MIT Sloan School of Management

MIT Sloan Working Paper 4547-05

June 2005

Too Motivated?

Eric Van den Steen

Too Motivated?

Eric Van den Steen[*]

June 10, 2005

Abstract

I show that an agent's motivation to do well (objectively) may be unambiguously bad in a world with differing priors, i.e., when people openly disagree on the optimal course of action. The reason is that an agent who is strongly motivated is more likely to follow his own view of what should be done. As a result, the agent is more willing to disobey his principal's orders when the two of them disagree on the right course of action.

This effect has a number of implications. First of all, agents who are subject to authority will have low-powered incentive pay. Second, intrinsically motivated agents will be more likely to disobey and less likely to be subject to authority. Firms with intrinsically motivated agents will need to rely on other methods than authority for coordination. Moreover, an increase in intrinsic motivation may decrease all players' expected utility, so that it may be optimal for a firm to look for employees with low intrinsic motivation. Finally, subjective performance pay may be optimal, even when the true outcome of the project is perfectly measurable and contractible.

Through this analysis, the paper identifies an important difference between differing priors and private benefits (or private information): with differing priors, pay-for-performance can *create* agency problems rather than solving them.

1 Introduction

Motivation is supposed to be a good thing: motivated people work hard and exert themselves to make good decisions. The purpose of this paper is to show that motivation can also be unambiguously bad. In particular, I will show that intrinsic and extrinsic motivation may create, rather than solve, agency problems when people openly disagree on the right course of action.

The idea is as follows. Consider a principal-agent setting in which principal and agent openly disagree on what course of action will lead to a success, i.e., they have differing priors. If the agent is paid a fixed wage and has no intrinsic motivation, then he is willing to follow the principal's orders with minimal inducement, even when he believes that these orders are the wrong thing to do. If, on the contrary, the agent has high intrinsic or extrinsic motivation to achieve a success, and if he disagrees with the principal's order, then he will be very tempted to disobey his principal. Motivation thus creates an agency problem.

[*] MIT-Sloan School of Management (evds@mit.edu). I thank Bob Gibbons for his extensive help and support, Bengt Holmstrom, John Roberts, Paul Oyer, and Ravi Singh for the detailed feedback and conversations, and Joe Chen, Jerker Denrell, Mathias Dewatripont, Oliver Hart, Bentley MacLeod, Jean Tirole, and the seminar participants at Harvard-MIT, INSEAD, Toulouse, ULB, USC, and the CRES Strategy Conference for the many useful suggestions.

I derive these results in a model where the agent has to choose between two actions, X and Y, and the principal and agent openly disagree on which action is most likely to succeed. After the players contract on a wage and a share of the revenues for the agent, the agent chooses his action. With some probability, the principal can then overturn the agent's decision, at a cost to both the agent and the principal. A central issue in the analysis is whether or not the agent 'obeys' the principal, i.e., whether or not the agent does what the principal wants him to do, instead of following his own beliefs.

The paper then shows the following. The agent is less likely to obey as he has higher-powered incentive pay, more intrinsic motivation (modelled as a private benefit from success), and stronger beliefs. Two equilibrium regimes emerge: one in which the agent has high-powered incentive pay and disregards what the principal wants him to do, and another in which the agent has low-powered incentive pay and obeys the principal (most of the time). Using authority in the sense of 'the power or right to give orders and enforce obedience' (Concise Oxford English Dictionary), I will then say that the principal has (interpersonal) authority over the agent.[1]

The model then makes the following predictions.

1. People with high-powered incentives or with strong intrinsic motivation are less likely to obey orders.

2. Employees who are subject to authority will typically have low-powered incentives.

3. People with high intrinsic motivation are less likely to be subject to authority

4. An increase in the agent's intrinsic motivation may lead to a Pareto inferior outcome.[2]

5. Subjective bonuses may be optimal, even when true performance is perfectly measurable and contractible. Subjective bonuses will go together with authority.

The model thus explains or predicts the following observations: employees typically have low-powered incentives, salespeople on commission are difficult to manage or control, subjective bonuses are used much more within firms than between firms, and some firms avoid people with strong views. An interesting consequence is also that firms with intrinsically motivated people will need to rely on mechanisms other than authority to coordinate their employees, such as 'hiring for fit' or 'socialization'.

An important aspect of this analysis is that differing priors are a necessary ingredient to obtain these results. Even stronger: differing priors and private benefits lead to *opposite* conclusions in

[1] Merriam-Webster Online defines authority as 'power to influence or command thought, opinion, or behavior', which is also consistent with its use in this paper. In particular, I will say that the principal has (interpersonal) authority when she can tell the agent what to do and the agent obeys with positive probability (but would have done something different if it weren't for the principal's order). While this is consistent with the use of 'authority' by, for example, Simon (1951), other parts of the economic literature, such as Aghion and Tirole (1997), have used authority more in the sense of a 'right to make an (impersonal) decision', such as setting a salary or allocating budget to a project. Such decisions, however, often need to get implemented by other players, in which case there is an implicit assumption that the principal also has interpersonal authority over these people. Note that the process of giving orders is eliminated from the main model, but is made explicit in appendix B.

[2] Efficiency comparisons in this paper are based on subjective expected utility of the players. It is this measure of utility that determines what contract the players negotiate, and thus what we will observe. However, it seems that many of the efficiency effects would continue to hold if we used reference beliefs to measure utility and considered a more extended model with coordination issues. For a discussion of different ways do measure utility in models with differing priors, see Van den Steen (2005).

this context. While giving an agent residual income reduces agency problems caused by private benefits, it exacerbates the agency problems caused by differing priors.

Some of the paper's implications, especially the prevalence of low-powered incentives in firms, have been derived in other contexts. The most important theory in this respect is that of multi-tasking. Holmstrom and Milgrom (1991), in their seminal contribution, argue that pay for performance on one activity can reduce effort on other activities that compete for attention, and as a consequence may bias effort towards more measurable activities. Holmstrom and Milgrom (1994) argue that low-powered incentives will go hand-in-hand with the ability to forbid employees to engage in outside activities. These multi-tasking models differ clearly from the current paper. Multi-tasking is about the relative measurability of different aspects of the overall outcome, and how giving incentives for one aspect may inefficiently reduce effort on other aspects that compete for attention. The main model of this paper, on the contrary, has only one task, one decision, and only one outcome.[3] Moreover, authority and its enforcement, which are among the core elements of this paper, play no real role in the multi-tasking literature. Further differences include the absence of differing priors in the multi-tasking models and the absence of risk aversion in the current model. The predictions also differ. Multi-tasking has no predictions on intrinsic motivation or disobedience, while the current theory has no implications on the impact of relative task measurability.

There are also two arguments for low-powered incentives in firms that are based on more traditional agency theory. First, if authority and pay-for-performance are substitute mechanisms to get the agent to do the right thing, and each carries a fixed cost, then there will be a tendency to use only one of the two. This idea has been present informally in the literature on franchising, such as Brickley and Dark (1987) or Martin (1988), but has recently been formalized by Prendergast (2002) as part of his analysis of the tenuous trade-off between risk and incentives. Second, Baker (1992) shows that incentives will be weaker when the measures of performance deteriorate. To see that both these theories differ from the current one, note that intrinsic motivation is always good in these traditional agency models. Moreover, section 6 studies a variation on this paper's model in the spirit of Prendergast (2002) and shows that the effect of pay for performance in that model is exactly the opposite from its effect here.

The paper is also related to the literature on authority as a way to control agency problems. The most important result is the idea of efficiency wages (Shapiro and Stiglitz 1984), which was studied in more detail in MacLeod and Malcomsom (1989) and MacLeod and Malcomsom (1998).[4] These papers relate to the efficiency-wage model in appendix C, although there are substantially differences both in focus and in results.

Within the behavioral finance literature, Barberis and Thaler (2003) note that 'since [overoptimistic managers] think that they are already doing the right thing, stock options or debt are unlikely to change their behavior.' Their argument is thus that equity-based pay loses its ability to solve agency problems. The current paper, on the contrary, implies that stock options *will* change an overoptimistic manager's behavior, but in the wrong direction. The manager may, for example, more forcefully resist limits imposed by the board. In other words, in this paper, equity-based pay does not simply lose its ability to solve agency problems, it creates new ones. There are also more distantly related contributions, such as Manove and Padilla (1999), who show that the signaling function of collateral breaks down when there may be overoptimistic entrepreneurs or managers.

[3] Since we can interpret action X as 'do nothing' and action Y as 'do something', both the action space and the outcome space are clearly uni-dimensional in this model.

[4] Legros and Newman (2002) also show how the ability of players to jam the signals of their opponents to a judge in a legal dispute may lead to authority as the optimal solution.

The key contribution of this paper is to show that pay-for-performance and intrinsic motivation, which usually alleviate agency concerns, may instead exacerbate the agency conflict introduced by differing priors. In particular, such motivation to achieve good results will hinder authority. The theory provides a novel explanation for the prevalence of low-powered incentives in firms. It also provides an explanation for the informally observed correlation between subjective performance pay and authority, and makes new predictions regarding, for example, the effects of intrinsic motivation.

The next section considers a slightly simplified model to expose clearly the intuition and the analysis behind the paper. Section 3 then analyzes the full model and also shows that intrinsic motivation may reduce Pareto-efficiency, while section 4 shows that subjective bonuses may be optimal. Section 5 shows that differing priors give very different results from private benefits or private information, while section 6 considers some more technical issues. Section 7 considers the implications for governance, while section 8 concludes. The appendices contain some proofs and study useful extensions of, and variations on, this model.

2 Simple Setting

To make the intuition behind the paper transparent, I will start by analyzing a simplified version of the model that results in a very short and clear analysis. Section 3 then studies the full model to derive the paper's predictions.

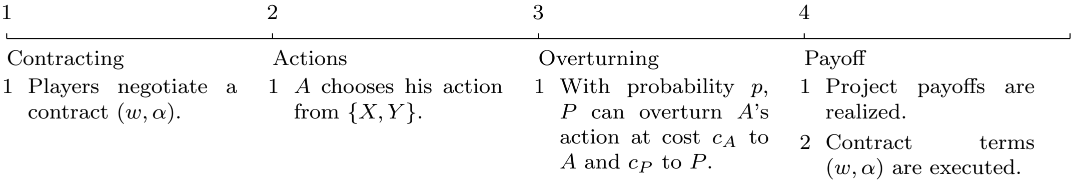
2.1 Model

Consider a setting in which a principal P hires an agent A to execute a project. As part of the project, A has to choose a course of action from the set $\{X, Y\}$. The agent's decision can be either right or wrong, resulting in a project revenue of respectively 1 or 0. The decision is right if and only if it fits the state of the world, which is either x or y. That state of the world is unknown, however, and each player i has his or her own subjective belief μ_i that the state is x. The players have differing priors, i.e., μ_A and μ_P may differ even though no player has private information.[5] For simplicity, assume that the players disagree on the optimal course of action, and that their beliefs are common knowledge.[6] In particular, let $\mu_P > .5 > \mu_A$ so that the principal believes that state x is most likely, while the agent believes that y is most likely. Let, finally, ν_i denote the strength of belief of player i, $\nu_i = \max(\mu_i, 1 - \mu_i)$, so that ν_i is also each player's belief in the state that he or she considers most likely.

