The Wrong Kind of Transparency

Andrea Prat[1]
London School of Economics[2]

January 9, 2003

[1]I thank Heski Bar-Isaac, Roman Inderst, Alessandro Lizzeri, George Mailath, Margaret Meyer, Stephen Morris, Marco Ottaviani, David Stavasage, and audiences in Edinburgh, Exeter, Newcastle, Nuffield College, Rochester (Wallis Workshop 2002), Stockholm University, and Toulouse for useful discussions.

[2]STICERD, Houghton Street, London WC2A 2AE, United Kingdom. Email: a.prat@lse.ac.uk. Homepage: econ.lse.ac.uk/staff/prat.

Abstract

In a model of career concerns for experts, when is a principal hurt from observing more information about her agent? This paper introduces a distinction between information on the consequence of the agent's action and information directly on the agent's action. When the latter kind of information is available, the agent faces an incentive to disregard useful private signals and act according to how an able agent is expected to act a priori. This conformist behavior hurts the principal in two ways: the decision made by the agent is less likely to be the right one (discipline) and ex post it is more difficult to evaluate the agent's ability (sorting). The paper identifies a necessary and sufficient condition on the agent signal structure under which transparency on action is detrimental to the principal. The paper also shows the existence of complementarities between transparency on action and transparency on consequence. The results on the distinction between transparency on action and transparency on consequence are then used to interpret existing disclosure policies in politics, corporate governance, and delegated portfolio management.

1 Introduction

There is a widespread perception, especially among economists, that transparency is a beneficial element in agency relationships because more information about the agent makes the agent more accountable to the principal. Holmström [18] has shown that in moral hazard principal-agent problems more information about the agent is never detrimental to the principal, and, under mild assumptions, it is strictly beneficial. Should one conclude that whenever it is technologically feasible and not extremely expensive the principal should observe everything that the agent does?

Before asking whether transparency may not be optimal, let us note that in practice we observe systematic deviations from transparency in agency relationships in delegated portfolio management, corporate governance, and politics.

In delegated portfolio management, one might expect a high degree of transparency between the principal (the fund manager) and the agent (the investor). Instead, investors are typically supplied with limited information on the composition of the fund they own. Currently, the US Securities and Exchange Commission requires disclosure every six months, which consists of a portfolio snapshot at a particular point in time and can easily be manipulated by re-adjusting the composition just before and after the snapshot is taken – a practice known as "window dressing". It would be easy and almost costless to have more frequent disclosure by requiring mutual funds to publicize their portfolio composition on the internet. Yet there is strong resistance from the industry to proposals in the direction of more frequent disclosure (Tyle [39]).

In corporate governance, violations to the transparency principle are so widespread that some legal scholars argue that secrecy is the norm rather than the exception in the relation between shareholders and managers (Stevensons [38, p. 6]): "Corporations – even the largest among them – have always been treated by the legal system as 'private' institutions. When questions about the availability of corporate information have arisen, the inquiry has typically begun from the premise that corporations, like individuals, are entitled to keep secret all information they are able to secure physically unless some particular reason for disclosure [...] could be adduced in support of a contrary rule. So deeply embedded in our world view is this principle that it is not at all uncommon to hear serious discussions of a corporate 'right to privacy'."

In politics, the principle of open government has made great inroads in the last decades but there are still important areas in which public decision-making is, by law, protected by secrecy. In the United States, the "executive privilege" allows the president to withhold information from the Congress, the courts, and the public (Rozell [35]). While the executive privilege cannot be used arbitrarily and fell in disrepute during the Watergate scandal, the Supreme Court recognized its validity (US *vs.* Nixon, 1974). In the European Union, the most powerful legislative body, the Council, has a policy of holding meetings behind closed doors and not publishing the minutes

(Calleo [4]). Over thirty countries have passed Open Government codes, which establish the principle that a citizen should be able to access any public document. There are, however, important types of information, such as pre-decision material, that are often exempt from this requirement (Frankel [16]).[1]

Are the observed deviations from transparency in some sense optimal, or are they just due to inefficient arrangements, that survive because of institutional inertia or resistance from entrenched interests? To answer this question, we need to establish what arguments can be made against transparency.

