the same for all θ (i.e., choose the lowest such bonus), and so some pooling must occur.

[12] In the two–type example, for instance, the observer will see the agent working with positive probability (perhaps even a relatively 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 ($\theta = \theta_L$), it does have a significant impact on motivation.

a cumulative distribution $G(\gamma \mid c)$ with the MLRP. An attractive task is one with a low c, and a high signal γ makes this more likely. The principal's objective function is then

$$\theta \left[1 - G\left(\gamma^*(b) \mid c\right)\right][W - b],$$

while the agent exerts effort if and only if

$$\theta (V + b) \geq \mathsf{E}[c \mid \gamma, b].$$

The same proof as above shows that a *higher reward is, in equilibrium, associated with a less attractive task;* therefore, bonuses reduce intrinsic motivation. Conversely, "forbidden fruits" are the most appealing ones. Note that, under either interpretation of the model, 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 –thereby "signalling" a pleasurable activity rather than a chore:[13]

> *"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... Tom... 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).

- *Paternalism: altruism towards a time–inconsistent agent.*

One 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

[13] We are grateful to Ilya Segal for reminding us of this example.

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 analyze here to manipulate the agent's perceptions of himself and of the tasks he faces –"for his own good".[14]

- *Retrospective justification.*

Combined with imperfect memory, the previous result has an interesting implication for situations where currently available information provides only insufficient justification for a certain course of action.[15] Suppose that in the future the agent 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, but not his intrinsic interest in the task (and the later observation of c). 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. (The situation envisioned is one where financial and career rewards are small). He may then reflect that since he chose to embark on this project after completing other, similar ones, the personal satisfaction enjoyed from previous completions (and which, at this later and perhaps somewhat depressed stage, he cannot quite recall) must have been significant. Hence it is worth persevering on the chosen path. The result that $E[c \mid \gamma, b] < E[c \mid \gamma, b']$ for $b' > b$ can provide an explanation for this kind of *ex–post rationalization* (Bem 1967, Staw 1977).

2.4 Empowerment and motivation

We showed earlier that the principal signals her trusts in the agent's ability (high θ) or intrinsic motivation (low c) through the use of a low–powered incentive scheme. We now investigate the use of delegation or empowerment to induce an agent to carry out the

[14] Formally, W in this case is equal to V/β, where $\beta \in (0,1)$ is the agent's quasi–hyperbolic discount factor (Strotz 1956, Phelps and Pollack 1968, Laibson 1997); equivalently $1/\beta$ measures the salience of the effort cost c for the agent, at the time when he must incur it.

[15] See Benabou and Tirole (1999) for a model of rational selective memory, awareness, or attention. The present argument requires only that memory be imperfect, especially with regard to one's past feelings and emotions (hedonistic payoffs).

objectives of the principal (Miles 1965).[16] In a nutshell, the principal demonstrates trust in the agent's ability (or, alternatively, his intrinsic motivation) by delegating control of the task to him, and thereby makes it more likely that the agent exerts effort.

In this section, we abstract from rewards by assuming a non–contingent compensation ($b = 0$). 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. These reduced forms could be derived from a situation where the principal decides to either relinquish some control rights to the agent, or put in place a monitoring technology or supervisor. As earlier, we assume that the principal knows the agent's probability of success θ, while the agent has a signal σ drawn from a cumulative distribution $G(\sigma \mid \theta)$, with density $g(\sigma \mid \theta)$ satisfying the monotone likelihood ratio property (2). The agent's utility is, as usual, $\theta V - c_d$, $d \in \{0, 1\}$. 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.

We make the following assumptions:

Assumption 1

$$\frac{d}{d\theta}\left(\frac{\mathcal{W}_1(\theta)}{\mathcal{W}_0(\theta)}\right) > 0 \quad \text{and} \quad \frac{\mathcal{W}_1(0)}{\mathcal{W}_0(0)} < 1 < \frac{\mathcal{W}_1(1)}{\mathcal{W}_0(1)}.$$

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 ($\theta = 1$). The assumption implies that there exists θ^* in $(0, 1)$ such that, under symmetric information, it is efficient to delegate the task if $\theta > \theta^*$, and not to delegate it if $\theta < \theta^*$.

[16] 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 won't interfere too much with his suggestions. In Dessein (1999), the principal delegates so as to enable the final decision to reflect the agent's information better than is the case when the agent communicates it strategically and the principal decides. Delegating to the agent the principal signals a greater *congruence* of their objectives. Salancik (1977) proposes yet another viewpoint, the "co-optation of personal satisfaction": "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."

Assumption 2 *For all* (σ_0, σ_1),

$$\frac{d}{d\theta}\left[\left(\frac{1-G(\sigma_1|\theta)}{1-G(\sigma_0|\theta)}\right)\left(\frac{\mathcal{W}_1(\theta)}{\mathcal{W}_0(\theta)}\right)\right] > 0.$$

Assumption 2 imposes an upper bound on the information effect arising from the correlation between the signals θ and σ received by the principal and the agent (and which was the focus of the previous section).[17] It substantially simplifies the analysis by guaranteeing that, in choosing whether to delegate, the principal is more concerned about the relative damage which the agent may cause when undertaking the task under each regime (Assumption 1), than about the potential impact of the delegation decision on the likelihood that the agent will undertake the task.

Proposition 2 *In equilibrium, under Assumptions 1 and 2:*
(i) Empowerment is good news for the agent about his ability, and therefore changes his attitude towards the task.
(ii) There is more empowerment than under symmetric information: the principal delegates if and only if $\theta > \theta^{**}$*, where* $\theta^{**} < \theta^*$.

Proof. (i) 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 the agent exerts effort if and only if $\sigma \geq \sigma_d^*$. The principal chooses $d \in \{0,1\}$ so as to maximize

$$[1 - G(\sigma_d^* \mid \theta)]\, \mathcal{W}_d(\theta).$$

Assumption 2 then implies that the principal delegates if and only if $\theta \geq \theta^{**}$ for some θ^{**}. Furthermore, $\sigma_1^* < \sigma_0^*$, for two reasons: first, delegation directly makes the task more attractive, by assumption ($c_1 \leq c_0$). Second, delegation is good news about ability.

(ii) Obvious. ■

[17] For example if $G(\sigma|\theta) = 1 - \exp\{-\sigma/(\theta + k)\}$ for $\sigma \in [0, +\infty)$, Assumption 2 amounts to imposing a lower bound on k.

Proposition 2 and its premise 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.*"[18]

Remark (when delegation backfires): Assumptions 1 and 2 are restrictive. To see this, suppose for instance that $\mathcal{W}_0$ and $\mathcal{W}_1$ are proportional to θ, $\mathcal{W}_1 = \theta W_1$ and $\mathcal{W}_0 = \theta W_0$, and that delegation is always costly: $W_1 < W_0$. Delegation to the agent is then equivalent to giving him a bonus of $c_1 - c_0$, at a cost to the principal of $\theta (W_0 - W_1)$. Following the steps of the proof of Proposition 1, one can show that, in equilibrium, delegation always represent *bad news* about the agent's ability .