The timing of the game is indicated in figure 1. In period 1, the players negotiate a compensation contract for A that consists of a wage w and a share of the project revenue $\alpha \in [0, 1]$. (P's

[5] Differing priors do not contradict the economic paradigm: while rational agents should use Bayes' rule to update their prior with new information, nothing is said about those priors themselves, which are primitives of the model. In particular, absent any relevant information agents have no rational basis to agree on a prior. Harsanyi (1968) observed that 'by the very nature of subjective probabilities, even if two individuals have exactly the same information and are at exactly the same high level of intelligence, they may very well assign different subjective probabilities to the very same events'. For a more extensive discussion, see Morris (1995) or Van den Steen (2005).

[6] To obtain a transparent and focused analysis, I eliminate here and in section 3 some elements from the model that would add realism but that are not necessary to explore the effects at the core of this paper. Appendix B presents a more complete model in which the players sometimes agree and sometimes disagree on the optimal course of action, and their beliefs are private information but can be communicated through cheap talk. Such communication about their beliefs are in effect 'orders' of one player to the other. While introducing such elements increases the realism of the model, doing so also adds complexity, without affecting its essential results.



1 Contracting

1 Players negotiate a contract (w, α).

2 Actions

1 A chooses his action from $\{X, Y\}$.

3 Overturning

1 With probability p, P can overturn A's action at cost c_A to A and c_P to P.

4 Payoff

1 Project payoffs are realized.

2 Contract terms (w, α) are executed.

Figure 1: Time line of simplified model

compensation is then $-w$ and the complementary share $(1-\alpha)$ of the project revenue.) Negotiation is according to axiomatic Nash bargaining with bargaining power λ and $1-\lambda$ for P and A, and outside options of 0 for both. I impose the no-wager condition that $\alpha \in [0,1]$ since the players would otherwise bet on the state and, in doing so, generate infinite utility. This no-wager condition would follow endogenously if players had the ability to sabotage the project, i.e., if each player had the ability to make sure that the project fails. In that case, any contract with $\alpha \notin [0,1]$ would give one of the players a strict incentive to sabotage the project. Anticipating that, the other would never accept the 'bet'. Alternatively, the condition could be defended based on legal provisions against gambling, or based on risk aversion. To maintain generality and simplify the analysis, I simply impose the condition as an assumption.

In period 2, A publicly chooses his action from the set $\{X, Y\}$. This decision is non-contractible, and the ultimate control over the decision is always in the hands of the agent and cannot be contracted or otherwise moved around. It follows that the decision will always be taken by the agent, who will choose the action that is best from his perspective given his beliefs and the contract negotiated in period 1.

In period 3, the principal has, with probability $p \in (0,1)$, the opportunity to overturn A's action, i.e., to force A to change his action. If P decides to overturn A's action, then A and P incur costs respectively $c_A, c_P > 0$.

In period 4, the state gets realized, the principal receives the project's revenue and pays the agent according to the contract (w, α). To study the effects of the agent's intrinsic motivation, I will also allow that A gets a private benefit $\gamma_A \geq 0$ when the project is a success (and 0 otherwise).[7] The idea here is that intrinsic motivation can be captured as a private benefit from success (or a private cost from failure). For notational simplicity, I will denote player i's total benefit as α_i, so that $\alpha_P = 1-\alpha$ and $\alpha_A = \alpha + \gamma_A$.

To simplify the analysis, I will also assume that $c_P < (2\nu_P - 1)$ and $c_A > \frac{(1-p)}{p}\gamma_A(2\nu_A - 1)$, which allows me to exclude some outcomes that aren't very interesting in the current context, but that complicate the analysis.[8]

The focus of the analysis is on whether or when A will do what P wants him to do, i.e., whether or when A will choose whatever action P thinks is best, rather than what he himself thinks is best. I will interpret this as A 'obeying' P, despite the lack of communication. It can be checked in

[7] Note that the restriction $\alpha \in [0,1]$ prevents players from contractually eliminating the effect of γ_A. Such contractual elimination would obviously defeat the purpose of introducing γ_A. Note, moreover, that a little bit of randomness in the value of γ_A would make it impossible to 'neutralize' its effect through a smart choice of α outside of $[0,1]$. It is also easy to check that all results (that are not about intrinsic motivation) go through with $\gamma_A = 0$, so that this private benefit is in no way necessary for the result.

[8] Without the first assumption, it will never be optimal for P to overturn any decision, while the effect of the second assumption is to exclude equilibria in which there is pure conflict between the agent and the principal: the agent always chooses Y, and the principal overturns the agent's decision whenever she gets the chance to do so.

appendix B that in a more complete model this corresponds indeed to A obeying P's orders.

2.2 Partial Equilibrium: The Effect of Motivation on Obedience

As a first step in the analysis, I take the compensation contract (α, w) as exogenous and consider under what conditions A will do what P wants him to do. In other words, I look for equilibria of the subgame starting in period 2, in which A 'obeys'. This builds intuition for the later results, but also delivers one of the main insights of the paper.

Starting with period 3, note that P will overturn A's decision if and only if $c_P \leq (1-\alpha)(2\nu_P - 1)$. Moving now back to period 2, if A believes that P will overturn his Y-decisions with probability q (which combines p with possible randomness in P's action), then he will choose X if

$$\alpha_A(1-\nu_A) + w \geq (1-q)(\alpha_A \nu_A + w) + q(\alpha_A(1-\nu_A) + w - c_A)$$

or $c_A \geq \sigma\alpha_A(2\nu_A - 1)$ where $\sigma = \frac{(1-q)}{q}$. This condition is weakest when $p = q$, i.e., when P overturns a Y-decision whenever she can. In that case the condition becomes $c_A \geq \theta\alpha_A(2\nu_A - 1)$ where $\theta = \frac{(1-p)}{p}$. Overall, there thus exists an equilibrium (for the subgame starting in period 2) in which A 'obeys' P if and only if

$$\begin{aligned} c_P &\leq (1-\alpha)(2\nu_P - 1) \\ c_A &\geq \theta\alpha_A(2\nu_A - 1) \end{aligned}$$

where $\theta = \frac{(1-p)}{p}$.

The second of these two conditions is central to the further analysis. It is the incentive compatibility constraint for the agent, i.e., the condition under which he is willing to 'obey' the principal. It says that obedience obtains only if the agent's cost, or penalty, from begin overturned is high enough.

The key insight here is that the minimal penalty to keep the agent honest *increases* in α_A, the agent's benefit from a success.[9] The reason is that, as α_A increases, the agent cares more about making the right decision, thus increasing his temptation to disobey when he is asked to do something he disagrees with, and thus reducing his 'zone of indifference' (Barnard 1938) or 'zone of acceptance' (Simon 1947). As I will show later, this is the opposite result from a model with private benefits: with private benefits, the minimal penalty to keep the agent honest *decreases* in the agent's benefit from success, α_A.

This result has some important implications. First of all, all else constant, agents with high pay-for-performance are more likely to disobey their principal and just do what they themselves consider optimal. This can manifest itself as either visible disobedience or as restraint by the principal in giving orders (since she knows she will be disobeyed). Either way, it implies a loss of control for the principal. This is one of the distinguishing predictions of this paper. It is also consistent with the management literature on sales compensation, which cites 'loss of control' (of the manager over her salespeople) and disregard of authority among the most important negative effects of sales commissions. Oliver and Anderson (1994), for example, using methods from psychology, show that sales people who are evaluated on outcome, which includes pay-for-performance incentive pay, are

[9] In the model of appendix C, this result takes a slightly different form. There, the principal has to pay an efficiency wage to make the agent obey (Shapiro and Stiglitz 1984). The key result then is that the efficiency wage increases in α, the agent's stake in the project, and in γ_A. In a model with private benefits instead of differing priors, the efficiency wage would *de*crease in α and γ_A. MacLeod and Malcomsom (1989) and MacLeod and Malcomsom (1998) study efficiency wage models in much detail, including the issue of commitment by the principal.

'less accepting of authority/direction'. Barry and Henry (1981) found that 'lack of control over salesforce' is one of the two most cited disadvantages of commission plans, as cited by firms that actually use commission plans. This effect that high-powered incentives cause a loss of control and disobedience will lead to the result in section 3 that, in equilibrium, agents who are subject to authority will have low-powered incentives.

A second implication of this result is that intrinsic motivation may sometimes be a bad thing. In particular, intrinsic motivation can be interpreted as the non-monetary benefit γ_A from achieving success, which will thus weaken the principal's authority and exacerbate the agency problem. The result implies that people with higher intrinsic motivation will be more difficult to control in case of disagreement. It will also lead to the result in section 3 that, in equilibrium, people with high intrinsic motivation will be less subject to authority.

A third implication of this result is that people are more likely to obey when they are not held responsible or accountable for the outcome, in the sense that they do not have to take the blame when things go wrong or do not get the praise when things go right.

Fourth, the condition implies some comparative statics that will be further formalized in section 3: the agent is more likely to obey when ν_A and γ_A are low and p is high. If ν_A decreases, then the agent cares less about what course of action he follows and thus has less reason to disobey. This comparative static may explain why some firms hire straight from college since such hires 'come with an open mind.' If γ_A decreases then the agent gains less from a success and thus has again less reason to disobey the principal. If the probability that the principal can overturn the agent's action, p, increases, then the expected cost of disobeying (getting overturned) increases while its expected benefit (doing what you think is right) decreases, both making disobedience less likely. Section 3 derives what these results imply for the equilibrium predictions.

Before moving to the equilibrium analysis, consider the other condition for obedience, i.e., $c_P \leq (1-\alpha)(2\nu_P - 1)$. This condition is the incentive compatibility for the principal to overturn the agent's decision. It says that the principal's cost of overturning the agent's decision may not be too high, with the maximal cost increasing in the principal's stake in the project and his conviction ν_P. It follows that obedience will be more prevalent when the principal has a high stake in the project and strong beliefs.

2.3 Equilibrium Contracts and Obedience

I can now move back one period and determine the full equilibrium of the simplified model. Since the contract gets determined by Nash bargaining, the players will select α to maximize the total surplus and divide that surplus among them by choosing w appropriately. It follows that the equilibrium is completely determined by finding the α that maximizes total utility.

Consider first the case in which $c_P > (1-\alpha)(2\nu_P - 1)$, so that P will never overturn any decision. In that case, A always chooses Y, so that P has no authority. I will denote this the *NAt* equilibrium (for 'No Authority'). The total utility is $(\alpha + \gamma_A)\nu_A + (1-\alpha)(1-\nu_P)$, which is maximized, and the condition maximally relaxed, at $\alpha = 1$. In that case, the total utility becomes $U_{NAt} = (1+\gamma_A)\nu_A$. The equilibrium requires $c_P > 0$, which is true by assumption.