One obvious candidate explanation is that information revealed to the principal would also be revealed to a third party who will make use of it in ways that hurt the principal. In the political arena, voters may choose to ignore information pertaining to national security to prevent hostile countries from learning them as well. In the corporate world, shareholders may wish to keep non-patentable information secret rather than risk that competitors learn it. In delegated portfolio management, *real time* disclosure could damage a fund because its investment strategy could be mimicked or even anticipated by competitors.[2]

The "third-party rationale" for keeping information secret presumably entails a tradeoff between damage from information leaks and weaker incentives for the agent. This paper will instead look for an "agency rationale": a desire for secrecy that stems purely from incentive considerations. The conjecture is that in some circumstances revealing more information about the agent makes the agent's interest less aligned with the principal's interest. Holmström's [18] results on the optimality of information revelation in moral hazard problems suggest that the agency rationale should be explored in contexts in which, for exogenous reasons, there is no full contracting on observables. We will focus our attention on career concern models (Holmström [19]), in which the principal and the agent can sign only short-term non-contingent contracts.[3]

The agency literature has already identified instances in which more information can hurt the principal. Holmström [19] noted that more precise information about the agent's type reduces the incentive for the agent to work hard in order to prove his worth. Dewatripont, Jewitt and Tirole [10] present examples in which the agent works harder if the principal receives a coarser signal on agent performance rather than observing performance directly. Crémer [8] shows that in a dynamic contracting model where renegotiation is possible the principal may be hurt by observing a precise signal on agent performance because it makes the commitment to non-renegotiation less

[1] Section 7 returns to these non-disclosure policies and re-interprets them in the context of the present model.

[2] However, the SEC proposed reform allows for a time lag – usually sixty days – that is judged to be sufficient to neutralize free riding and front running.

[3] As Gibbons and Murphy [17] show, there are still strong career concern incentives even when contracts are contingent on observables. Thus, the crucial assumption we make is that long-term contracts are not available.

credible. In these three instances, more information is bad for discipline (the agent works less) but it is good for sorting (it is easier to identify agent type).

The rationale for secrecy considered in the present paper is entirely different. It does not hinge on the risk that the agent exerts less effort, like in the papers above, but rather on the possibility that the agent disregards useful private signals. In a nutshell, we show that the availability of a certain kind of information may induce the agent to behave in a conformist way. This hurts the principal both through discipline (the agent's action is less aligned with the principal's interest) and sorting (it is impossible to discern the agent's ability). In the following paragraphs, we provide a brief, informal description of the model and the main findings.

This paper employs a model of career concerns for experts (Scharfstein and Stein [37], Prendergast and Stole [34], Ottaviani and Sørensen [29]). What differentiates a good agent from a bad agent is his ability to understand the state of the world, which can be interpreted as expertise, intelligence, or vision. Expert models have been used to represent agency relationships in each of the three areas we draw examples from: politics, corporate governance, and financial intermediation. There are two periods: the current period and the future period. In the current period, an agent (the expert) is in charge of taking an action on behalf of the principal.[4] The agent has no intrinsic preferences among possible actions, i.e. there is no moral hazard in a classical sense. The agent receives a signal about the state of the world, whose precision depends on the agent's type. For now we assume that the agent does not know his own type. The action, together with the state of the world, determines a consequence for the principal. At the end of the current period, the principal forms a posterior about the agent's type, based on information available, and she decides whether to keep the current agent or replace him with another, randomly drawn, agent. In the future period, the agent who is in charge faces a similar decision problem. The wage of the agent cannot be made contingent on the agent's current performance. The agent maximizes the probability of keeping his job. The principal cares about the consequence in the current period (discipline) and the consequence in the second period, which in turn depends on the ability of the principal to screen agents by type (sorting).

We distinguish between two kinds of information that the principal can observe: information about the action that the agent takes and information about the consequence of the agent's action. Suppose for now that the principal always observes the consequence but may or may not observe the action (and that a consequence can be generated by more than one action-state pair, so the principal cannot deduce the action from the consequence).

For example, in delegated portfolio management the relevant state of the world is a vector of future asset prices. The agent (the fund manager) receives a private signal on changes in asset

[4] The formalization we use can be interpreted as a reduced form of either a model with only one principal or a model with a market of principals. For now, we adopt the first interpretation.

prices and selects a portfolio of assets on behalf of the investor. A good fund manager differs from a bad one in his ability to make correct predictions on future asset prices, through which he can generate higher returns for his investor. In the second period, the investor is more likely to retain the fund manager if the posterior on his predictive ability is high.[5] In the first period, the fund manager selects the portfolio in order to show his predictive ability. If the investor is risk neutral, the distinction between action and consequence is straightforward. The action is the portfolio that the fund manager selects; the consequence is the return on the portfolio. Suppose for now that the investor always observes the return (which is available on newspapers for major funds). The question we ask is: should the investor also observe the composition of the fund she owns?

A first result is that more information about the action can hurt the principal. To understand this, first note that even if the principal knows the consequence of the agent's action perfectly she still stands to gain from knowing the action because knowing which particular action-state pair has generated the observed consequence helps the principal understand the agent's type. Direct information on the agent's action thus has a potential positive sorting effect. This effect, however, is based on the assumption that the agent's behavior is constant, but clearly an agent who realizes that his action is going to be observed faces a different incentive structure. A crucial observation is that, in a generic model, the possible realizations of the agent's signal can be ranked in order of *smartness*, that is, according to the posterior on the agent's type given the realization of the signal. Good agents are more likely than bad agents to receive smart signals. If in equilibrium the agent's action is informative of his signal, then also all the possible actions can be ranked in order of smartness. The posterior on the agent's type depends on the consequence but also on the smartness of the action. This can create a contradiction. If the smartness component is too strong, the only possible equilibrium is one in which actions cannot be ranked in order of smartness, i.e. an uninformative equilibrium. The agent disregards his private signal and acts in a purely conformist way. If this is the case, the principal is clearly better off committing to keep the action concealed.