This remark leads to a more general point. What we have said about rewards and empowerment applies more generally to any type of policy that is less costly to the principal when the agent is more talented. This *sorting condition* implies that the principal can demonstrate trust and boost the agent's self–confidence. The same effect may not hold when the principal simply encourages the agent by telling him how able he is and how pleasant the task can be. Unless the principal is known to have a strong aversion to lying or has built a reputation for honest appraisals, such encouragements are cheap talk and may be perceived by the agent as self–serving.

That the sorting condition is needed in order for the principal to boost the agent's self–confidence is further 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 to motivate 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 to comfort them.[19]

[18] Cited in Baron and Kreps (1999), who provide a further illustration at Hewlett-Packard.

[19] It would be interesting to assess in this light the evidence on the role of expectancies. 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' behavior changes, the students' self-confidence is unaffected (Darley and Fazio 1980).

3 Discussion and robustness

Let us now step back and discuss the relevance, scope, and methodology of our approach.

3.1 Relevance

We first return to the hidden cost of rewards. We showed that:

- Rewards impact intrinsic motivation. While under symmetric information the intrinsic (θV) and extrinsic (θb) motivations can be cleanly separated, under asymmetric information they cannot. In particular, the "intrinsic motivation" $\hat{\theta}(\sigma, b) V$ *decreases* with the level of the bonus. Similarly, when the agent does not know how costly or exciting the task is, his perception $\mathsf{E}[c \mid \gamma, b]$ is affected by the level of the reward.

- A reward is a positive reinforcer in the short–term, but always decreases future motivation.

While adopting an economic approach, our analysis is well in line with a branch of the social psychology literature. 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' behavior. Both approaches emphasize the information impact of rewards. As Deci (1975, p142) 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 interesting conjecture: 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. 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 would seem like a small price to pay to get the current job done. Properly exploring this idea will require an explicit dynamic model, however, and is therefore left for future work. Our current framework really corresponds to settings where the agent's successive interactions or "engagements" are with different principals (e.g., successive teachers or managers), or between myopic players.[20]

Our results are also consistent (in sign, perhaps not in magnitude) with Etzioni's (1971) claim that workers find control of their behavior via incentives "alienating" and "dehumanizing", with Kohn's (1993) argument that incentive schemes make people less enthusiastic about their behavior, and with Deci and Ryan's (1985) view that rewards change the locus of causality from internal to external and make employees bored, alienated, and reactive rather than proactive.

We should, however, issue two important caveats here. The first 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.*"

Our model may clarify the difference between what we would label "promised" or "'ex ante" contingent rewards, and "discretionary" or "ex post" rewards. The model of section 2 is about the *control* of behavior through rewards. The principal selects a reward for a

[20] In particular, if the agent and the principal are in a long–term relationship (the engagement and re–engagement take place with the same actors), there is a second ratchet effect to take into account: the agent may have an incentive to shirk at the initial stage, in order to signal that he has bad information (a low σ), and thereby elicit a higher bonus in the future.

well–defined 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 his self–esteem or intrinsic motivation, because of a different learning effect: the worker or the child learns from the reward that the task was generally 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 repeating this performance is worthwhile. Thus, giving ex post a bicycle to a hard–working child or a special pay raise or early promotion to a productive assistant professor will boost rather than hurt their self–confidence. The agent then does not infer that his behavior 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 did not know how to interpret his performance. The reward is then an indirect measure of performance for the agent.

The second caveat is that our economic analysis also unveils necessary conditions for rewards to have a negative impact on self–confidence. The first key condition is that the principal has *information* about the agent or the task that the agent does not have. This may explain why the existence of hidden costs of rewards is less controversial in an educational setting 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.). By contrast, the structure of rewards in the workplace is 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 this information 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 the case 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–pay job and given the leadership of a new project or division, together with a pay–for–performance scheme. In this example, which can be thought of as a convex combination of Sections 2.3 and 2.4, 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 thus boost the manager's self–confidence.

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.

3.2 Scope

Section 2 studied rewards and delegation as specific policies impacting the agent's self–confidence. Several other types of social interactions are worth studying. Some involve mere reinterpretations of the model, other unveil richer patterns:

a) *Help.*

Suppose that the principal offers to contribute a level of help h (at private cost h) in case the agent decides to undertake the task. This help improves the probability of the agent's 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 $\mathsf{E}\left[P(\theta, h) \mid \sigma^*(h), h\right] V = c$, and $\sigma^*(h)$ is a decreasing function. Ignoring rewards, the principal's payoff is

$$U_P = \left[1 - G(\sigma^*(h) \mid \theta)\right] \left[P(\theta, h)W - h\right].$$

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:

$$\frac{\partial^2 \ln\left(P(\theta, h)W - h\right)}{\partial\theta\partial h} < 0.$$

Intuitively, this means that help makes more of difference for weak agents than for strong ones. Following the steps in the proof of Proposition 1, one can show that in equilibrium, *a trusting principal helps less*: if $\theta_1 > \theta_2$ and h_i is an optimal level of help for type θ_i, then $h_1 \leq h_2$. Conversely, a high level of help is bad news for the agent, permanently weakening his intrinsic motivation for the task.

This observation may explain why help (like rewards or lack of delegation) can be detrimental to self–esteem. 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 et al. 1983, p233). 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 and 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").[21] 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/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 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).

b) *Disclosure of information.*

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.[22] Successful coaches are viewed as those who build up others' confidence (Kinlaw 1997). Although section 3.4 will argue that coaches may not always want to boost self–confidence, the complementarity between effort and talent makes it clear why even a selfish coach will often benefit from doing so. Formally, the principal's policy p is here the disclosure (or absence of disclosure) to the agent of information relevant to this self, that is, of a signal held by the principal and covarying with the agent's ability. The release of a signal

[21] 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.

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

covarying positively (negatively) with θ boosts (lowers) the agent's self–confidence. In the model of section 2, the principal indeed wants to release good signals and conceal bad ones.

Taking it for granted that the principal wants to boost the agent's self–esteem, it is interesting to note that, in some circumstances, the released signal may have ambiguous consequences. Suppose that the agent failed previously. The principal may then try to convince him that the link from talent and effort to performance is rather random (e.g., the probability of success was $\varepsilon\theta e$, where ε is noise) or, relatedly, that the agent was discriminated against. Offering such *excuses*[23] may sometimes prove self–defeating. 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 the agent.[24]

We have not yet discussed the credibility of the principal's disclosure. Credibility is no issue if the information is "hard" (i.e., if the agent can verify the veracity of the information). Often, however, the principal's information is "soft", in that the agent cannot verify it. The agent then may not believe proclaimed "good news" because he understands that the principal's ulterior motivation is to boost his self–esteem. The credibility of announcements with soft information may be restored, however, if the principal succeeds in building a reputation for not exaggerating her claims. By contrast, a professor who tells all her students and colleagues how great they are may do little to their egos.[25]

3.3 Methodology

Our approach, in the tradition of economics and some of cognitive psychology, focuses on the individual's motivation. An alternative viewpoint, along the lines of the behaviorist school (Hull 1943, Skinner 1953) would shunt the inner process and posit a direct link from stimulus to response. The agent in our context would just exhibit an instinctive, aversive reaction to being offered a contingent reward.