Consider next the case where $c_P < (1-\alpha)(2\nu_P - 1)$, so that P will always overturn any Y-decision. If $c_A > \theta\alpha_A(2\nu_A - 1)$, then A will always obey and the principal has authority, in the sense discussed earlier. I will denote this the *At* equilibrium (for 'Authority'). The total utility is $(\alpha + \gamma_A)(1-\nu_A) + (1-\alpha)\nu_P$, which is maximized, and all constraints maximally relaxed, at $\alpha = 0$. The total utility is then $U_{At} = \gamma_A(1-\nu_A) + \nu_P$, and the equilibrium requires that $c_P < (2\nu_P - 1)$ and $c_A > \theta\gamma_A(2\nu_A - 1)$, which are both satisfied by assumption.

The case where $c_P < (1-\alpha)(2\nu_P - 1)$ and $c_A < \theta\alpha_A(2\nu_A - 1)$ (so that A always chooses Y and P overturns whenever possible) is strictly dominated by either *At* or *NAt*, and the same is true for any mixed strategy profiles.

Overall, there are thus two possible equilibria:

- *At* with $\alpha = 0$ and $U_{At} = \gamma_A(1-\nu_A) + \nu_P$.
- *NAt* with $\alpha = 1$ and $U_{NAt} = (1+\gamma_A)\nu_A$.

The overall equilibrium is the one with the largest total utility. I will discuss these results after first extending the model.

3 Complete Model

While the simplified model makes the basic analysis and intuition very transparent, it misses some elements that are useful for interpreting the results of the model. In this section, I will therefore add two elements to the model and then derive and discuss the model's implications.

3.1 Model

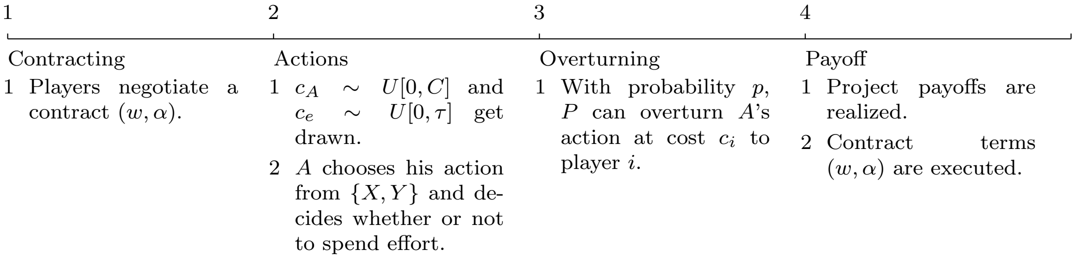
The timing of the complete game is indicated in figure 2. The first addition is to introduce a random element in the agent's decision. In particular, instead of being a fixed parameter, A's cost of being overturned, c_A, will be a random variable with probability distribution F. I will assume F to be uniform on $[0, C]$ with $C \geq \frac{(1-p)}{p}(1+\gamma_A)(2\nu_A - 1)$. The value of c_A will be publicly drawn at the start of period 2, i.e., after the contract negotiation but before the agent's action choice. This change to the model allows me to talk in a meaningful way about the likelihood that the agent obeys or disobeys, and how that affects equilibrium outcomes.

The second addition to the model is to introduce an independent moral hazard component that will affect the optimal level of incentive pay. In particular, I will assume that project success depends not only on the agent's choice of action but also on whether or not the agent spends effort. Formally, assume that simultaneously with his choice of action (X or Y), the agent also decides whether or not to spend effort. The cost of effort to the agent is a random variable c_e with a uniform distribution on $[0, \tau]$, where $\tau \in (0, 1)$. The value of c_e will be drawn, and revealed to both A and P, simultaneously with, but independently of, the value of c_A. With probability $(1-\tau)$, the project is, as before, a success if and only if the agent's decision matches the state of the world. With the complementary probability τ, however, the project is a success if and only if the agent spent effort. The parameter τ thus captures the relative importance of effort versus decision making. The effect of this change is to make α (sometimes) take values other than the extremes (0 and 1), so that I can say meaningful things about how incentive pay affects behavior in equilibrium.

3.2 Analysis

I now repeat the analysis of section 2 for the complete model. The first proposition considers again when the agent will do what the principal thinks is right, i.e., when A 'obeys' P.

Proposition 1 [Partial Equilibrium] *If P always overturns a Y-decision (when she can), then A will choose Y, and thus disobey, with probability $\frac{\theta\alpha_A(2\nu_A - 1)}{C}$ where $\theta = \frac{(1-p)}{p}$. The likelihood that A disobeys increases in α, γ_A, ν_A, and decreases in p.*



1 Contracting

1 Players negotiate a contract (w, α).

2 Actions

1 $c_A \sim U[0,C]$ and $c_e \sim U[0,\tau]$ get drawn.

2 A chooses his action from $\{X, Y\}$ and decides whether or not to spend effort.

3 Overturning

1 With probability p, P can overturn A's action at cost c_i to player i.

4 Payoff

1 Project payoffs are realized.

2 Contract terms (w, α) are executed.

Figure 2: Time line of complete model

Proof : If P always overturns Y (when she can), then A will choose X iff

$$\alpha_A(1-\nu_A) + w \geq (1-p)(\alpha_A \nu_A + w) + p(\alpha_A(1-\nu_A) + w - c_A)$$

or $c_A \geq \theta\alpha_A(2\nu_A - 1)$ where $\theta = \frac{(1-p)}{p}$. This implies the first part of the proposition, while the second part follows immediately. ■

Since this result was discussed in section 2, let me turn immediately to the overall equilibrium of the game.

There will be again two equilibrium regimes. In the first regime, A always chooses Y, i.e., A always disregards what the principal wants him to do and just follows his own beliefs. The principal never overturns any of the agent's decisions. The principal thus has no interpersonal authority over the agent (in the sense discussed earlier). In this equilibrium, the agent will have very high-powered incentives. As before, I will denote this type of equilibrium as *NAt*, which stands for 'No Authority'.

In the second regime, A chooses X with strictly positive probability, i.e., he sometimes does as the principal wants him to do, going against his own beliefs. Whenever the agent chooses Y, the principal will try to overturn that decision (which succeeds with probability p). In this case, the principal thus has (some) interpersonal authority over the agent. The agent will typically have low-powered incentives. I will again denote this type of equilibrium as *At*, which stands for 'Authority'.

To state the proposition formally, let me define

$$\begin{aligned} f &= \tau - (1-\tau)(\nu_A + \nu_P - 1) + (1-\tau)(1-p)\frac{\theta}{C}(2\nu_A - 1)\left(\gamma_A(2\nu_A - 1) + (\gamma_A - 1)(2\nu_P - 1)\right) \\ g &= \tau - (1-\tau)(1-p)\frac{\theta}{C}(2\nu_A - 1)\left((2\nu_A - 1) + 2(2\nu_P - 1)\right) \end{aligned}$$

Proposition 2 [Equilibrium] *There exists $\hat{\nu}_P$ such that the equilibrium is of type At if $\nu_P \geq \hat{\nu}_P$, and of type NAt otherwise.*

- *If the equilibrium is At and either $g \leq 0$ or $f \leq 0$ (which includes $\tau = 0$), then $\hat{\alpha} = 0$.*
- *If the equilibrium is At and $f, g > 0$, then $\hat{\alpha} = \min(f/g, 1 - c_P/(2\nu_P - 1)) < 1$.*
- *If the equilibrium is NAt, then $\hat{\alpha} = 1$.*

The value of $\hat{\nu}_P$ increases in ν_A, γ_A, and τ.

Proof : The proof is in appendix A. ■

To understand this result and the intuition behind it, consider first the case without the extra moral hazard component.

Corollary 1 *If $\tau = 0$, then the equilibrium is At with $\alpha = 0$ when $\nu_P > \hat{\nu}_P$, while it is NAt with $\alpha = 1$ when $\nu_P < \hat{\nu}_P$. The value of $\hat{\nu}_P$ increases in ν_A and γ_A.*

Proof : Fix $\tau = 0$. In that case, $g = -(1-p)\frac{\theta}{C}(2\nu_A - 1)\left((2\nu_A - 1) + 2(2\nu_P - 1)\right) < 0$, so that, if the equilibrium is *At*, then $\hat{\alpha} = 0$. The rest of the proposition follows immediately from proposition 2. ■

In this base case, there are clearly two extreme regimes: one with low-powered incentives and authority of the principal over the agent, and one with high-powered incentives and no authority. Consider first the authority regime. There are three reasons why authority goes together with low-powered incentives. First, low-powered incentives minimize the probability that the agent disobeys, and thus enhances authority. Second, low-powered incentives for the agent means a high stake for the principal, which commits the principal to overturning decisions whenever the agent 'disobeys'. Third, since most decisions follow the principal's beliefs, she will value the returns more than the agent does, which favors shifting residual income to the principal. The reasons for high-powered incentives in *NAt* are exactly the opposite. I will discuss the comparative statics in terms of γ_A and ν_A later.

When τ increases, the moral hazard component comes into play, which favors higher pay for performance for *A*. At first, the only effect is to make *NAt* more attractive and thus more prevalent (by increasing $\hat{\nu}_P$), since *NAt* does have stronger incentives than *At*. For large enough values of τ, there may in some cases even be pay for performance in the *At* regime. For one such case, the evolution of α as a function of τ is depicted in figure 3. Note the rise in α at larger values of τ and then the sudden jump when the regime shifts to *NAt*. The reason for the jump to *NAt* is that pay for performance is very costly under *At* since it causes disobedience and shifts income to the player who values it less. Note also that the parameter values are extreme: $\nu_P = .9$ and $\nu_A = .55$. This is no coincidence: pay for performance under *At* is limited to cases where a high need for effort (τ) is combined with very asymmetric beliefs (high ν_P and low ν_A) that make authority very attractive. In most cases, $\alpha = 0$ everywhere in the *At* equilibrium. And even when $\alpha > 0$, it stays relatively low-powered except in the most extreme of cases.

The proposition thus predicts that people who are subject to authority will usually have low-powered incentive pay, or often even fixed salaries. Since nearly all employees are subject to authority, it thus provides a new explanation for the (informally observed) lack of high-powered incentives in firms. In the other direction, the model predicts that agents with high-powered incentives should be less subject to authority. A nice illustration of this phenomenon can be found in the HBS case on Lincoln Electric (Berg and Fast 1983), a firm famous for its high-powered incentive systems: both employees who are interviewed about the company start out by saying how much they like being their 'own boss' or their 'own man'. This prediction also fits the finding, mentioned earlier, that salespeople who are evaluated on outcomes are 'less accepting of authority/direction' (Oliver and Anderson 1994).

The result also establishes that there are two regimes: one in which the principal decides on the course of action and bears the risk of her decisions, and another in which the agent decides on the course of action and bears the risk of his decisions. I will discuss this result and its potential relevance for the theory of the firm in more detail in section 7.