To make this line of reasoning more concrete, let us return to the delegated portfolio management example. Suppose that the fund manager has two possible investment strategies: a portfolio oriented toward blue chips and one which is heavy on high-tech smaller firms.[6] There

[5] Delegated portfolio management in the US fits well with the career concern setup because the Investment Advisers Act of 1940 prevents mutual fund managers from receiving additional payments contingent on good returns (Das and Sundaram [9]).

[6] With only two assets, an investor who observes the portfolio return and stock prices can deduce the portfolio composition. So, if we want to keep open the possibility that the investor does not learn portfolio choices, we need to assume that she does not observe stock prices. With at least three assets, this kind of deduction is in general no longer possible. In practice, fund managers are able to construct portfolios by combining hundreds of stocks,

are two states of the world, according to whether blue chips or high techs will yield better returns. The agent receives a private signal on asset prices with two possible realizations, "blue chips" or "high techs," which tells him which portfolio choice maximizes expected return. The crucial assumption is that, in the terminology introduced above, the high-tech realization is the smart one. This means that the fund manager's ability is more important in understanding when to invest in high-tech stocks (perhaps because prospects of small technology firms are more difficult to evaluate). If the investor could observe the fund manager's signal directly, she would form a higher posterior when the realization is "high tech" rather than "blue chip". If the portfolio composition is observed and if the fund manager acts according to the realization of his signal, for any possible return the investor forms a higher posterior if the investor chooses high techs rather than blue chips. But then, the fund manager may have an incentive to choose high techs even when her signal suggests blue chips. In other words, transparency on portfolio induces the fund manager to behave not according to his private signal but according to the investor's prior on how an able fund manager is likely to behave. If this is the case, the only equilibrium is for the fund manager to always pick the same portfolio – a very negative situation for the investor who gets an uninformed choice in the first period and is not able to sort fund managers based on ability. Under these circumstances, the investor should commit not to observe the portfolio composition, in which case the fund manager follows his private signal in making the investment decision. In Section 7 we will relate this theoretical result to evidence suggesting that mutual funds (in which investors only observe returns and bi-annual snapshots) may outperform pension funds (in which investors have access to portfolio composition information).

The core result of the paper is a necessary and sufficient condition under which revealing the agent's action leads to conformism. The condition has to do with the relative smartness of the realizations of the agent's signal. If one realization is much more smart than the others, then the chain of negative effects described above takes place and there are only conformist equilibria. In mathematical terms, the condition is expressed as a bound on the relative informativeness of the different realizations of the agent's signal. This condition implies that the more advantageous it is for the principal to commit to concealment ex ante, the more advantageous it is for her to renege on her commitment ex post and observe the agent's action for sorting purposes.

We also show that there is complementarity between transparency on action and transparency on consequence. The optimal probability that action is observed is nondecreasing in the probability that the consequence is observed. This is because an agent who pretends to have observed the smart realization by playing the action corresponding to the smart realization has a lower probability of obtaining a good consequence than an agent who actually observed the smart re-

and the presence of an inference problem becomes unrealistic.

alization. Thus the cost of pretending to have observed the smart realization is increasing in the probability that consequence is observed.

The main results of the paper are obtained under the assumption that the agent does not know his own type. If the agent has better a priori information about his own type than the principal (self-knowledge), things become more complex because the agent's action reflects both his self-knowledge and his signal. In an extension, we show that self-knowledge reduces the risk that action revelation leads to full conformism. This is because an agent who knows he is good is more reluctant than an agent who knows he is bad to take an action that goes against his signal. If we parameterize self-knowledge along one dimension, we can show that if self-knowledge is below a certain threshold then the results proven in the rest of the paper are still valid.

At this point, it is important to stress that the main message of this paper is not that transparency is a bad thing. There are powerful and sensible arguments in favor of a presumption that full transparency is the optimal policy in most agency contexts. As most Open Government codes recognize, the burden of proof should rest on those who want to limit transparency. This paper identifies one well-defined set of circumstances in which revealing one particular kind of information may be detrimental to the agency relationship. Section 7 will return to this point and argue that the paper can also be used to rebut a certain kind of argument in favor of opaqueness.