Unsurprisingly, we feel reluctant to adopt such a "reduced form" approach. Certainly,

[23] See Snyder et al. (1983) for a broad discussion of excuses.

[24] We are just saying that, while an outside observer could interpret the excuse as an exercise in confidence building (an increase in $\widehat{\theta}$), the relevant ability ($\varepsilon\theta$) includes the noise.

[25] See Baker et al (1997) for an illustration of the role of the principal's reputation in relationship contracts.

individuals do not constantly compute perfect Bayesian equilibria when trying to figure out the cognitive implications of their environment's actions. Indeed economists are content with the idea that individuals are boundedly rational and use rules of thumb and analogies in order to economize time and thinking, as long as the resulting inferences and behavior are not too much at odds with their self–interest. In particular, from casual experience, we feel that individuals are quite sophisticated at drawing inferences from the behavior of people they interact with (with some variations in the population[26]).

In our view, the cognitive/economics approach delivers two benefits. First, it helps understand *why* the response to the stimulus is the way it is. Second, it makes testable predictions as to *when* rewards may have a hidden cost and when the hidden cost of rewards is likely to be a myth. We refer the reader to section 3.2 for these predictions.

3.4 Undermining the other's ego

Our premise (and that of much of the human resources literature for example) has been that one benefits from boosting the self–esteem of one's spouse, child, colleague, coauthor, subordinate, or teammate. But while benefits from others' self–confidence are indeed a pervasive aspect of social interactions, it is also a fact that people often criticize or downplay the achievements of colleagues and relatives, and disclose information that is detrimental to their ego. The study in section 2 must therefore be part of a broader construct, in which the principal may sometimes benefit from repressing the agent's ego. This section lists some potential motivations for such behaviors.

a) *Direct competition.*

A rather trivial reason for why someone may want to bash another person's ego is that the two are in direct competition (for a job, a mate, a discovery, a title, and so forth). Then the former is directly hurt when the latter succeeds. In the context of our model, W is negative.

b) *The risk of "coasting".*

One of the basic equations in social psychology (and, consequently, our starting point in section 2) is that the marginal payoff to an individual's effort is increasing in his talent.

[26]For example, children exhibit different speeds of learning how to interpret social signals. And adults usually have more experience in the matter than children.

In certain situations, however, effort and ability are *substitutes* rather than complements, creating the risk that the agent may reduce effort when feeling more self–confident ("resting on his laurels"). This situation arises in particular when the agent's private payoff for performance is of a "pass–fail" nature. For example, a pupil whose only ambition is to pass an exam may cram less if he feels talented. Similarly, an individual who aims at little more than keeping his spouse and takes her for granted won't put much effort into being attractive to her.[27] The teacher or parent may then want to downplay the pupil's achievements, and the spouse may tell him that he is not that 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 and select 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 $\left(\widehat{\theta}\right)$ is (inverted) U–shaped, as opposed to constantly increasing as posited in section 2: an increase in the agent's self–esteem helps up to a point, where it 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 other agents.

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

3.5 Ego bashing and battles for dominance

Many circumstances in private life or at the workplace are characterized by power relationships. Egos clash as individuals try to establish dominance over each other along some

[27] 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.

dimension (e.g., intellectual). What matters in such situations is one's relative standing in the relationship, rather than any absolute standing. Shattering the other's self–confidence in the relevant dimension may then increase one's power in the relationship.

To illustrate this, consider a pair of individuals, 1 and 2. They must take a joint decision (they share the "formal control right" over the decision). Each comes with one idea or project, but only one project 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 B is natural, since individuals are more likely to search for (or reveal) ideas that favor them; B admits several interpretations: the task may be easier for individual i, may have positive spillovers over i's other activities, or bring him outside credit for having had the idea.

Let us assume for simplicity that θ_1 is known while θ_2 can take two values θ_2^L and θ_2^H, with $\theta_2^H > \theta_2^L$, and

$$\theta_2^H V + B > \theta_1 V > \theta_2^L V + B. \tag{6}$$

Individual 1 knows θ_2, while individual 2 does not. In our terminology, individual 1 can thus be viewed as the principal and individual 2 the agent, even though there is *no hierarchy* in terms of control rights. The principal has no hard information about θ_2 when $\theta_2 = \theta_2^H$, but when $\theta_2 = \theta_2^L$ she does, and can choose to disclose to the agent these "bad news" about his talent.[28] We rule out monetary transfers between the two individuals for simplicity. The timing goes as follows:

Stage 1: The principal learns θ_2 and (if $\theta_2 = \theta_2^L$) chooses whether to disclose the information.

Stage 2: Both come up with an 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 project offer, i.e., chooses the project.

It is easy to see that when $\theta_2 = \theta_2^L$ the principal wants to convey these bad news to the agent, because she thereby establishes dominance: from (6), even if the agent gets to

[28] Thus θ_2^H corresponds to "no bad news", and θ_2^L to individual 1 learning "bad news".

propose a course of action he then defers to the principal, which he would not do if he were more self–confident. By lowering the other's self–confidence, individual 1 enjoys *real authority* despite sharing *formal authority* over decisions with individual 2.[29]

The situation described above may still be viewed as a relatively tame and efficient version of the "battle of the egos", as 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 he may feel disappointed, but should recognize he is being saved from making a costly mistake. A slight variant of the model, however, can yield a much less harmonious and efficient outcome.

Suppose now that when $\theta_2 = \theta_2^L$ the principal has no hard information, but when $\theta_2 = \theta_2^H$ he does.[30] Let $\bar{\theta}_2$ denote the agent's initial self–confidence, i.e. his prior about his own ability in the absence of information. Finally, instead of (6), assume now that:

$$\theta_1 V + 2B > \theta_2^H V + B > \theta_1 V > \bar{\theta}_2 V + B, \tag{7}$$

The first inequality states that the principal would like to be in control even when $\theta_2 = \theta_2^H$. The second and third ones mean that the agent will not submit if he becomes aware that he is of high ability, but will yield if he remains uninformed. The principal will then systematically censor positive signals about the agent's ability, and would even be willing to spend resources in order to prevent them from reaching the agent. In contrast to the earlier case, the principal's undermining 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 seems to correspond 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).

The two types of ego–bashing behavior could also be combined into the same model, by allowing agent 1 to pay a cost in order to try and find out (with some probability) the value of θ_2. If he learns that $\theta_2 = \theta_2^L$ he will disclose it, but if he finds out that $\theta_2 = \theta_2^H$

[29] We thank Isabelle Brocas for suggesting the analogy with real authority and the Aghion and Tirole (1997) paper.

[30] Thus θ_2^H now corresponds to individual 1 learning "good news", and θ_2^L to "no good news".

he will stay mum, and feign ignorance. 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.

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 in other, more important tasks) due to time–inconsistent preferences, which results in his attaching negative value to information about his ego (Benabou and Tirole 1999). 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 its helps him by making sure that the efficient project is selected.