Consider next the comparative statics on the prevalence of authority. In particular, *At* becomes more prevalent (in the sense of obtaining in a strictly larger subset of the appropriate section of

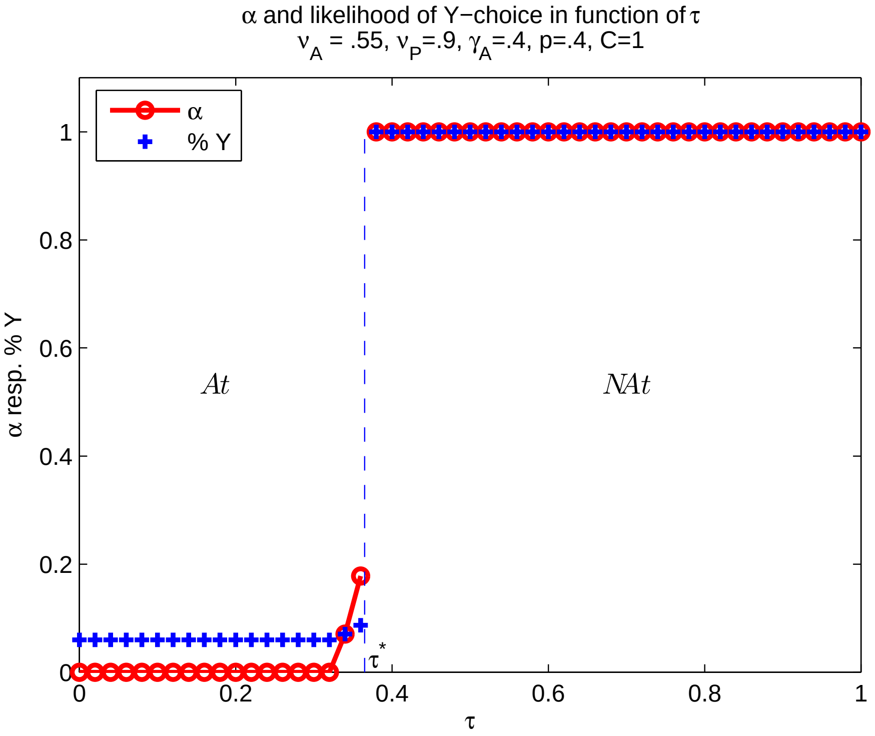


Figure 3: Evolution of α and the likelihood that A choose Y, as a function of τ.

the $(\nu_A, \nu_P, \gamma_A, p, c_A, c_P)$ parameter space) as γ_A, ν_A, and τ are smaller and as ν_P is larger.

Consider first the comparative static on ν_A, the agent's belief strength. If ν_A is higher, then A cares more about which action is taken and is thus less willing to obey P, making authority more difficult to implement. In other words, people with strong beliefs about the right course of action have difficulties following the ideas and 'orders' of others and will, in equilibrium, be less subject to authority. This predicts that people with strong beliefs should be more likely to become entrepreneurs. This is consistent with the evidence that entrepreneurs have relatively strong beliefs about being right (Cooper, Dunkelberg, and Woo 1988, Landier and Thesmar 2004). It also predicts that people with strong beliefs are more likely to be in leadership or management positions, while people with weak beliefs are more likely to be in the position of follower or subordinate.

Consider next the comparative static on γ_A. The private benefit from success, γ_A, gives A a reason to decide according to his own insights, and thus to disobey P. The most useful interpretation of γ_A is that it captures intrinsic motivation, e.g. personal satisfaction from achieving success or career concerns. The theory then predicts that people with high intrinsic motivation are less likely to be subject to authority. Note that this result is *not* caused by the fact that motivated people need less supervision, but by the fact that it is costly to make them obey. This prediction is consistent with McClelland's (1964) theory that entrepreneurs will be people with a high need for achievement.

Another implication of the result that *NAt* is more likely when γ_A is large, is that a firm with highly motivated employees will need to rely on other methods than authority to achieve coordination. One important alternative is to hire people with similar beliefs (Van den Steen 2004). The prediction would thus be that firms that put a lot of weight on intrinsic motivation when hiring new employees will also hire more on 'fit' than other firms and will invest more in socialization and training.

Consider next the result that At becomes less prevalent as τ increases. As mentioned earlier, the parameter τ captures the relative importance of effort versus decision making. The result thus predicts that At gets more prevalent as the importance of making the right decisions increases relative to the importance of effort. This is discussed further in section 7.

The fact, finally, that authority is more likely when the principal has stronger beliefs is caused by the fact that stronger beliefs give him more reason to make sure his beliefs are followed, and thus to overturn the agent's decision when needed. The result predicts that people with strong beliefs are most effective at exercising authority.

3.3 Counter-Productive Intrinsic motivation

Propositions 1 and 2 suggest that intrinsic motivation may be dysfunctional in a world with differing priors, since it makes interpersonal authority less effective. I will show here that an increase in the intrinsic motivation of A may indeed decrease the players' expected utility, despite the fact that the direct effect of an increase in intrinsic motivation is to increase A's utility.

The following proposition says that an increase in A's intrinsic motivation γ may decrease both players' utilities.

Proposition 3 *There exists $(\nu_A, \nu_P, \gamma_A, p, \tau)$ such that each player's utility strictly decreases in γ_A.*

Proof : Consider the derivative of U_{At}

$$\frac{dU_{At}(\alpha)}{d\gamma_A} \quad = \quad (1-\tau)(1-\nu_A) + (1-\tau)(1-p)\frac{\theta}{C}(2\nu_A - 1)\left(\alpha_A(2\nu_A-1) - \alpha_P(2\nu_P - 1)\right) + (1+\gamma_A)\tau$$

At $\gamma_A = 0$, $C = \theta(1+\gamma_A)(2\nu_A - 1)$, and $\tau = 0$ (so that also $\hat{\alpha} = 0$), this becomes

$$\frac{dU_{At}(\alpha)}{d\gamma_A} = (1-\nu_A) - (1-p)(2\nu_P - 1)$$

which is strictly negative for sufficiently small p and sufficiently large ν_P. Moreover, $U_{At} > U_{NAt}$ as long as $\nu_P > \nu_A$ (for p small enough). This proves the proposition. ■

The intuition is as follows. While an increase in γ_A indeed increases A's utility, this gain will be small when A always obeys P (which is the case when α and γ_A equal zero) and A has strong beliefs (so that he is nearly sure the project will fail). This arbitrarily small gain is outweighed by the cost of A disobeying more when γ_A increases. The intuition thus also suggests that this utility-lowering effect of intrinsic motivation is most likely when both A and P have strong beliefs, and A's intrinsic motivation is small to begin with.

The key implication of this section on counter-productive intrinsic motivation, then, is that it is not always optimal for a firm to try to hire employees with high intrinsic motivation.

4 Subjective Performance Pay

Up to this point, I have considered pay-for-performance based only on objective measures of true performance. There are, however, some elements in the analysis that suggest that subjective performance measures (where the principal pays the agent according to how well she thinks the agent performed) may be more effective in this context. In particular, the agency problem originates here in the fact that the principal and the agent have different perspectives on the same issue. The problem might be solved if the agent were to see the problem 'through the eyes of the principal'. It is here that subjective performance pay might be useful. The purpose of this section is to consider that suggestion more formally.[10]

To study this issue, I will assume that the contract between the principal and the agent may, apart from w and α, also contain a provision that the agent gets a bonus at the end of period 3 that equals β times the principal's expected value (at that time), $\beta E_P[R]$, with $\beta \in [0, (1-\alpha)]$ and R the revenue from the project.[11] I abstract here from the (very important) questions how the principal could commit to such a payment and whether other subjective compensation schemes may be even better. My only purpose is to show that, if the parties can find a way to implement this scheme, such subjective measures of performance may completely dominate the objective ones. In other words, I want to show that subjective bonuses are not just a second-best solution when objective bonuses are unfeasible, but may actually be better even when objective payments are available.

The following proposition says indeed that objective pay for performance will never be used in an At equilibrium when subjective pay for performance is available, and vice versa for *NAt*. For simplicity, I assume that there are no private benefits from success, i.e., $\gamma_A = 0$. I will discuss the intuition in more detail after the proposition and its proof.

Proposition 4 *Any contract $(w, \alpha, \beta) = (v, a, b)$ that induces an At equilibrium and that is not dominated by a NAt equilibrium, is Pareto dominated by a contract $(v, 0, a + b)$, which also induces At and is not dominated by NAt.*

Any contract $(w, \alpha, \beta) = (v, a, b)$ that induces a NAt equilibrium is Pareto-dominated by a contract $(v, a + b, 0)$, which also induces a NAt equilibrium.

Proof : Note first that the effort part of the payoffs and of the analysis is not affected by shifts between objective and subjective performance pay (since these effects are independent of the beliefs that are used to evaluate payoffs).

I will now first show that the contract $(v, 0, a + b)$ always induces an At equilibrium (and uniquely so when $a + b > 0$). Consider a contract (w, α, β). When P overturns Y-decisions with probability q (which combines p with the likelihood that P tries to overturn, and which may thus be zero) the IC constraint that makes A obey is

$$c_A \geq \sigma[\alpha(2\nu_A - 1) - \beta(2\nu_P - 1)]$$

with $\sigma = \frac{(1-q)}{q}$. When $\alpha = 0$ and $\beta = a + b$, then this is always satisfied (even when P never overturns any of A's decisions).

Consider now a contract (v, a, b) that induces an At equilibrium and that is not dominated by a *NAt* equilibrium, which implies that $\nu_P \geq \nu_A$. The above argument implies that $(v, 0, a + b)$ also induces At, with A always obeying. Total utility then equals ν_P. Since the total utility under the original contract is either $a(1 - \nu_A) + (1 - a)\nu_P$ or

$$(1-q)\frac{z}{C}(a\nu_A + (1-a)(1-\nu_P)) + (1 - (1-q)\frac{z}{C})(a(1-\nu_A) + (1-a)\nu_P) - q\int_0^z u\frac{1}{C}\,du$$

[10] There is a considerable literature on subjective bonuses, such as Baker, Gibbons, and Murphy (1994), MacLeod and Malcomsom (1998), or Suvorov and van de Ven (2005), but these papers typically assume that the objective outcome is non-contractible, as opposed to here.

[11] The restriction that $\beta \leq (1 - \alpha)$ is again the no-wager condition, but now from the subjective perspective of the principal: the principal cannot promise more than what he expects to get from the project.

for some $z \in [0, C]$ and for some $q \in (0, 1)$, the total utility increases when going from (v, a, b) to $(v, 0, a + b)$. This proves the first part of the proposition.