The plan of the paper is as follows. The main argument is developed in a simplified environment in which there are two states of the world, two actions, two consequences, and two realizations of the agent signal. This allows for a full characterization of the equilibrium set, which in turn leads to precise welfare results. Later in the paper, we study a very general model and we prove extended versions of some results obtained in the binary model. Section 2 introduces the baseline career concern game and shows how it can be interpreted as the reduced form of two economic situations. Section 3 contains the analysis of the baseline model. We begin with a simple example in which revealing the agent's action generates complete conformism to the detriment of discipline and sorting. The main technical result is a characterization of the set of perfect Bayesian equilibria under the two information scenarios, concealed action and revealed action, which is then used to perform a welfare analysis. Section 4 studies the complementarity between action observation and consequence observation. Section 5 allows the agent to know his own type or have at least some information about it. Section 6 analyzes the general version of the model. Section 7 concludes by using the results of the paper to interpret some existing institutional arrangements in politics, corporate governance, and delegated portfolio management.

1.1 Related literature

There are many works that are somehow related to how agency relationships are affected by changes in the underlying information structure. In "classical" moral hazard principal-agent problems, the question has been resolved by Holmström [18]. Observing an additional signal can never hurt the principal and it is strictly beneficial if and only if the principal does not already observe a signal that is a sufficient statistic for the additional signal.

On the contrary, the literature on career concerns has already several examples in which more information about the agent's behavior hurts the principal. There are three main approaches to model career concerns, depending on whether the agent's type is seen as ability to exert effort (Holmström [19]), congruence of preference with the principal (the multi-period version of Crawford and Sobel [7]), or ability to observe a signal about the state of the world (see references below). In the first approach, actions are costly from the point of view of the agents and the ones that are more beneficial to the principal are more costly for the agent. The second and third approaches typically give rise to cheap talk models, in which the agent faces no direct cost when he takes an action (obviously, the agent can face an indirect cost through the reaction of the principal).

For the first approach, Holmström [19] already provides an example of one kind of information that worsens discipline. If the principal has more prior information about the agent's type, the agent has less incentive to work hard in the current period to signal his type. When one focuses attention to information about the agent's performance, rather than the agent's type, the question of comparing information structures is studied in a general way by Dewatripont, Jewitt and Tirole [10]. They first present two examples in which a more precise signal about the agent's performance reduces discipline. They then find general sufficient conditions under which an additional signal increases effort.

Not unrelated to the first approach is Crémer [8], who studies optimal transparency when contracts are renegotiable. He shows that, in a two-period agency model where renegotiation is possible, the principal may be hurt by a decrease in the cost of observing the agent's performance. This is because improving the ex post information of the principal makes a commitment not to renegotiate less credible.

None of the papers in the first approach present examples in which more information worsens *both* discipline and sorting. There appears to be a trade-off between discipline and sorting. For instance, in Holmström [19] knowing the agent's type destroys effort exertion but makes screening trivial. In the model that is used in the present paper there is no effort exertion and the rationale behind the potentially negative effect of transparency in entirely different.

For the second approach, the question of comparing information structures is briefly discussed

by Morris [26, p 18-19]. There, an agent observes a signal about the state of the world and makes a report to the principal. The principal makes a decision after hearing the agent's report. Then, the state of the world is revealed. A market then forms a posterior on the basis of the agent's report and on the observation of the state. Morris compares this situation with the situation in which the market observes neither the signal nor the state (because principals are short-lived). He shows that observing the state and the signal improves sorting and may improve or worsens the current period decision: while the bad type's decision is more aligned with the principal's preference, the good type may be induced to take an extreme action to separate himself from the bad type (the "political correctness" effect).[7]

The third approach – the expert agent model (Scharfstein and Stein [37], Zwiebel [40], Prendergast and Stole [34], Ottaviani and Sørensen [28] [29], Levy [23]) – is the one that is used here. To the best of my knowledge, there is no paper in this stream comparing the effect of revealing different kinds of information. It is typically assumed that the principal (or the market) observes the agent's action. For instance, in Prendergast and Stole [34], the agent's action – the investment decision – is publicly observed, and in Ottaviani and Sørensen [29] the agent's "action" is the message that the expert sends to the evaluator and it is, by definition, observed. However, there are also models, like Zwiebel [40], in which the agent's action is not observable.[8]

Prendergast [33] analyzes an agency problem in which the agent exerts effort to observe a variable which is of interest to the principal. The principal too receives a signal about the variable and the agent receives a signal about the signal that the principal received. This is not a career concern model, and the principal can offer payments conditional on the agent's report. Prendergast shows that the agent uses his information on the principal's signal to bias his report toward the principal's signal. Misreporting on the part of the agent causes a loss of efficiency. For this reason, the principal may choose to offer the agent a contract in which pay is independent of action. This will induce minimum effort exertion but also full honesty. While the setup is entirely different, the present work shares Prendergast's insight that when the principal attempts to gather information on the agent's signal the agent may have an incentive to distort his signal report. The two works are complementary in that Prendergast focuses on comparing compensation schemes while we are interested in comparing information structures.