Ego–bashing may also have costs for the principal. As shown earlier, the agent's lack of self–confidence may reduce his initiative in coming up with (searching for) good projects initially, as well as his motivation for putting effort into the joint endeavor later on. Another cost may stem from individual 2's drawing more complex inferences about individual 1's ulterior motivation. Suppose that individual 2 cares not only about this project, but also about individual 1's altruism/friendship/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.

Nonetheless, individuals may often be willing to incur such costs in order to establish dominance. Because this may result in very inefficient outcomes, an interesting avenue for future research is how people try, in practice, 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 the allocation of 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 his self). 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.

4 Self–presentation: signaling one's self–esteem

4.1 Introduction

A large literature in social psychology addresses self–presentation, namely the set of behaviors and attitudes (self–promotion, excuse–making, supplication, intimidation, ingratiation, etc.) that are strategic and aim at manipulating others people's beliefs about oneself. For instance, Baumeister (1998) defines self–presentation as "attempts to convey information about or images of oneself to others". This topic has also been widely explored in economics, albeit with a very different range of applications, under the heading of signaling theory (Spence 1974).

While sections 2 and 3 were concerned with the principal's private information, this section focuses on the impact of self–knowledge on self–presentation, that is, on the relationship between the *"inner self"* and the *"outer self"*. We will of course not attempt to cover this huge territory here, and only wish to illustrate the formal approach on a specific issue, namely the strategic nature of the relationship between depressed individuals and their environment.

Depression has long been diagnosed as a disorder of self–esteem (Bibring 1953, Freud 1957). Its symptoms are well–known: poor self–image, inhibition of all activity, public admission of one's weaknesses, lack of interest in the outside world, low tolerance for frustration, etc. Of particular interest for our analysis are the self–esteem maintenance and self–presentation strategies used by depressed individuals to cope with their condition.[31]

First, depressed individuals sometimes "blackmail" others for attention. They want to verify that they are not unworthy nor unloved, and therefore look for sympathy and reassurance (Cohen 1954, Coyne 1976). Yet, paradoxically, they are often unreceptive to the positive feedback which others may offer (see, e.g., Hill et al. 1986). Finally, they are willing to incur disapproval costs in order to avoid demands to perform. That is, through acts and words, they confess their weakness in order to ask for leniency on a web of obligations, and attempt to lower others' standards or expectancies (Shaw 1982, Hill et al. 1986).[32]

[31] See, e.g., Hill et al (1986) and Snyder et al. (1983) for reviews.

[32] Relatedly, individuals who are not depressed may also ask for a milder form of moratorium, for example by reporting test anxiety to set up excuses for possible failure. It is interesting to note in this respect that self-reports of test anxiety do not occur when the incentives for success are high or when

These behaviors and attitudes can be analyzed from the perspective of our self–confidence model. Depressed individuals lack self–esteem (believe they have a low θ), and are therefore in search of good news about their self.[33] They realize, however, that the news may be favorable, but may also push them deeper into depression and helplessness. In economics parlance, depressed individuals are willing to "gamble for resurrection".

There are various ways of obtaining this self–relevant information. First, the individual may undertake activities to try and obtain reassurance about his self; but precisely because of the lack of self–confidence, this route is very costly, and indeed depressed individuals face inhibition in most activities ("what's the use?"). Second, the individual may attempt, usually with the help of others, to recover certain repressed or unconscious information about the self. Indeed, the general goal of Freudian theory and traditional psychoanalysis is to make the unconscious conscious. This strategy also has substantial costs. For example, an adult may get reassurance about his talent by remembering that he failed in school partly because of abuse by his parents. Third, the individual can turn to others and ask for encouragement, reassurance, and similar boosts to his self–esteem: "Will I ever make it? Do you still love me?", and so forth.

Alternatively, a depressed person may "admit defeat" by openly confessing (sometimes even exaggerating) his low self–worth, which then justifies asking for help or indulgence with respect to the tasks he face. Under this strategy, the depressed person calls for others to lower their expectations of his performance. Because of the existence of costs which we discuss later, however, neither leniency nor effective reassurance will come about easily. As we discussed in section 2, *a sorting condition on the helper's side* must be satisfied. In particular, the depressed face an attributional dilemma in evaluating positive responses from others (Wortman and Linsenmeier 1977). Due to the arousal of guilt and the presence of spillovers, people do not like their child, spouse, parent, friend or colleague to be depressed. They therefore have a vested interest in boosting his or her self–esteem, in order to reduce these costs imposed on them by the individual's depressed state and associated behavior. This, in turn, may explain why the depressed often question the

subjects are told that test anxiety does not affect performance on the particular test (see Greenberg et al. 1986 for an overview).

[33] This need for ego–relevant information is particularly acute when low self–esteem combines with time–inconsistent preferences (e.g., hyperbolic discounting) to undermine motivation, resulting in constant procrastination and lack of efficacy. See Benabou and Tirole (1999) for an analysis of individuals' demand for self-knowledge which incorporates this additional element.

optimistic feedback that is given to them; they want to make sure that the encouraging evaluations and other "pep talks" are genuine, and therefore are willing to challenge them.

Conversely, and provided that others are willing to incur costs to credibly help a depressed individual by boosting his morale and/or lowering their standards, *a sorting condition on the individual's side* must also hold . Otherwise, a person could always fake a depression in order to manipulate his environment to his own advantage. We shall now use the principal–agent framework to illustrate the logic of this last point in the specific context of standard–setting.

4.2 Standards and calls for a lowering of expectancies

4.2.1 Preliminaries: the costs and benefits of standards.

A student comes to his advisor's office and exposits an idea that reflects serious effort but only moderate promise. Should the advisor tell the truth or should she praise and encourage the student to pursue the idea and suggest some improvements? In this decision, the advisor implicitly selects a (loose or tough) standard. In this respect, she faces a choice similar to that facing parents setting school performance standards for their children, or managers setting work standards for their subordinates.

There are various ways of setting standards. One, suggested above, is to disclose information to the agent, telling him that his current performance is not acceptable: the payoff in the task pursued (V) is low ("you will never enter a good university or find a good job with such grades"), or the principal has high expectancies ("I will be disappointed if you don't enter a good university or don't get a good job", "your father would have wished you to be more ambitious"). A simpler form of setting standards is to limit the agent's choice ("you are not allowed to take this route"). For the sake of simplicity, this section describes standard–setting as the principal restricting the agent's choice set. More subtle standards, based on persuasion, would be worth studying as well.

Standards have costs and benefits: on the one hand, a standard forces or persuades the agent to align his goals with those of the principal. On the other hand, the agent's self–confidence may not be adequate for the lofty goal set by the principal.