Consider next a *NAt* equilibrium. Note that the total utility under (v, a, b) equals

$$a\nu_A + b(1 - \nu_P) + (1 - a - b)(1 - \nu_P) = a\nu_A + (1 - a)(1 - \nu_P)$$

which increases to $(a + b)\nu_A + (1 - a - b)(1 - \nu_P)$ under $(v, a + b, 0)$. This proves the proposition. ■

To see the intuition behind the result, note first that disagreement is costly, especially in an At equilibrium: not only is one of the players always disappointed with the decision, and thus expects a low payoff, but disagreement also creates direct costs in At, when decisions get overturned. If, however, A gets paid a subjective bonus, then he will evaluate actions and outcomes as if he had the principal's beliefs. This alignment of beliefs will make the agent do as the principal wants him to do, eliminating costly reversals. Moreover, it will also make A evaluate payoffs using P's beliefs so that he will expect a high payoff from such decisions. In a *NAt* equilibrium, on the contrary, where A always chooses Y and P never overturns A's decisions, subjective bonuses can never be optimal. On the one hand, a subjective bonus will be smaller (in the eyes of the agent) than the equivalent objective one, since the principal believes that the project is likely to fail and thus has a low expected value. On the other hand, making the agent think more like the principal also doesn't help as long as the equilibrium stays *NAt*.[12]

This result thus establishes two things. First, subjective pay-for-performance may sometimes be optimal, rather than a second-best solution when outcomes are difficult to measure. Second, the result also predicts that authority (for the principal) will go together with subjective pay-for-performance (for the agent). This prediction is consistent with the observation that subjective bonuses are often used within firms, where managers exert authority over employees, but are only rarely used between firms, where authority is often much weaker or non-existent.

5 Differing Priors versus Private Benefits or Private Information

When working with differing priors, it is useful to consider whether the results are unique to differing priors or could also be obtained in a model where agency problems are modelled as private benefits or private information. There are at least two reasons for this. First, although differing priors may sometimes be a more intuitive or more relevant way to model agency problems, the argument for modelling with differing priors is obviously stronger if the results cannot be obtained otherwise. Second, answering this question deepens our understanding of the underlying mechanism, and the role that differing priors play in it.

In this case, the result is striking: differing priors and private benefits give essentially *opposite* results, while private information gives no result like the one derived in this paper. This is an important outcome from a methodological point of view, since it implies that results obtained with private benefit models do not necessarily extend to the case of differing priors.

[12] At first sight, it may seem that there is an alternative intuition. In particular, it may seem that subjective bonuses implicitly allow 'contracting on actions'. The complete model in appendix B and the efficiency-wage model in appendix C make clear that this is not the right intuition. In both cases, the subjective bonus results holds. However, in the first, it is not known in advance what action the principal will believe is right, so contracting on actions could never replicate a subjective bonus. In the second, the difference is even stronger, since P often does not know A's action, only what A is supposed to do. Both imply that the result is really about A looking through the eyes of P, rather than finding indirect ways to contract on actions.

5.1 Private benefits

To compare differing priors with private benefits, I will analyze a variation on the simple model of section 2 in the style of Prendergast (2002). As in section 2, assume that the agent has to choose a course of action from $\{X, Y\}$, and that the action is a success, giving payoff 1 instead of 0, if and only if it fits the state of the world, which is either x or y. However, to make this a common-prior model, now assume that the probability that the state is x is commonly known to be $\rho > .5$. The players thus agree that action X is more likely to succeed than action Y. The agent, however, has a commonly known private benefit b from undertaking Y. The timing of the game remains that of figure 1, including the principal's ability to overturn, with probability p, the agent's decision at cost c_i to player i.

I focus again on the contract terms under which the agent will do as his principal prefers. The following proposition identifies two sets of conditions. I discuss these conditions and their implications after the proof.

Proposition 5 *There exists an equilibrium for the subgame starting in period 2 in which A does what P wants him to do if and only if*

- *either* $c_A \geq \theta(b - \alpha(2\rho - 1))$ *and* $\alpha \leq 1 - \frac{c_P}{(2\nu_P - 1)}$,
- *or* $\alpha(2\rho - 1) \geq b$.

Proof : As before, P will overturn a Y-decision only if $(1 - \alpha)(2\nu_P - 1) \geq c_P$ or $\alpha \leq 1 - \frac{c_P}{(2\nu_P - 1)}$. Assume first that this is satisfied and that P overturns a Y-decision with probability q (which combines p and any potential randomization by P), then A will choose X if

$$\alpha\rho + w \geq (1 - q)(\alpha(1 - \rho) + b + w) + q(\alpha\rho + w - c_A)$$

or $c_A \geq \sigma(b - \alpha(2\rho - 1))$ where $\sigma = \frac{(1-q)}{q}$. In this case, a (subgame) equilibrium in which A obeys is possible if and only if $c_A \geq \theta(b - \alpha(2\rho - 1))$.

Assume next that $\alpha > 1 - \frac{c_P}{(2\nu_P - 1)}$, so that P never overturns any decision. Now A will choose X if and only if $\alpha(2\rho - 1) \geq b$. ■

Consider the first set of conditions. They identify the subgame equilibrium that is the analogue of the subgame equilibrium identified in section 2.2 in which the agent obeys. In particular, in both cases A chooses the action that P considers best, based on a credible 'threat' by P that she will overturn the decision otherwise. The condition that $c_A \geq \theta(b - \alpha(2\rho - 1))$ is the IC constraint that makes A 'obey': it specifies the minimal penalty upon getting overturned that induces obedience. In section 2.2, this minimal penalty increased in α, so that pay for performance hindered authority. Here, on the contrary, the minimal penalty *decreases* in α, so that pay for performance now facilitates obedience. That is the key result of this subsection: differing priors and private benefits have *opposite* effects.

The fact that pay-for-performance has a very different effect when the agency issue is private benefits is even more clear from the second set of conditions: when A's stake α is high enough, then A obeys even without P's threat to overturn the decision. The exact expression for 'high enough' is the condition that $\alpha(2\rho - 1) \geq b$. In other words, pay-for-performance aligns the objectives of principal and agent when the agency issue is private benefits, while it further misaligns their objectives when the agency issue is differing priors.

5.2 Private Information

At first sight, differing priors may also seem similar to private information that cannot be communicated: in both cases, the principal and agent have different beliefs about the right course of action. It may seem, in particular, that an agent with private information may also care more about control when he has a stake in the outcome. While it is true that incentive pay will make the agent care more about the decision and about the allocation of control, private information in itself causes no agency conflict in the sense that there is no conflict in objectives and that there can never be open disagreement between the principal and the agent on the optimal decision or on the optimal allocation of control. For example, the principal and the agent always agree that whoever has the best information should make the decision.

Agency considerations come into play where there is both private information and private benefits. The typical situation is one in which the agent has superior information about the optimal course of action, but also has private reasons to deviate from that course of action. Consider, for example, the following variation on the private-benefits model of section 5.1. Assume that principal and agent both start with the prior that both actions are equally likely to succeed, and that the agent gets, at the start of period 2, a signal about the state of the world which is correct with probability ρ. At the start of period 3, the principal observes the agent's signal with probability p and can overturn the agent's decision if he wants to. The agent has a commonly known private benefit b from undertaking action Y. Otherwise, the setting is identical to that of section 5.1.

It turns out that this setting is completely isomorph with the one in section 5.1. In particular, the following proposition shows that the same result applies.

Proposition 6 *There exists an equilibrium for the subgame starting in period 2 in which A does what P wants him to do if and only if*

- *either $c_A \geq \theta(b - \alpha(2\rho - 1))$ and $\alpha \leq 1 - \frac{c_P}{(2\nu_P - 1)}$,*
- *or $\alpha(2\rho - 1) \geq b$.*

Proof : P will overturn a Y-decision only if she gets information about the state (which happens with probability p), and then only if $(1 - \alpha)(2\rho - 1) \geq c_P$ or $\alpha \leq 1 - \frac{c_P}{(2\rho - 1)}$.
Assume first that this is satisfied and that P overturns a Y-decision with probability q (which combines p and any potential randomization by P), then A will choose X if

$$\alpha\rho + w \geq (1 - q)(\alpha(1 - \rho) + b + w) + q(\alpha\rho + w - c_A)$$

or $c_A \geq \sigma(b - \alpha(2\rho - 1))$ where $\sigma = \frac{(1-q)}{q}$. In this case, a (subgame) equilibrium in which A obeys is possible if and only if $c_A \geq \theta(b - \alpha(2\rho - 1))$.
Assume next that $\alpha > 1 - \frac{c_P}{(2\nu_P - 1)}$, so that P never overturns any decision. Now A will choose X if and only if $\alpha(2\rho - 1) \geq b$. ■

The same conclusions as before thus apply: the combination of private information and private benefits has the opposite effect of differing priors.

6 Some Technical Considerations

This section addresses some issues that arise quite naturally from the model and its analysis.

6.1 Contractible Action Choice

Is non-contractibility of actions a necessary ingredient for the effect in the paper? It turns out that this is not the case, but that non-contractibility does make the paper's effects more important and more powerful.

To see this, consider the original model with two modifications. First, actions are always perfectly observable and contractible. Second, the payoff upon failure is $-k < 0$. In this case, it is efficient for the agent to do what the principal considers best iff $\alpha_P(2\nu_P - 1) \geq \alpha_A(2\nu_A - 1)$. Moreover, for high enough k and ν_A, the principal's payoff then decreases in the agent's stake α_A. So the principal will sometimes prefer an agent with lower motivation for success. The intuition is similar to the earlier effect: getting the agent to agree to do something that he expects to fail is more costly as the agent has a higher stake in the failure, and thus gets hurt more by agreeing.

This effect is less general than the main effect of the paper since it requires that payoffs from failure are negative. It does show, however, that the underlying ideas of the paper are not limited to a context with non-contractible actions.

6.2 Renegotiation

As with many solutions to agency problems, there is an issue of renegotiation in this model. In particular, it will sometimes be ex-post inefficient for the principal to overturn the agent's decision, for example when c_A is very high. This is a well-known issue in agency theory: once the agent has made his decision, some of the measures that were meant to affect the agent's action are now ex-post inefficient and would be renegotiated if the players got the chance to do so. In a traditional model of moral hazard, for example, where the agent has to choose costly actions but is risk-averse, it is typically ex-post Pareto optimal to renegotiate the agent's compensation plan once he has chosen his actions. But since the players will anticipate this renegotiation, the agency problem gets worse (Fudenberg and Tirole 1990).

There are two reasons why the issue is smaller here than in some other agency models. First of all, overturning the decision will often actually be the efficient thing to do (even taking into account c_A and c_P), and will thus not be renegotiated. In these cases, the analysis goes through without a change and renegotiation is simply not an issue. Second, since overturning decisions is typically off the equilibrium path, it should be easier for the principal to build a reputation for not renegotiating. Beyond this, the typical defenses apply: renegotiation may be impossible due to time constraints or other reasons. Overall, renegotiation is an important issue, but does not seem to be more of an issue here than in other agency settings.

7 Potential Implications for Governance and Theory of the Firm

Grocers versus Employees Alchian and Demsetz (1972) famously argued that a manager exerts no more authority over his employee than a customer exerts over his grocer. In particular, you can 'fire' your grocer, just as you can fire an employee. Some people have suggested even worse: that contractors are *more* likely to obey than employees, since firing an employee can be difficult, while you can easily walk away from your grocer.