Avery and Meyer [1] ask whether in a career concerns for advisors who may be biased (second

[7] Maskin and Tirole [25] use a career concern model of the second kind to explore the issue of the optimal degree of accountability for public decision-makers. In their model, the principal observes the agent's choice with certainty and the consequence of the choice with a certain probability.

[8] In Ottaviani and Sorensen [29], it is immaterial to think about the agent's decision as message transmitted to the principal or an action taken on behalf of the principal. This is not true anymore in the present model when the action is not observed.

approach) it is beneficial from the point of view of principal to keep track of the advisor's past recommendations. They argue that in certain circumstances observing past recommendations worsens discipline and does not improve sorting. Although the setup is quite different, the intuition bears a connection to the present paper. If the advisor knows that his recommendations affect his future career prospects, he may have an incentive to pool on one type of recommendation independently of his private information.[9]

Ely and Välimäki [13] and Ely, Fudenberg, and Levine [12] ask under what conditions incentives for reputation are bad. Ely and Välimäki construct a model with a long-lived expert who can be either a strategic type who is good or a commitment type who is bad and they reach the striking result that an increase in the reputation motive (i.e. a lower discount rate for the agent) reduces the payoffs of all players. Ely, Fudenberg, and Levine generalize Ely and Välimäki's setup to identify a set of conditions under which the bad effects of reputation arise.

Finally, it is worth mentioning a link with the bargaining literature. Perry and Samuelson [30] analyze how the outcome of delegated bargaining depends on whether offers are observable to principals or not. Fingleton and Raith [15] study career concerns for delegated bargaining when the type of bargainers determine their ability of understanding the opponent's valuation. They ask whether bargaining behind closed doors is better or worse from the viewpoint of the principal. If bargaining occurs secretly, the principal is not able to observe offers but only acceptances. Thus, also their paper questions the optimality of transparency in expert models. However – besides the fact that their model is developed in a context of bargaining – their distinction between acceptance and offer does not correspond to our distinction between consequence and action (if the offer is accepted, everything is observed).[10]

2 Model

We first write the agency problem in a detail-free reduced form. We then show how the reduced form corresponds to two economic situations ("expanded forms"), one in which the bargaining power is on the principal side, the other in which it is more on the agent side.

The model presented in this section restricts attention to a binary action space, state space, signal space, and consequence space. Section 6 examines the general case.

[9] Avery and Meyer assume that only the action is observable, not the consequence. The question of whether the distinction between information on action and information on consequence is crucial also in the second approach is still open. See Conclusions.

[10] See also Seidmann [36] for a complete information model of collective decision making in which a non-disclosure policy leads to better outcomes.

2.1 Reduced form

There are a principal and an agent. In this reduced form, the agent is the only one taking an action; the principal's role is limited to forming an expectation. The agent's type $\theta \in \{g, b\}$ is unknown to both players. The probability that $\theta = g$ is $\gamma \in (0, 1)$ and it is common knowledge. The state of the world is $x \in \{0, 1\}$ with $\Pr(x = 1) = p \in (0, 1)$. The random variables x and θ are mutually independent. The agent selects an *action* $a \in \{0, 1\}$. The *consequence* $u(a, x)$ is 1 if $a = x$ and 0 otherwise.

The principal does not know the state of the world. The agent receives a private signal $y \in \{0, 1\}$ that depends on the state of the world and on his type. Let $q_{x\theta} = \Pr(y = 1|x, \theta)$. We assume that

$$0 < q_{0g} < q_{0b} < q_{1b} < q_{1g} < 1. \tag{1}$$

This means that the signal is informative ($\Pr(x = 1|y)$ is increasing in y and $\Pr(x = 0|y)$ is decreasing in y) and that the signal is more informative for the better type ($\Pr(x = y|y, g) > \Pr(x = y|y, b)$).

These assumptions alone are not sufficient to guarantee that the signal is useful. For instance, if the prior p on x is very high or very low, it is optimal to disregard y. To make the problem interesting, we also assume that the signal y is *decision-relevant*, that is:

$$(q_{1g}\gamma + q_{1b}(1 - \gamma))\, p + ((1 - q_{0g})\gamma + (1 - q_{0b})(1 - \gamma))(1 - p) > \max(p, 1 - p). \tag{2}$$

We can show:

Proposition 1 *Condition (2) holds if and only if*

$$\Pr(x = 1|y = 1) > \Pr(x = 0|y = 1),$$

and

$$\Pr(x = 0|y = 0) > \Pr(x = 1|y = 0).$$

Proof. We have

$$\begin{aligned} \Pr(x = 1|y = 1) &> \Pr(x = 0|y = 1) \\ (q_{1g}\gamma + q_{1b}(1 - \gamma))\, p &> (q_{0g}\gamma + q_{0b}(1 - \gamma))(1 - p), \end{aligned}$$

yielding

$$(q_{1g}\gamma + q_{1b}(1 - \gamma))\, p + ((1 - q_{0g})\gamma + (1 - q_{0b})(1 - \gamma))(1 - p) > 1 - p,$$

and

$$\begin{aligned} \Pr(x=0|y=0) &> \Pr(x=1|y=0) \\ ((1-q_{0g})\gamma + (1-q_{0b})(1-\gamma))(1-p) &> ((1-q_{1g})\gamma + (1-q_{1b})(1-\gamma))p, \end{aligned}$$

yielding

$$((1-q_{0g})\gamma + (1-q_{0b})(1-\gamma))(1-p) + (q_{1g}\gamma + q_{1b}(1-\gamma))p > p.$$

■

The mixed strategy of the agent is a pair $\alpha = (\alpha_0, \alpha_1) \in [0,1]^2$, which represents the probability that the agent plays $a = 1$ given the two possible realizations of the signal.

We consider two cases: *concealed* action and *revealed* action. In the first case, the principal observes only the consequence u. In the second case, she observes also the action a.[11]

The principal's posterior probability that the agent's type is g is $\pi(I)$, where I is the information available to the principal. With concealed action, the posterior is

$$\tilde{\pi}(u) = \Pr(\theta = g|u) = \frac{\gamma \Pr(a = x|\alpha, x, \theta = g)\Pr(x)}{\Pr(a = x|\alpha, x)\Pr(x)}.$$

With revealed action, the principal is able to infer x from a and u. The agent's posterior, assuming that a is played in equilibrium with positive probability, is

$$\pi(a,x) = \Pr(\theta = g|a,x) = \frac{\gamma \Pr(a,x|\theta = g)\Pr(x)}{\Pr(a,x)\Pr(x)}.$$

If action a is not played in equilibrium, perfect Bayesian equilibrium imposes no restriction on $\pi(a,x)$.

The payoff to the agent is simply the posterior $\pi(I)$. The payoff to the principal depends on the consequence and on the posterior: $u(a,x) + v(\pi(I))$, where v is a convex function of π (as we shall see in the "long forms", convexity is a natural assumption if the principal uses the posterior for her hiring and firing decisions). Given any equilibrium strategy α^*, the ex ante expected payoff of the agent must be γ, while the ex ante expected payoff of the principal is $w(\alpha^*) = E_{a,x}(u(a,x) + v(\pi(I))|\alpha^*)$. As the agent's expected payoff does not depend on α^*, the expected payoff of the principal can also be taken as total welfare.

A perfect Bayesian equilibrium of this game (whether the action is concealed or revealed) is a mixed-strategy profile (α_0^*, α_1^*) and a posterior $\pi(I)$ for all possible information sets I, such that α_0^* is a best-response for an agent with $y = 0$, α_1^* is a best-response for an agent with $y = 1$, and $\pi(I)$ is consistent with Bayesian updating given (α_0^*, α_1^*). We sometimes refer to a perfect

[11] Section 4 will allow for hybrid disclosure policies, in which the action is observed with a probability ρ_a and the consequence is observed with ρ_u.

Bayesian equilibrium simply as an "equilibrium". An equilibrium is *informative* if $\alpha_0^* \neq \alpha_1^*$ and *pooling* if $\alpha_0^* = \alpha_1^*$. An informative equilibrium is *separating* if either $\alpha_0^* = 0$ and $\alpha_1^* = 1$ or $\alpha_0^* = 1$ and $\alpha_1^* = 0$. An informative equilibrium is *semi-separating* if it is not separating, i.e. if at least one of the two agents uses a mixed strategy. An informative equilibrium is *perverse* if the agent chooses the 'wrong' action given his signal: $\alpha_0^* > \alpha_1^*$.

Let $E_{revealed}$ and $E_{concealed}$ be the sets of perfect Bayesian equilibria in the two possible information scenarios. Given the existence of babbling equilibria, it is clear that the sets are nonenmpty. Let $W_{revealed}$ be the supremum of $w(\alpha^*)$ in $E_{revealed}$ and let $W_{concealed}$ the corresponding value when the action is concealed. The main question that we shall ask is whether $W_{revealed} \geq W_{concealed}$.

Attention should be drawn to two assumptions. First, assuming that the agent maximizes the posterior $\pi(I)$, rather than an arbitrary function of the posterior $\pi(I)$, is not without loss of generality (see Ottaviani and Sørensen [29] for a discussion of this point). As we shall see, the assumption is arbitrary in Expanded Form I but it is somewhat more natural in Form II. The assumption is made by most papers in career concerns because it makes the analysis simpler.[12]

Second, the agent does not know his own type (again, Ottaviani and Sørensen [29] discuss this point). If the agent knew his own type, he could use his action choice as a costly signal of how confident he is of his own information. Section 5 looks at an extension in this sense.