Suppose there are two tasks. Task i involves private cost c_i to the agent, and, if completed successfully, yields V_i to the agent and W_i to the principal; either task yields

0 to both parties in case of failure. Task 1 is the easy task in that

$$0 < V_1 < V_2, \quad 0 < W_1 < W_2, \text{ and } \quad 0 < c_1 < c_2. \tag{8}$$

We further assume that:

$$\theta_1 \equiv \frac{c_1}{V_1} < \theta_2 \equiv \frac{c_2 - c_1}{V_2 - V_1} < 1. \tag{9}$$

The probability of success, θ, is the same in both tasks, for simplicity. In this section, the agent knows θ and the principal does not. The prior cumulative distribution of θ on $[0, 1]$ is denoted $F(\theta)$, with density $f(\theta)$. If left free by the principal to choose between the two tasks (or not exerting effort at all), the agent solves

$$\max \{0, \theta V_1 - c_1, \ \theta V_2 - c_2\}.$$

He thus shirks if $0 \leq \theta < \theta_1$, chooses task 1 if $\theta_1 \leq \theta < \theta_2$, and task 2 if $\theta_2 < \theta \leq 1$.

Suppose now that the principal can *forbid one of the tasks.* The first observation is that the principal never forbids the hard task. If she did, the only change in the agent's behavior would be that the agent would "select" task 1 rather than task 2 when $\theta \geq \theta_2$. But the principal's payoff in task 2, θW_2, is higher than that, θW_1, in task 1. Would the principal want to forbid task 1? With this standard, the agent exerts effort if and only if $\theta V_2 - c_2 > 0$, or

$$\theta \geq \frac{c_2}{V_2} \equiv \theta^*, \tag{10}$$

Note that $\theta_1 < \theta^* < \theta_2$. The tradeoff faced by the principal is that a standard makes types in $[\theta^*, \theta_2]$ *more ambitious*, but makes weaker types in $[\theta_1, \theta^*]$ *give up.* The principal wants to forbid task 1 if and only if the net gain from imposing the standard is positive:

$$S(\theta_1, \theta_2) \equiv \left(\int_{\theta^*}^{\theta_2} \theta f(\theta) d\theta \right) (W_2 - W_1) - \left(\int_{\theta_1}^{\theta^*} \theta f(\theta) d\theta \right) W_1 > 0. \tag{11}$$

Remark: A similar tradeoff might be present if the principal were to set a standard by disclosing "hard" information rather than by constraining task selection. Suppose that the principal discloses information about the benefit V_1 being low for the agent ("I don't

like this job for which you are studying", "A low performance won't do", "This project won't lead to a paper publishable in a first–rate journal"). Then, the agent may react by becoming more ambitious or else by being discouraged, and so a similar pattern holds.

4.2.2 Call for leniency

Let us assume that (11) is satisfied, so that the principal is eager to impose a standard. The agent prefers not to face such a constraint (at least weakly), and so may be willing to incur a cost to induce the principal to remove it. In practice, this cost may take several forms, from mild (procrastination, which raises the cost of task completion) to severe (drug use, self–mutilation). We will focus on another common one: self–deprecation. When asking for a lenient treatment or a helping hand, the agent *admits his weakness*; this admission has a negative impact on future relationships with the principal or with other parties, who will disapprove of the agent or will not turn to him for new interactions. As Hill et al. (1986, p. 219) argue in their discussion of depressive self–protection,

"by emphasizing his or her weakness or illness, the depressive, then, may risk short–term disapproval and may even deprecate his or her present accomplishments in order to avoid altogether future demands to perform, or at least to avoid the embarrassment that may result for unanticipated future negative performance outcomes (i.e., the depressive may risk a short–term loss of esteem in order to avoid any further losses). Unfortunately, the avoidance of future performance likely serves only to maintain the depressive's self–doubts and shaky self–confidence."

Similarly, Aronson and Carlsmith (1962) state that

> *"the individual may self–deprecate in order to lower expectancies by rejecting an unexpected success".*

We assume that the endogenous *self–deprecation cost* incurred by the agent is proportional to the difference between his expected ability conditional on accepting the standard, $\mathsf{E}_s[\theta]$, and his expected ability conditional on asking for a leniency, $\mathsf{E}_m[\theta]$ (where "m" stands for "moratorium"), with coefficient of proportionality μ:

$$\mu\left(\mathsf{E}_s[\theta] - \mathsf{E}_m[\theta]\right). \tag{12}$$

While we could entertain alternative functional forms (the appropriate one depending on the exact nature of future interactions), the important features of this self–deprecation cost are that: a) the cost is associated with the change in others' perceptions of the agent's self, and b) this cost is *endogenous*, and depends on rational attributions by others.[34]

Note, on the other hand, that we restrict this inference to be as simple as possible: it reflects only whether or not the individual asks to be allowed to perform the easy task, and not his performance in the task ultimately chosen. This can be rationalized by assuming that outcomes are revealed only at a later stage, after the principal and other observers have already made their decisions concerning future interactions (or lack thereof) with the agent. Note also that an individual whose ability is so low that he would not exert effort under either regime will not request the lower standard, but pool instead with high ability agents (until his performance is ultimately observed), by nominally choosing task 2 but putting zero effort into it. By focussing on this particular timing of signals and decisions we are, once again, leaving to further research the more complex learning issues which would arise in a truly dynamic model.

Given the assumptions stated above, we thus analyze a simple two–stage game:[35]

Stage 1: *The agent asks or does not ask for a lowering of expectancies with respect to his performance.*

Stage 2: *The principal chooses whether to impose (or maintain) the standard.*

We will first look at a pure–strategy perfect Bayesian equilibrium, then later on discuss mixed strategies.[36]

[34] In particular, the agent could offer a "bribe" to the principal (money, favors, friendship) in exchange for the moratorium. By doing so he would signal that he is of a relatively low type, and thus bear a cost qualitatively similar to (12), in addition to the direct cost of the bribe.

[35] The model described in this section is one of "cheap talk", as in Crawford and Sobel (1982). Its payoff structure, however, differs from that of Crawford and Sobel, where a monotonicity condition implies that the sender's message reveals that his type belongs to a certain interval. In our model, by contrast, pooling will generally occur over a non–connected set: agents with very high and very low ability will not ask for a lowering of standards, while those in some intermediate range will. See Lemma 1 below.

[36] We shall ignore the usual "babbling equilibrium" in which the agent's request is simply ignored and, conversely, the request is totally uninformative about his type.

Lemma 1 *In an equilibrium in which there is a positive probability that the principal lowers the standard, there exists* $\widehat{\theta}_1$ *and* $\widehat{\theta}_2$*, with*

$$\theta_1 < \widehat{\theta}_1 < \theta^* < \widehat{\theta}_2 < \theta_2,$$

such that the agent asks for leniency (lifting of the standard) if and only if his self–confidence lies in $\left[\widehat{\theta}_1,\, \widehat{\theta}_2\right]$.

Proof. First, note that asking for a lenient treatment, or extra help, must be associated with a reputation cost ($\mathsf{E}_s[\theta] > \mathsf{E}_m[\theta]$); otherwise everyone would ask for it, since the agent always prefers to have more choice, ceteris paribus. But then the principal would always impose the standard, by condition (11).

Second, the agent has no incentive to ask for leniency if he ends up not exerting effort or choosing task 2, since the option of choosing task 1 then has no benefit, while the plea for indulgence has a positive self–deprecation cost. In particular, types in $[0, \theta_1]$ and $[\theta_2$,1] have a strict preference for accepting the standard.