The analysis in this paper suggests that this argument overlooks an important aspect: the fact that a typical employee has a very limited financial stake in the outcome of the project, relative to a contractor. As a consequence, employees and contractors will obey *under different circumstances.*

If the principal's 'order' does not require private effort, but affects the probability of success of the project, and there is disagreement on whether the order is a good idea, then a contractor will be more reluctant to obey than an employee, since the contractor bears a large share of the consequences while the employee doesn't. The fact that an employee may be willing to go along with an order that he thinks is wrong, is sometimes reflected in a complaint in the following style: 'This is a stupid decision. But you know what? I don't care. It's his project and if he wants it this way, we'll do it that way.' If, on the other hand, the 'order' by the principal requires private effort from the agent or causes a private cost for the agent, then a contractor may well be more responsive than an employee.

In line with this argument, section 3 concluded that authority will be more prevalent as the importance of making the right decisions increases relative to the importance of effort. This suggests that in equilibrium we will observe employees when such obedience is important.

Two Distinct Regimes The paper has a second potential implication for the theory of the firm. In particular, the analysis shows that two qualitatively different regimes emerge:

1. A gets nearly all the residual income and P does not interfere with A's decisions.

2. P gets nearly all the residual income and tells A what to do.

Like Holmstrom and Milgrom (1994), then, this paper suggests that there are two different systems that resemble to some degree the 'firms versus markets' distinction. Moreover, in this paper the key elements of these systems are interpersonal authority and residual income, which fits well with the intuitive view that many people have of the distinction between firms and markets.

8 Conclusion

High-powered incentives and intrinsic motivation may make authority less effective. As a consequence, employees subject to authority will typically have low-powered incentives, while firms with intrinsically motivated employees may have to rely on alternative means of coordination, such as hiring people with similar beliefs or preferences. Moreover, subjective bonuses may actually be optimal, not just a second-best solution when outcomes are difficult to measure.

The analysis also has an important methodological implication: differing priors can have very different, even opposite, results from private benefits. We can therefore not assume that differing priors are just an extension of private benefits, and that results derived under private benefits extend to a context with differing priors. This opens up important new areas of research, since open disagreement is an important aspect of organizational life.

A Proofs

Proof of Proposition 2: Note that in the Nash bargaining in period 1, the players will agree on the value of α that gives the highest total continuation utility, and bargain over w to allocate that total utility between them. It thus suffices to determine which α gives the highest total utility.

For backwards induction, consider first the choice of effort (which is independent of any decision or overturning). The agent will spend effort if $\tau\alpha_A \geq c_e$, so the agent's utility from the effort portion is

$$\int_0^{\tau\alpha_A} (\tau\alpha_A - u)\,\frac{1}{\tau}\,du = \tau\alpha_A^2 - \frac{\tau\alpha_A^2}{2} = \frac{\alpha_A^2}{2}\tau$$

while the principal's utility from the effort portion is

$$\int_0^{\tau\alpha_A} (\tau\alpha_P)\,\frac{1}{\tau}\,du = \tau\alpha_A\alpha_P$$

so that the total utility from effort equals

$$\frac{\alpha_A^2 + 2\alpha_A\alpha_P}{2}\tau$$

When $\alpha > 1 - \frac{c_P}{(2\nu_P - 1)}$ (i.e., $c_P > \alpha_P(2\nu_P - 1)$), then the unique equilibrium is for P never to overturn any of A's decisions, and thus for A to always choose Y. This is the *NAt* equilibrium. The total utility equals

$$U_t = (1-\tau)(\alpha_A\nu_A + \alpha_P(1-\nu_P)) + \frac{\alpha_A^2 + 2\alpha_P\alpha_A}{2}\tau$$

with derivative

$$\begin{aligned}\frac{dU_t}{d\alpha} &= (1-\tau)(\nu_A + \nu_P - 1) + \frac{2\alpha_A + 2\alpha_P - 2\alpha_A}{2}\tau \\ &= (1-\tau)(\nu_A + \nu_P - 1) + (1-\alpha)\tau\end{aligned}$$

which is maximized at $\alpha = 1$ with utilities

$$U_{NAt} = (1-\tau)(1+\gamma_A)\nu_A + \frac{(1+\gamma_A)^2}{2}\tau$$

When $\alpha < 1 - \frac{c_P}{(2\nu_P - 1)}$, P will always overturn Y-decisions (when she can). Proposition 1 implies that A will choose X with probability $1 - \frac{\theta\alpha_A(2\nu_A - 1))}{C}$ where $\theta = \frac{(1-p)}{p}$. This is the At equilibrium. Let now $z_\alpha = \theta(\alpha + \gamma_A)(2\nu_A - 1)$, which is smaller than C by assumption. The total utility now equals,

$$\begin{aligned}U_t(\alpha) &= (1-\tau)\int_0^{z_\alpha} \left[(1-p)\left(\alpha_A\nu_A + \alpha_P(1-\nu_P)\right) + p\left(\alpha_A(1-\nu_A) + \alpha_P\nu_P - u\right)\right]\frac{1}{C}\,du \\ &\quad + (1-\tau)\left(1 - \frac{z_\alpha}{C}\right)\left[\alpha_A(1-\nu_A) + \alpha_P\nu_P\right] + \frac{\alpha_A^2 + 2\alpha_A\alpha_P}{2}\tau \\ &= (1-\tau)\left(\alpha_A(1-\nu_A) + \alpha_P\nu_P\right) + (1-\tau)(1-p)\frac{\theta}{C}\alpha_A(2\nu_A - 1)\left(\frac{1}{2}\alpha_A(2\nu_A - 1) - \alpha_P(2\nu_P - 1)\right) \\ &\quad + \frac{\alpha_A^2 + 2\alpha_A\alpha_P}{2}\tau\end{aligned}$$

Note that

$$\begin{aligned}U_{At}(\alpha = 1) &= (1-\tau)(1+\gamma_A)(1-\nu_A) + (1-\tau)(1-p)\frac{\theta}{C}(1+\gamma_A)^2(2\nu_A - 1)^2\frac{1}{2} + \frac{(1+\gamma_A)^2}{2}\tau \\ &< (1-\tau)(1+\gamma_A)\frac{1}{2} + \frac{(1+\gamma_A)^2}{2}\tau < U_{NAt}\end{aligned}$$

The derivative is

$$\begin{aligned}
\frac{dU_{At}(\alpha)}{d\alpha} &= (1-\tau)\left((1-\nu_A)-\nu_P\right)+(1-\tau)(1-p)\frac{\theta}{C}(2\nu_A-1)\left(\frac{1}{2}\alpha_A(2\nu_A-1)-\alpha_P(2\nu_P-1)\right) \\
&\quad +(1-\tau)(1-p)\frac{\theta}{C}\alpha_A(2\nu_A-1)\left(\frac{1}{2}(2\nu_A-1)+(2\nu_P-1)\right)+\frac{2\alpha_A+2\alpha_P-2\alpha_A}{2}\tau \\
&= \tau-(1-\tau)\left(\nu_A+\nu_P-1\right)+(1-\tau)(1-p)\frac{\theta}{C}(2\nu_A-1)\left(\gamma_A(2\nu_A-1)+(\gamma_A-1)(2\nu_P-1)\right) \\
&\quad +\left((1-\tau)(1-p)\frac{\theta}{C}(2\nu_A-1)\left((2\nu_A-1)+2(2\nu_P-1)\right)-\tau\right)\alpha \\
&= f-g\alpha
\end{aligned}$$

where

$$\begin{aligned}
f &= \tau-(1-\tau)\left(\nu_A+\nu_P-1\right)+(1-\tau)(1-p)\frac{\theta}{C}(2\nu_A-1)\left(\gamma_A(2\nu_A-1)+(\gamma_A-1)(2\nu_P-1)\right) \\
g &= \tau-(1-\tau)(1-p)\frac{\theta}{C}(2\nu_A-1)\left((2\nu_A-1)+2(2\nu_P-1)\right)
\end{aligned}$$

The second derivative is

$$\frac{d^2U_{At}(\alpha)}{d\alpha^2}=(1-\tau)(1-p)\frac{\theta}{C}(2\nu_A-1)\left((2\nu_A-1)+2(2\nu_P-1)\right)-\tau=-g$$

If $g \leq 0$, then the solution is either $\alpha = 0$ or $\alpha = 1$.[13] Since $U_{At}(\alpha = 1) < U_{NAt}$, it follows that, using $\hat{\alpha}$ to denote the optimal α, $\hat{\alpha} = 0$ whenever the equilibrium is *At*.

Consider next $g > 0$, so that the unrestricted optimal α is unique and determined by the FOC, and would thus equal f/g. If $f \leq 0$, then $\hat{\alpha} = 0$. When $f > 0$, then the optimal α is restricted by the condition that $\alpha \leq 1 - \frac{c_P}{(2\nu_P-1)}$, so that $\hat{\alpha} = \min(f/g, 1 - \frac{c_P}{(2\nu_P-1)})$. Finally, when $\alpha = 1 - \frac{c_P}{(2\nu_P-1)}$ both *At* and *NAt*, with $\hat{\alpha}$ as derived above, are possible, but mixed strategy profiles are always strictly dominated. All this implies the first part of the proposition.

I will now show that $U_{At}(\hat{\alpha}) - U_{NAt}$ strictly decreases in ν_A and γ_A. Moreover, whenever $U_{At}(\hat{\alpha}) = U_{NAt}$, $U_{At}(\hat{\alpha}) - U_{NAt}$ strictly increases in ν_P and strictly decreases in τ. These four results imply the second part of the proposition follows. Note that the envelope theorem applies, so that it is not necessary to consider how the optimal α changes with any of these parameters.