2.2 Expanded form I: Competing agents.

This form is suited to represent a political game, in which agents are competing parties or candidates and the principal is the electorate (see Persson and Tabellini [31] for a discussion of retrospective voting models). In this two-period model, there are two agents and one principal. One agent, the *incumbent*, is available in the first period. The other agent, the *challenger*, appears at the end of the first period. The type of the incumbent is $\theta \in \{g, b\}$, where the probability that $\theta = g$ is γ. The type of the challenger is $\theta_c \in \{g, b\}$, where the probability that $\theta = g$ is γ_c. The principal, as well as the two agents, do not observe the agents types. While γ is known, γ_c is itself a stochastic variable with distribution f, which is revealed at the end of the first period.

In the first period, the incumbent is in charge of a binary policy decision $a \in \{0, 1\}$. The state of the world is $x \in \{0, 1\}$. The agent observes a signal $y \in \{0, 1\}$ according to the conditional probability q described above. We make an additional assumption on q:

$$q_{1b}p + (1 - q_{0b})(1 - p) > \max(p, 1 - p). \tag{3}$$

[12] See Footnote 18 on page 38 (and the appendix) for a discussion of which results are likely to be unaffected if one drops the linearity assumption.

This guarantees that, even in the worst-case scenario (when it is learnt that the agent is for sure a bad type), the signal y is decision-relevant. Condition (3) implies the decision-relevance condition (2). Without this assumption, it may be the case that second-period efficient decision making requires choosing the same action independently of the signal.

The consequence u is 1 if the action matches the state and zero otherwise. At the end of the first period the challenger appears and γ_c is learnt. The principal observes the consequence, and possibly the action as well. She then chooses whether to keep the incumbent or replace him with the challenger.

In the second period, the agent that has been retained faces a decision problem that is similar to the first period. He selects action $\hat{a} \in \{0,1\}$ to match state $\hat{x} \in \{0,1\}$, where the probability that $\hat{x} = 1$ is still p. The agent receives $\hat{y}$ a signal about $\hat{x}$ that is distributed according to $q_{y\theta}$ described above. The consequence $\hat{u}$ is 1 if the action matches the state and zero otherwise.

The payoff to the principal is $u + \delta\hat{u}$, where $\delta \in (0, \infty)$, which captures both the discount rate and the relative importance of the two periods. A $\delta > 1$ occurs when the second period is more important than the first. The payoff to each agent is 1 if he is hired for the second period and zero otherwise (the benefit that the incumbent receives in the first period is normalized to zero). Clearly, this model describes a world of very incomplete contracts. An agent who is hired gets a fixed rent that the principal cannot control. In particular, the principal cannot offer transfers that are conditional on observed performance.

We assume that at the beginning of the first period the interim probability on the challenger's type, γ_c, is uniformly distributed on the unit interval – that is f_c is a uniform distribution with support $(0,1)$. This restriction guarantees that the payoff of the incumbent is linear in the posterior.

To summarize, the timing is as follows:

1. The incumbent observes signal y and selects action a.

2. The consequence u is realized. The challenger appears: his prior γ_c is realized and observed by all. In the concealed action case, the principal observes u. In the revealed action case, the principal observes a and u. The principal forms a posterior π on the incumbent's type and chooses between the incumbent and the challenger.

3. The agent that has been retained observes signal $\hat{y}$ and selects action $\hat{a}$.

4. The consequence $\hat{u}$ is realized.

We start by analyzing the two last stages, which are straightforward. In the second period, the agent that is retained has no career concerns and he is indifferent with regards to the action

he takes. Thus, any strategy is a continuation equilibrium. In line with the rest of the literature on career concerns, we restrict attention to the continuation equilibrium in which the agent acts in the interest of the principal. Given (3), independently of his belief on his own type, the agent selects $\hat{a} = \hat{y}$. Let $\hat{\gamma}$ be the probability that the agent that is retained for the second period is good, as computed by the principal at the beginning of the second period ($\hat{\gamma} = \pi$ if the incumbent is confirmed, $\hat{\gamma} = \gamma_c$ if the challenger is hired). The second-period expected utility of the principal is:

$$\begin{aligned} E(\hat{u}|\hat{\gamma}) &= \Pr(\hat{y} = \hat{x}|\hat{\gamma}) = \hat{\gamma}((1-p)(1-q_{0g}) + pq_{1g}) + (1-\hat{\gamma})((1-p)(1-q_{0b}) + pq_{1b}) \\ &= (1-p)(1-q_{0b}) + pq_{1b} + \hat{\gamma}((1-p)(q_{0b} - q_{0g}) + p(q_{1g} - q_{1b})) \\ &= \bar{Q} + Q\hat{\gamma}. \end{aligned} \tag{4}$$

Thus, $E(\hat{u}|\hat{\gamma})$ is linear and increasing in $\hat{\gamma}$. The principal chooses to retain the agent with the higher probability of being a good type. Therefore, $\hat{\gamma} = \max\{\pi, \gamma_c\}$, which is a convex function of π. Thus, the principal's payoff is a convex function of π. Then, we have proven that given the continuation payoff in the second period, the subgame in the first period can be represented by the reduced form presented above.