Third, the agent solves

$$\max\left\{0, \theta V_1 - c_1 - \mu\left(\mathsf{E}_s[\theta] - \mathsf{E}_m[\theta]\right), \theta V_2 - c_2\right\}. \tag{13}$$

This shows that the set of types asking for a lowering of expectancies is an interval $\left[\widehat{\theta}_1,\, \widehat{\theta}_2\right]$, as described in the lemma. ■

The lemma implies that

$$\mathsf{E}_s[\theta] - \mathsf{E}_m[\theta] = D\left(\widehat{\theta}_1, \widehat{\theta}_2\right) \equiv \frac{\int_0^{\widehat{\theta}_1} \theta f(\theta)\, d\theta + \int_{\widehat{\theta}_2}^1 \theta f(\theta)\, d\theta}{F\left(\widehat{\theta}_1\right) + 1 - F\left(\widehat{\theta}_2\right)} - \left(\frac{\int_{\widehat{\theta}_1}^{\widehat{\theta}_2} \theta f(\theta)\, d\theta}{F\left(\widehat{\theta}_2\right) - F\left(\widehat{\theta}_1\right)}\right). \tag{14}$$

The self–deprecation cost is $\mu\left(\mathsf{E}_s[\theta] - \mathsf{E}_m[\theta]\right) = \mu D\left(\widehat{\theta}_1, \widehat{\theta}_2\right)$.[37] The lemma also implies that an equilibrium with a positive probability of a plea for indulgence is fully described by the following two equations in two unknowns $(\widehat{\theta}_1,\ \widehat{\theta}_2)$:

[37] In the case of a uniform distribution ($F(\theta) = \theta$), for instance, $D\left(\widehat{\theta}_1,\, \widehat{\theta}_2\right) = \frac{1}{2}\left(\frac{1-\widehat{\theta}_1-\widehat{\theta}_2}{1+\widehat{\theta}_1-\widehat{\theta}_2}\right)$.

$$\mu D\left(\widehat{\theta}_1,\,\widehat{\theta}_2\right)=\widehat{\theta}_1 V_1-c_1, \tag{15}$$

$$\mu D\left(\widehat{\theta}_1,\,\widehat{\theta}_2\right)=\widehat{\theta}_2\left(V_1-V_2\right)-\left(c_1-c_2\right), \tag{16}$$

which define the indifference points in (13). Finally, it must be the case that the plea for leniency is effective. That is, the gain for the principal of lifting the standard and inducing types in $\left[\widehat{\theta}_1,\,\theta^*\right]$ to undertake task 1 rather than doing nothing must be greater than the loss associated to the switch from task 2 to task 1 by types in $\left[\theta^*,\,\widehat{\theta}_2\right]$. Thus, whereas in the absence of information about θ the net value of imposing the standard was positive ($S(\theta_1,\theta_2)>0$ in (11)), it must now be negative:

$$S\left(\widehat{\theta}_1,\,\widehat{\theta}_2\right)=\left(\int_{\theta^*}^{\widehat{\theta}_2}\theta f\left(\theta\right)d\theta\right)\left(W_2-W_1\right)-\left(\int_{\widehat{\theta}_1}^{\theta^*}\theta f\left(\theta\right)d\theta\right)W_1<0. \tag{17}$$

An equilibrium is thus a solution $\left(\widehat{\theta}_1,\,\widehat{\theta}_2\right)$ to (15)–(16), or equivalently to

$$\widehat{\theta}_2=\frac{c_2-\widehat{\theta}_1 V_1}{V_2-V_1}, \tag{18}$$

$$\widehat{\theta}_1 V_1-c_1=\mu D\left(\widehat{\theta}_1,\,\widehat{\theta}_2\right), \tag{19}$$

with $\theta_1<\widehat{\theta}_1<\theta^*$, and such that (17) is satisfied. Note from condition (18) that, as $\widehat{\theta}_1$ increases from θ_1 to θ^*, $\widehat{\theta}_2$ decreases from θ_2 to θ^*.

We shall now solve for the equilibrium (and verify that (17)–(19) are consistent with the earlier assumption (11)) in the case where μ is relatively small, meaning that self–deprecation is relatively cheap. As (18)–(19) make clear, the equilibrium thresholds $\left(\widehat{\theta}_1,\widehat{\theta}_2\right)$ are then uniquely determined and close to the cutoffs (θ_1,θ_2) corresponding to the no–standard case. Indeed, a simple Taylor approximation yields:

$$\widehat{\theta}_1 \;\approx\; \theta_1+\mu D\left(\theta_1,\,\theta_2\right)/V_1 \tag{20}$$

$$\widehat{\theta}_2 \;\approx\; \theta_2-\mu D\left(\theta_1,\,\theta_2\right)/\left(V_2-V_1\right), \tag{21}$$

where $\approx$ means that we neglect terms of higher order. Recall from (14) that μD is the cost of self–deprecation, which must be positive. From here on we shall therefore assume that

$D(\theta_1,\,\theta_2) > 0$. As intuition suggests, the range of types $\left[\widehat{\theta}_1,\,\widehat{\theta}_2\right]$ who ask to be allowed to perform the easier task is smaller: a) the higher the implied signalling cost $\mu D(\theta_1,\,\theta_2)$; b) the lower the relative payoff to that task compared to the alternative chosen under the standard (the smaller V_1 for those with relatively low ability; the higher $V_2 - V_1$ for those with relatively high ability).

The last step is to verify that the equilibrium pair satisfies $S\left(\widehat{\theta}_1,\,\widehat{\theta}_2\right) < 0$ (condition (17)), while making sure that this is compatible with the previous requirement that $S(\theta_1,\,\theta_2) > 0$ (condition (11)). To that effect, we shall assume that:

$$\left(\frac{W_2 - W_1}{W_1}\right)\left(\frac{V_1}{V_2 - V_1}\right) > \frac{\theta_1 f(\theta_1)}{\theta_2 f(\theta_2)}. \tag{22}$$

Proposition 3 *Assume that $D(\theta_1,\theta_2) > 0$, that (22) holds, and that $S(\theta_1,\,\theta_2)$ is positive but relatively small. Then there exist $\underline{\mu}$ and $\overline{\mu}$ such that for all $\mu \in \left[\underline{\mu},\overline{\mu}\right]$, there is a unique pure strategy PBE, $(\widehat{\theta}_1,\widehat{\theta}_2)$, with positive probability of effective self–deprecation. As μ increases, $\widehat{\theta}_1$ rises and $\widehat{\theta}_2$ falls.*

Proof. For small μ, a Taylor expansion using (20)–(21) yields:

$$S\left(\widehat{\theta}_1,\,\widehat{\theta}_2\right) \approx S(\theta_1,\,\theta_2) - \mu D(\theta_1,\,\theta_2)\left[\theta_2 f(\theta_2)\left(\frac{W_2 - W_1}{V_2 - V_1}\right) - \theta_1 f(\theta_1)\left(\frac{W_1}{V_1}\right)\right]. \tag{23}$$

By assumption (22) the term in square brackets, which will be denoted ω, is strictly positive. Therefore, $S\left(\widehat{\theta}_1,\,\widehat{\theta}_2\right) > 0$ for all μ greater than $\underline{\mu} \equiv S(\theta_1,\,\theta_2)/\omega D(\theta_1,\,\theta_2) << 1$ but still small enough for the Taylor approximations to be valid. ■

Does there exist a mixed–strategy equilibrium in which the principal lifts the standard only with probability $x \in (0,1)$ when receiving a call for leniency? Such a mixed–strategy equilibrium is characterized in the following way: a) the left–hand sides of (15) and (16) are both multiplied by x, so that the function D is replaced everywhere by D/x; intuitively, the principal's mixed strategy amplifies the cost of self–deprecation; b) condition (17) is satisfied with equality. Equating the right–hand–side of (23) to zero, one easily sees that for $\mu \in \left[\underline{\mu},\overline{\mu}\right]$ the only equilibrium is the one in pure strategies described in Proposition 3, while for $\mu < \underline{\mu}$ the only equilibrium involves the principal using the mixed–strategy $x = \mu/\underline{\mu}$.