Consider first ν_A. Note that

$$\frac{dU_{NAt}(\alpha)}{d\nu_A}=(1-\tau)(1+\gamma_A)$$

while

$$\begin{aligned}
\frac{\partial U_{At}(\alpha)}{\partial \nu_A} &= (1-\tau)\left(-\alpha_A\right)+(1-\tau)(1-p)\frac{\theta}{C}(\alpha+\gamma_A)2\left(\frac{1}{2}\alpha_A(2\nu_A-1)-\alpha_P(2\nu_P-1)\right) \\
&\quad +(1-\tau)(1-p)\frac{\theta}{C}(\alpha+\gamma_A)(2\nu_A-1)\alpha_A \\
&= -(1-\tau)\alpha_A+2(1-\tau)(1-p)\frac{\theta}{C}\alpha_A\left(\alpha_A(2\nu_A-1)-\alpha_P(2\nu_P-1)\right)
\end{aligned}$$

If the last term is negative, then $\frac{dU_{At}(\hat{\alpha})-U_{NAt}}{d\nu_A} < 0$ since both terms are negative. If the last term is positive, then, using $z_\alpha < C$ or $\theta\alpha_A(2\nu_A-1) < C$,

$$\frac{\partial U_{At}(\alpha)}{\partial \nu_A} \leq (1-\tau)\left[2(1-p)\left(\alpha_A-\alpha_P\frac{(2\nu_P-1)}{(2\nu_A-1)}\right)-\alpha_A\right] \leq (1-\tau)\alpha_A\left[2(1-p)-1\right] < (1-\tau)(1+\gamma_A)$$

[13] Actually, if $g = f = 0$, then all $\alpha \in [0, 1]$ are solutions, but given indifference, it is sufficient to look at the extremes, and since $U_{At}(\alpha = 1) < U_{NAt}$, *At* will always be dominated in this case.

so that again $\frac{dU_{At}(\hat{\alpha})-U_{NAt}}{d\nu_A} < 0$.
Consider next γ_A. In this case,

$$\frac{dU_{NAt}(\alpha)}{d\gamma_A} = (1-\tau)\nu_A + (1+\gamma_A)\tau$$

while

$$\begin{aligned}
\frac{\partial U_{At}(\alpha)}{\partial \gamma_A} &= (1-\tau)\left((1-\nu_A)\right) + (1-\tau)(1-p)\frac{\theta}{C}\left(\alpha_A(2\nu_A-1)^2 - \alpha_P(2\nu_A-1)(2\nu_P-1)\right) + \frac{2\alpha_A+2\alpha_P}{2}\tau \\
&= (1-\tau)(1-\nu_A) + (1-\tau)(1-p)\frac{\theta}{C}(2\nu_A-1)\left(\alpha_A(2\nu_A-1) - \alpha_P(2\nu_P-1)\right) + (1+\gamma_A)\tau
\end{aligned}$$

If the second term is negative then $\frac{dU_{At}(\hat{\alpha})-U_{NAt}}{d\gamma_A} < 0$. If the second term is positive, then

$$\begin{aligned}
\frac{\partial U_{At}(\alpha)}{\partial \gamma_A} &\leq (1-\tau)(1-\nu_A) + (1-\tau)(1-p)(2\nu_A-1)\left(1 - \frac{\alpha_P(2\nu_P-1)}{\alpha_A(2\nu_A-1)}\right) + (1+\gamma_A)\tau \\
&\leq (1-\tau)\left[(1-\nu_A) + (1-p)(2\nu_A-1)\right] + (1+\gamma_A)\tau \\
&< (1-\tau)\nu_A + (1+\gamma_A)\tau
\end{aligned}$$

which implies again $\frac{dU_{At}(\hat{\alpha})-U_{NAt}}{d\gamma_A} < 0$.
Consider next ν_P. Note that U_{NAt} is independent of ν_P. So it suffices to show that $\frac{\partial U_{At}(\hat{\alpha})}{\partial \nu_P} > 0$ whenever $U_{At}(\hat{\alpha}) = U_{NAt}$. To this end, note

$$\begin{aligned}
\frac{\partial U_{At}(\alpha)}{\partial \nu_P} &= (1-\tau)\alpha_P + (1-\tau)(1-p)\frac{\theta}{C}\alpha_A(2\nu_A-1)(-2\alpha_P) \\
&= (1-\tau)\alpha_P\left(1 - 2(1-p)\frac{\theta}{C}\alpha_A(2\nu_A-1)\right) \\
&= (1-\tau)\alpha_P\left(1 - 2(1-p)\frac{z_\alpha}{C}\right)
\end{aligned}$$

When this is (weakly) negative, then $2(1-p)\frac{z_\alpha}{C} \geq 1$ so that (using also $z_\alpha \leq C$)

$$\begin{aligned}
U_{At}(\alpha) &= (1-\tau)\left(\alpha_A(1-\nu_A) + \alpha_P\nu_P\right) + (1-\tau)(1-p)\frac{z_\alpha}{C}\left(\frac{1}{2}\alpha_A(2\nu_A-1) - \alpha_P(2\nu_P-1)\right) + \frac{\alpha_A^2+2\alpha_A\alpha_P}{2}\tau \\
&\leq (1-\tau)\left(\alpha_A(1-\nu_A) + \alpha_P\nu_P\right) + (1-\tau)\alpha_A(\nu_A - \frac{1}{2}) - (1-\tau)(1-p)\frac{1}{(1-p)2}\alpha_P(2\nu_P-1) + \frac{(1+\gamma_A)^2}{2}\tau \\
&= (1-\tau)\frac{1+\gamma_A}{2} + \frac{(1+\gamma_A)^2}{2}\tau \\
&< (1-\tau)(1+\gamma_A)\nu_A + \frac{(1+\gamma_A)^2}{2}\tau = U_{NAt}
\end{aligned}$$

It follows that, whenever $U_{At}(\hat{\alpha}) \geq U_{NAt}$, $\frac{dU_{At}(\hat{\alpha})-U_{NAt}}{d\nu_P} > 0$. So it follows that there exists a $\hat{\nu}_P$ such that the equilibrium is *At* when $\nu_P \geq \hat{\nu}_P$, and it is *NAt* otherwise.
Consider finally τ. Note that we can write $U_{At} = (1-\tau)A + \tau B$ and $U_{NAt} = (1-\tau)E + \tau F$, where $F > B$. It follows that when $U_{At} = U_{NAt}$, $E < A$. The derivative can now be written

$$\frac{\partial U_{At}(\hat{\alpha}) - U_{NAt}}{\partial \tau} = -A + B + E - F < 0$$

The second part of the proposition then follows. ■

B A Model with Explicit Orders

The models in sections 2 and 3 trade off realism for transparency and simplicity. One of the key concessions in this respect is the assumption that the players' beliefs are common knowledge (and that they are known to disagree). In reality, people cannot read each others' minds, and beliefs are thus private information. Moreover, people sometimes agree and sometimes disagree, and many of the contentious issues arise during the execution of the project, long after the contract has been signed. The purpose of this appendix is to present an extension of the model in this sense and show that the results still hold.

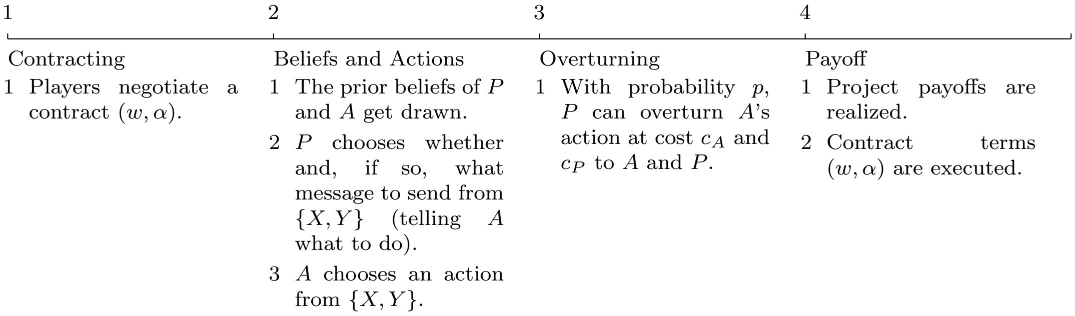
Consider therefore the model of section 2 with the following modifications. The most important modification is that beliefs are not common knowledge any more, but are randomly drawn and private information to the players. The players can communicate about these beliefs through cheap talk. In particular, at the start of period 2 the beliefs get drawn from a random distribution, with μ_i denoting i's belief that the state is x. The idea of having the beliefs being drawn after the contract negotiation, is that the contentious issues, on which the players may disagree, arise only after the project has been started, so that it is only at that time that it becomes clear which beliefs are relevant. To keep the analysis transparent and tractable, I will assume a very simple degenerate distribution for these prior beliefs. In particular, for some given parameters $\nu_i \in (.5, 1)$, μ_i will be either ν_i or $1 - \nu_i$, with equal probability. In other words, μ_i is drawn from a 2-point distribution with half its weight on ν_i and half on $1 - \nu_i$. It follows that the player always has the same strength of belief, ν_i, in the state that he considers most likely. Moreover, each player will believe half the time that the state is x, and half the time that the state is y. The prior beliefs will be independent draws, so that the players will disagree half the time. The expected revenue according to i is ν_i when i believes that the decision is right, and $(1 - \nu_i)$ otherwise.

The beliefs are originally private information. In the second step of period 2, however, P has a chance to tell A what to do. In particular, P chooses whether and, if so, what message to send from the set $\{X, Y\}$. These messages can best be interpreted as respectively 'you should do X' and 'you should do Y'. While sending these messages is costless, I will assume that P has a lexicographic preference for being obeyed. In particular, I will assume that P prefers 'giving an order and (always) being obeyed' over 'not giving an order', and prefers 'not giving an order' over 'giving an order and (ever) being disobeyed'. For simplicity, I will also limit attention to pure-strategy equilibria that are not Pareto-dominated, which will imply that the communications will reveal the players' beliefs truthfully, if at all. The full timing of the game is shown in figure 4.

I have to be careful here when I use the term 'obey'. In particular, I will reserve the term 'obey' for the case in which A does what P (literally) tells him to do. Note that this is not necessarily the same as what P wants him to do (since the order is cheap talk and may thus differ from P's true preferences). Whenever confusion is possible and A does what P *wants* him to do, I will simply say so.

As in section 2, I will assume that c_A is exogenously given, with $c_A > \theta\gamma_A(2\nu_A - 1)$, and that $\tau = 0$. I will now define the *At* and *NAt* equilibria as follows. In the *NAt* equilibrium, P never tells A what to do and A always chooses the action that he considers most likely to succeed. In the *At* equilibrium, P always tells A what to do (and the order is always what P wants him to do), A always obeys P's order, and if A ever does not obey, then P will overturn A's decision (when she can). The following proposition shows that the result is identical to that in section 2.

Proposition 7 *There exists a $\hat{\nu}_P$ such that the (pure-strategy, Pareto-optimal) equilibrium is At with $\alpha = 0$ when $\nu_P > \hat{\nu}_P$, while it is NAt with $\alpha = 1$ when $\nu_P < \hat{\nu}_P$. The value of $\hat{\nu}_P$ increases in*



1 Contracting

1 Players negotiate a contract (w, α).

2 Beliefs and Actions

1 The prior beliefs of P and A get drawn.

2 P chooses whether and, if so, what message to send from $\{X, Y\}$ (telling A what to do).

3 A chooses an action from $\{X, Y\}$.

3 Overturning

1 With probability p, P can overturn A's action at cost c_A and c_P to A and P.

4 Payoff

1 Project payoffs are realized.

2 Contract terms (w, α) are executed.

Figure 4: Time line of basic model with explicit orders

ν_A and γ_A.