4.3 Humbleness as posturing

"Don't be so humble – you are not that great."

Golda Meir (1898–1978) to a visiting diplomat.

In the situation described earlier, the agent self–deprecates in order to benefit from lower expectancies. In other situations humility may, on the contrary, be a way of signalling a high level of self–confidence. A researcher who insistently "pushes" his work on others may be perceived as insecure about the value of his contribution or professional recognition; as a result, the pitch may backfire. By contrast, one who is more humble in his self–presentation may thereby reveal that he knows (perhaps from his track record) the quality of his work to be such that it will, sooner or later, "speak for itself". Similarly, in professional and social interactions, very wealthy people can afford to "dress down"or even look grungy, and very famous people will eschew boasting and name–dropping. Both types are confident that their interlocutors will find out soon enough whom they are dealing with.

The signalling approach provides a simple explanation for such behaviors. Suppose that the agent knows the quality of his work, θ, and that a higher θ makes it more likely that the principal will later on receive a signal that the work is of high quality. Assume further that the agent may already have a piece of hard information that reflects positively on his work (whether accurately or not), but is less informative than the signal to be received later on. Upon meeting the principal, the agent chooses whether or not to disclose the favorable hard information.[38] As long as this disclosure is costly (it takes time or resources, or may involve criticizing someone else's work, which is unpleasant), the following behavior may emerge in equilibrium: a sufficiently self–confident agent does not disclose the information even when he has it, whereas an agent with lower self–confidence always does. The sorting condition results from the fact that the self–confident type knows the costly disclosure to be unnecessary, since a favorable signal is likely to be received later on anyway.[39]

[38]Keeping with the researcher example, he may or may not mention that his paper was cited by some famous and highly–regarded colleague, that it is under revision at some prestigious journal, etc.

[39]To completely rationalize the Golda Meir repartee, one may also allow the principal to have independent information about the agent's quality.

5 Concluding remarks

Psychologists, experts in human resource management and sociologists have long emphasized the central role played by self–esteem and self–perception in personal motivation and social interactions. People are quite capable at drawing inferences about their self from others' behavior, and at analyzing the impact of their own actions on others' feelings. This side of social psychology has been largely neglected by economists. Yet the tools of economic theory can help us understand that the strategies of social interaction emphasized there are often quite rational, and analyze when these strategies are effective or backfire.

Rather than stating again the main results of the paper, we would like to indicate a few avenues of research that we feel are particularly interesting. The first avenue is a more systematic investigation of the many strategies related to the looking–glass self (sections 2 and 3) and to self–presentation (section 4). The second will consist in combining the two forms of signaling.[40] The third relates to the long–term dynamics of self–confidence in dyads (two–person relationships), particularly the learning and ratchet effects suggested by our analysis. Fourth, while our modeling currently accommodates the possibility of feelings of altruism/friendship/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 behavior subtle inferences not only about abilities and task characteristics, but also about how much the other cares about them. Rich dynamics in the relationship might ensue. Last, the analysis ought to be extended to groups. One hears frequent complaints about workplace settings 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 positive questions, (e.g., whether these strategies are mutually reinforcing), as well issues of institutional design related to the optimal organization of educational and work environments.

[40] 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 oppose it. In so doing, the principal reveals information to the agent, which impacts his subsequent behavior.

APPENDIX 1

To further illustrate the hidden cost of rewards, let us derive equilibrium behavior in the two–type case. The agent's ability may be high, θ_H, with probability f_H, or low, θ_L, with probability f_L. Let b_k^*, $k \in \{L, H\}$, denote the minimum bonus that induces compliance when the agent is fully informed about his ability:

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

and assume

$$0 = b_H^* < b_L^* < W.$$

The conditional densities of the agent's signal are denoted $g_H(\sigma)$ and $g_L(\sigma)$, and the MLRP simply means that g_H/g_L is increasing. Let us assume that this likelihood ratio has full support $(0, +\infty)$.

To pin down the equilibrium, let us adopt the common refinement that the monotonicity of beliefs with respect to the signal (bonus), which we showed must necessarily hold on the equilibrium path, also holds off the equilibrium path. This condition implies that the lowest equilibrium bonus is equal to 0. On–the–equilibrium–path monotonicity, in turn, implies that when $\theta = \theta_H$ the principal offers no bonus in equilibrium. Any other equilibrium bonus therefore fully reveals that the principal has information θ_L, and therefore such a bonus must necessarily be b_L^*. Next, note that when $\theta = \theta_L$ the principal must randomize between $b = 0$ and $b = b_L^*$: if she played only b_L^*, then $b = 0$ would induce compliance with probability 1, and therefore would be preferred to b_L^*. The equilibrium is then described by two parameters: $x^* \in (0, 1]$, the probability that a principal observing type θ_L selects bonus 0 (pools), and σ^*, the cut–off signal when bonus 0 is offered. These parameters are given by:

$$\theta_L\left(W - b_L^*\right) = \theta_L\left[1 - G_L(\sigma^*)\right]W \tag{24}$$

and

$$\left(\frac{f_H g_H(\sigma^*)}{f_H g_H(\sigma^*) + f_L g_L(\sigma^*)x^*}\theta_H + \frac{f_L g_L(\sigma^*)x^*}{f_H g_H(\sigma^*) + f_L g_L(\sigma^*)x^*}\theta_L\right)V = c. \tag{25}$$

The second equation can be rewritten as

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

so by the above assumptions on the likelihood ratio, for any $x^* > 0$ there exists a unique solution $\sigma^* = s(x^*)$ to (25), with $s' > 0$. Substituting into (24), the principal's net incentive to offer bonus 0 when $\theta = \theta_L$ and he is expected to randomize with probability x^* is:

$$\theta_L \left[1 - G_L(s(x^*))\right] W - \theta_L \left(W - b_L^*\right).$$

This function is increasing in x^*, and negative at $x^* = 0^+$. So either the function has a unique zero on $(0, 1)$,which then defines the principal's mixing strategy; or else the function is non–positive on all of $(0, 1]$, in which case the principal's equilibrium strategy is $x^* = 1$, which means 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). ■

References

[1] Aghion, P. and J. Tirole (1997) "Formal and Real Authority in Organizations," *Journal of Political Economy*, 105: 1–29.

[2] Aronson, E. and Carlsmith, J.M. (1962) "Performance Expectancy as a Determinant of Actual Performance," *Journal of Abnormal Psychology, 65: 178–182.*

[3] Baker, G., Gibbons, R. and K. Murphy (1997) "Relationship Contracts and the Theory of the Firm," mimeo, MIT.

[4] Baron, J. and D. Kreps (1999) *Strategic Human Resources*, New York: John Wiley.

[5] Baumeister, R. (1998) "The Self," in *The Handbook of Social Psychology*, edited by D. Gilbert, S. Fiske and G. Lindzey, Boston: McGraw–Hill.

[6] Benabou, R. and J. Tirole (1999) "Self–Confidence: Intrapersonal Strategies," mimeo, IDEI, Toulouse and NYU, June.

[7] Bem, D. (1967) "Self–Perception: An Alternative Interpretation of Cognitive Dissonance Phenomena," *Psychological Review,* 74: 183–200.

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

[9] Cameron, J. and D. Pierce (1994) "Reinforcement, Reward, and Intrinsic Motivation: A Meta–Analysis," *Review of Educational Research*, 64: 363–423.

[10] Carrillo, J., and T. Mariotti (1997) "Strategic Ignorance as a Self–Disciplining Device," forthcoming, *Review of Economic Studies.*

[11] Cohen, M. (1954) "An Intensive Study of Twelve Cases of Manic–Depressive Psychosis," *Psychiatry*, 17: 103–137

[12] Condry, J. and J. Chambers (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.

[13] Cooley, C. (1902) *Human Nature and the Social Order*, New York: Scribner's.

[14] Coyne, J. (1976) "Toward an Interactional Description of Depression," *Psychiatry*, 39: 24–40.

[15] Crawford, V. and Sobel, J. (1982) "Strategic Information Transmission," *Econometrica*;, 50(6), November: 1431–51.

[16] Darley, J. and R. Fazio (1980) "Expectancy Confirmation Processes Arising in the Social Interaction Sequence," *American Psychologist*, 35: 867–881.

[17] Deci, E. (1975) *Intrinsic Motivation*, New York: Plenum Press.

[18] Deci, E., and R. Ryan (1985) *Intrinsic Motivation and Self–Determination in Human Behavior*, New York: Plenum Press.

[19] Dessein, W. (1999) "Authority and Communication in Organizations," ULB mimeo.

[20] Eisenberger, R., and J. Cameron (1996) "Detrimental Effects of Reward: Reality or Myth?," *American Psychologist*, 51: 1153–1166.

[21] Etzioni, A. (1971) *Modern Organizations*, Englewood Cliffs, N.J.: Prentice–Hall.

[22] Gibbons, R. (1997) "Incentives and Careers in Organizations,' in D. Kreps and K. Wallis, eds., *Advances in Economic Theory and Econometrics*, vol.II, Cambridge University Press.

[23] Greenberg, J., Pyszczynski, T. and S. Solomon (1986) "The Causes and Consequences of a Need for Self–Esteem: A Terror ManagementTheory," in *Public Self and Private Self*, ed. By R. Baumeister, New York: Springer Verlag.

[24] Hill, M., Weary, G. and J. Williams (1986) "Depression: A Self–Presentation Formulation," in *Public Self and Private Self*, ed. By R. Baumeister, New York: Springer Verlag.

[25] Holmström B. and P. Milgrom (1991) "Multi–Task Principal–Agent Analyzes: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics and Organization*, Vol.7, special issue, 24–52.

[26] Hull, C.L. (1943) *Principles of Behavior*, N.Y.: Appleton–Century–Crofts.

[27] Kinlaw, D. (1997) *Coaching: Winning Strategies for Individuals and Teams*, Gower Publishing, U.K.

[28] Kohn, A. (1993) *Punished by Rewards*, New York: Plenum Press.

[29] Korman, A.K. (1970) "Toward an Hypothesis of Work Behavior," *Journal of Applied Psychology*, 54: 31–41.

[30] —— (1971) "Expectancies as Determinants of Performance," *Journal of Applied Psychology*, 55: 218–222.

[31] Kreps, D.(1997) "Intrinsic Motivation and Extrinsic Incentives, " *American Economic Review*, 87(2): 359–364.

[32] 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.

[33] Kruglanski, A., Friedman, I. and G. Zeevi (1971) "The Effect of Extrinsic Incentives on Some Qualitative Aspects of Task Performance," *Journal of Personality*, 39: 608–617.

[34] Laffont J–J. and J. Tirole (1988) "Repeated Auctions of Incentive Contracts, Investment and Bidding Parity, with an Application to Takeovers," *Rand Journal of Economics*, 19: 516–537.

[35] Laibson, D. (1997) "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, 112: 443–478.

[36] Lazear, E. (1996) "Performance, Pay and Productivity,' mimeo.

[37] Lepper, M. and D. Greeene (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.

[38] Lepper, M., Greene, D., and R. Nisbett (1973) "Undermining Children's Interest with Extrinsic Rewards: A Test of the 'Overjustification Hypothesis'," *Journal of Personality and Social Psychology*, 28: 129–137.

[39] Luthans, F., and R. Kreitner (1985) *Organizational Behavior Modification and Beyond*, London: Scott, Foresman and Co.

[40] Merton, R. (1948) "The Self–Fulfilling Prophecies," *Antioch Review*, 8: 193–210.

[41] Miles, R. (1965) "Human Relations and Human Resources," *Harvard Business Review*, July/August.

[42] Pfeffer, J. (1994) *Competitive Advantage Through People: Problems and Prospects for Change*, Boston: Harvard Business School Press.

[43] Phelps, E. and Pollack, R. (1968) "On Second–Best National Savings and Game–Equilibrium Growth," *Review of Economic Studies, 35: 185–199.*

[44] Rosenthal, R. and L. Jacobson (1968) *Pygmalion in the Classroom*, Holt–Rinehart–Winston.

[45] Schwartz, B. (1990) "The Creation and Destruction of Value," *American Psychologist*, 45: 7–15.

[46] Skinner, G.F. (1953) *Science and Human Behavior*, NY: MacMillan.

[47] Snyder, C., Higgins, R., and R. Stucky (1983) *Excuses: Masquerades in Search of Grace*, New York: John Wiley.

[48] Spence, M. (1974) *Market Signaling*, Cambridge, Mass.: Harvard University Press.

[49] 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.

[50] Steers, R., and L. Porter (1975) *Motivation and Work Behavior*, McGraw–Hill.

[51] Strotz, R. (1956) "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies*, 23: 165–180.

[52] Wilson, T., Hull, J. and J. Johnson (1981) "Awareness and Self–Perception: Verbal Reports on Internal States," *Journal of Personality and Social Psychology*, 40: 53–71.

[53] Wortman, C., and J. Linsenmeier (1977) "Interpersonal Attraction and Techniques of Ingratiation in Organizational Settings," in B. Staw and G. Salancik, eds, *New Directions in Organizational Behavior*, Chicago: St Clair Press.