Proof : Let B_i denote the action that player i believes is most likely to succeed and $\overline{B}_i$ the opposite action. Consider first the possible equilibria in which P makes an announcement. Since I consider only pure-strategy equilibria, there are four possibilities. The first two are that P always announces X or that P always announces Y. In either case, the announcement contains no information and is thus equivalent with no announcement at all, which I consider below. (From the analysis below, it also follows that P would be disobeyed half the time, and would thus prefer not to give any orders at all.) The two other possibilities are that P either always announces B_P or always announces $\overline{B}_P$. In either case, the announcement fully reveals P's beliefs, so that P's beliefs become common knowledge. The analysis of section 2 then goes through unchanged. In particular, there are again two possible equilibria: one in which $\alpha = 0$ and A always does what P believes is best, and another one in which $\alpha = 1$ and A always does as he himself believes is best. Moreover, when A does as P wants, the equilibrium in which P announces B_P Pareto-dominates the equilibrium in which she announces $\overline{B}_P$ (since in the former, her orders get obeyed while in the latter, her orders get disobeyed). Finally, in the case that A does what he believes is right, P's orders are always disobeyed half the time, so that P would prefer not to make any announcement (so that this equilibrium will be Pareto dominated by an equilibrium that follows below).

Consider now the case in which P never announces her beliefs. In that case, there are 4 possible strategies for A: always choose X, always choose Y, always choose B_A, always choose $\overline{B}_A$. An equilibrium with 'A always chooses B_A' Pareto-dominates all three others. There are two strategies for P: P can either try to overturn a $\overline{B}_P$-choice (when she gets the chance) or not, depending on whether $\alpha < 1 - \frac{c_P}{(2\nu_P - 1)}$ or not. Any potential equilibrium with $\alpha < 1 - \frac{c_P}{(2\nu_P - 1)}$ (so that P overturns when she can) is dominated either by the equilibrium (derived above) in which P always announces B_P and A chooses B_P, or by an equilibrium with $\alpha = 1$, which I discuss now. When $\alpha > 1 - \frac{c_P}{(2\nu_P - 1)}$ (so that P never overturns), total utility is maximized at $\alpha = 1$. This latter equilibrium dominates the earlier equilibrium in which P made announcements but A does as he himself believes is right (since in that case A always disobeys P's orders half the time).

So the only possible (pure strategy, Pareto-optimal) equilibria are *At* and *NAt* as defined above. Since the payoffs are $U_{At} = \gamma_A(1 - \nu_A) + \nu_P$ and $U_{NAt} = (1 + \gamma_A)\nu_A$, the rest of the proposition follows. ■

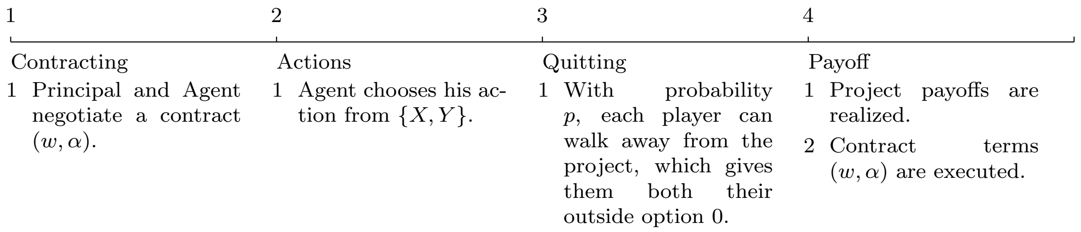

1
2
3
4
Contracting
1 Principal and Agent negotiate a contract (w, α).
Actions
1 Agent chooses his action from $\{X, Y\}$.
Quitting
1 With probability p, each player can walk away from the project, which gives them both their outside option 0.
Payoff
1 Project payoffs are realized.
2 Contract terms (w, α) are executed.

Figure 5: Time line of the efficiency-wage model

C An Efficiency-Wage Model

Sometimes, it is more natural to assume that the principal can stop the project or fire the agent, rather than being able to overturn the agent's decision. In this appendix, I show that very similar results obtain for such a situation. A key benefit of this model is that it easily relates to well-known ideas about authority, in particular to the ideas on efficiency wages (Shapiro and Stiglitz 1984). I will show, for example, that the efficiency wage will increase in the agent's share of the project: as you pay the agent more residual income, you also need to pay him a higher fixed wage.

The formal setting is identical to that of section 2, except for one change. In period 3, there is a probability p that employment becomes at will, i.e., with probability p each player can quit the project. If either player quits the project, then both get their outside option which equals zero. In that case, the project gets cancelled and the contract is void. Figure 5 depicts the timing of this game. For simplicity, I restrict the analysis to pure-strategy, Pareto-optimal equilibria.

As mentioned earlier, the negotiated wage will now also play the role of efficiency wage: the principal pays the agent more than the market wage, in order to have something to punish the agent. In particular, the following proposition identifies the contract terms that cause the agent to obey the principal. The proof is available from the author.

Proposition 8 *There exists a (pure-strategy, Pareto-optimal) equilibrium (for the subgame that starts in period 2) in which A does what P wants him to do, and neither player quits (in equilibrium) if and only if the following conditions are satisfied:*

$$\begin{aligned} w &\geq \alpha_A(\theta(2\nu_A - 1) - (1 - \nu_A)) && (1) \\ w &\geq \alpha_P(1 - \nu_P) && (2) \\ w &\leq \alpha_P \nu_P && (3) \end{aligned}$$

where $\theta = \frac{(1-p)}{p}$.

The first condition is the efficiency wage condition that makes it incentive compatible for the agent to 'obey' the principal. The second condition commits the principal to firing an agent who disobeys: it guarantees that the wage is so high that the principal only wants to continue if the agent did obey. The third condition is a simple individual rationality condition. Note that the efficiency wage indeed increases in α_A (for sufficiently low values of p), so that higher pay-for-performance makes it more difficult to generate authority and obedience.

There are again two equilibria. In analogy to before, I will use *At* ('Authority') to describe the following equilibrium:

- Principal and agent agree on a contract in which $\alpha = 0$.
- The agent does what the principal thinks is best.
- The principal quits if (and only if) she observes that the agent did not act as the principal wanted him to do.

and I will use *NAt* ('No Authority') to describe the following equilibrium:

- Principal and agent agree on a contract in which $\alpha = 1$.
- The agent chooses the action that he believes has the highest probability of success.
- Neither player quits in equilibrium or as a response to a deviation by the other.

The following proposition then says that these two equilibria are the only possible (pure-strategy, Pareto-dominant) equilibria, and that interpersonal authority will be *more* likely when the agent has weaker beliefs, less private benefits at stake, and when P has an easier time observing A's actions. The proof is again available from the author.

Proposition 9 *For any set of parameters, the (pure-strategy, Pareto-optimal) equilibrium exists and is either At or NAt. There exists $\hat{\nu}_P$ such that the (only) equilibrium is At when $\nu_P > \hat{\nu}_P$ and the (only) equilibrium is NAt when $\nu_P < \hat{\nu}_P$. The value of $\hat{\nu}_P$ increases as γ_A or ν_A increase.*

Since the equilibrium is very similar to proposition 2, the same comments and implications apply here.

References

Aghion, P., AND J. Tirole (1997): "Formal and real authority in organizations," *Journal of Political Economy*, 105(1), 1– 29.

Alchian, A. A., AND H. Demsetz (1972): "Production, Information Costs, and Economic Organization," *American Economic Review*, 62, 777– 795.

Baker, G., R. Gibbons, AND K. J. Murphy (1994): "Subjective Performance Measures in Optimal Incentive Contracts," *Quarterly Journal of Economics*, 109(4), 1125– 1156.

Baker, G. P. (1992): "Incentive contracts and performance measurement," *Journal of Political Economy*, 100(3), 598– 614.

Barberis, N. C., AND R. H. Thaler (2003): "A Survey of Behavioral Finance," in *Handbook of the Economics of Finance 1B*, ed. by G. M. Constantinides, M. Harris, and R. M. Stulz. Elsevier North Holland.

Barnard, C. (1938): *The Functions of the Executive*. Harvard University Press, Cambridge MA.

Barry, J. W., AND P. Henry (1981): *Effective Sales Incentive Compensation*. McGraw-Hill, New York.

Berg, N. A., AND N. D. Fast (1983): "Lincoln Electric Co," HBS Case 9-376-028.

Brickley, J. A., AND F. H. Dark (1987): "The Choice of Organizational Form: The Case of Franchising," *Journal of Financial Economics*, 18, 401–420.

Cooper, A. C., W. C. Dunkelberg, AND C. Y. Woo (1988): "Entrepreneurs' Perceived Chances for Success," *Journal of Business Venturing*, 3(3), 97– 108.

Fudenberg, D., AND J. Tirole (1990): "Moral Hazard and Renegotiation in Agency Contracts," *Econometrica*, 58(6), 1279– 1319.

Harsanyi, J. C. (1968): "Games with Incomplete Information Played by 'Bayesian' Players, I-III, Part III. The Basic Probability Distribution of the Game," *Management Science*, 14(7), 486– 502.

Holmstrom, B., AND P. Milgrom (1991): "Multi-Task Principal Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, and Organization*, 7, 24– 52.

——— (1994): "The firm as an incentive system," *American Economic Review*, 84(4), 972– 991.

Landier, A., AND D. Thesmar (2004): "Financial Contracting with Optimistic Entrepreneurs: Theory and Evidence," Working Paper NYU Stern - ENSAE.

Legros, P., AND A. Newman (2002): "Courts, Contracts, and Interference," *European Economic Review*, 46(4), 734– 744.

MacLeod, W. B., AND J. M. Malcomsom (1989): "Implicit Contracts, Incentive Compatibility, and Involuntary Unemployment," *Econometrica*, 57(2), 447– 480.

——— (1998): "Motivation and Markets," *American Economic Review*, 88, 388– 411.

Manove, M., AND A. J. Padilla (1999): "Banking (Conservatively) with Optimists," *Rand Journal of Economics*, 30(2), 324– 350.

Martin, R. (1988): "Franchising and Risk Management," *American Economic Review*, 78(5), 954– 968.

McClelland, D. C. (1964): *Power: The Inner Experience*. Irvington Publishers, New York.

Morris, S. (1995): "The Common Prior Assumption in Economic Theory," *Economics and Philosophy*, 11, 227– 253.

Oliver, R. L., AND E. Anderson (1994): "An Empirical Test of the Consequences of Behavior- and Outcome-Based Sales Control Systems," *Journal of Marketing*, 58(4), 53– 67.

Prendergast, C. (2002): "The Tenuous Trade-off between Risk and Incentives," *Journal of Political Economy*, 110(5), 1071– 1102.

Shapiro, C., AND J. E. Stiglitz (1984): "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review*, 74(3), 433– 444.

Simon, H. (1947): *Administrative Behavior*. Free Press, New York.

——— (1951): "A Formal Theory of the Employment Relationship," *Econometrica*, 19, 293– 305.

Suvorov, A., AND J. van de Ven (2005): "Discretionary Bonuses as a Feedback Mechanism," Working paper.

Van den Steen, E. J. (2004): "Culture Clash: The Costs and Benefits of Homogeneity," Working paper.

——— (2005): "Notes on Modelling with Differing or Heterogeneous Priors," Working Paper, MIT-Sloan.