2.3 Expanded form II: Competing principals

This form could be taken as a simple representation of a market for skilled labor. Several firms compete to hire a worker with a unique talent. If the firms were identical, it would not matter from an efficiency point of view which firm hires the worker. So, to make sorting relevant from a social viewpoint, we look at an asymmetric setup.

There are three firms (principals), A, B, and C, and one worker (agent). Again, there are two periods. In the first period there is only principal A. As before a, x, y, and u denote first-period variables while $\hat{a}$, $\hat{x}$, $\hat{y}$, and $\hat{u}$ are for the second-period. In the second period, the three principals compete to hire the agent. A principal who does not hire the agent gets a payoff of zero. Principals B and C are "small": they do not incur fixed costs and their payoff is 1 if the consequence matches the state and zero otherwise. Principal A is "large". In order to become active in the first period, she has to pay an upfront cost $f \in (0,1)$. If the action matches the state she gets 2. Otherwise she gets zero. We also keep assumption (3).

Timing is as follows:

1. The first-period state x is realized. The agent works for Principal A. He observes y and chooses a.

2. The consequence u is observed by everyone. In the revealed action case, also a is observed. Each principal makes a wage offer to the agent.

3. The agent chooses one of the three principals. The second-period state x is realized. The agent observes $\hat{y}$ and chooses $\hat{a}$. The consequence for the principal who hired the agent is $\hat{u} = 1$ if $\hat{a} = \hat{x}$ and zero otherwise. If the principal is A, she receives $2\hat{u} - f$. If the principal is B or C, she receives $\hat{u}$.

As before, we focus attention on continuation equilibria in which, whenever indifferent, the agent chooses his action in order to maximize the payoff of the principal who hired him. In the second period, the probability that the agents matches the action $\hat{a}$ to the consequence $\hat{u}$ is, similarly to (4), a linear function of the posterior of the agent π: $\bar{Q} + Q\pi$. In the bidding game at stage 2, Principal A is willing to pay up to $2\left(\bar{Q} + Q\pi\right) - f$, while the other two principals are willing to pay up to $\bar{Q} + Q\pi$. Excluding dominated strategies, the equilibrium bid is $\bar{Q} + Q\pi$. Principal A hires the agent if and only if

$$\pi \geq \frac{f - \bar{Q}}{Q}.$$

The expected payoff of A given π is $\max\left(\bar{Q} + Q\pi - f, 0\right)$. Thus, her expected payoff is convex in the agent's posterior. The agent's payoff is instead just the equilibrium bid $\bar{Q} + Q\pi$, and it is therefore linear in the posterior. Again, given the continuation payoffs in the second period, the first period is strategically equivalent to the reduced form above.

2.4 Smart realization

We introduce a notion that corresponds to a mental experiment. Suppose the principal could observe the agent signal y directly. Which of the two realizations of the signal y is better news about the agent type? This corresponds to comparing $\Pr(\theta = 1|y = 1)$ with $\Pr(\theta = 1|y = 0)$.

We exclude the nongeneric case in which the two probabilities are identical. In such a situation, the posterior about the agent must be equal to the prior and the signalling game is uninteresting. $\Pr(\theta = 1|y = 1) > \Pr(\theta = 1|y = 0)$ we say that $y = 1$ is the *smart realization* of the agent signal. If $\Pr(\theta = 1|y = 1) < \Pr(\theta = 1|y = 0)$, we say that $y = 0$ is the smart realization. The following result relates smartness to the primitives:

Proposition 2 *The smart realization is $y = 1$ if and only if*

$$\frac{q_{0b} - q_{0g}}{q_{1g} - q_{1b}} < \frac{p}{1 - p}.$$

Proof. Note that:

$$\begin{aligned}
\Pr(\theta|y) &= \frac{\Pr(y|g)\Pr(g)}{\Pr(y|g)\Pr(g) + \Pr(y|b)\Pr(b)} \\
&= \frac{\left(\Pr(y|1,g)\Pr(1) + \Pr(y|0,g)\Pr(0)\right)\Pr(g)}{\left(\Pr(y|1,g)\Pr(1) + \Pr(y|0,g)\Pr(0)\right)\Pr(g) + \left(\Pr(y|1,b)\Pr(1) + \Pr(y|0,b)\Pr(0)\right)\Pr(b)}
\end{aligned}$$

$$\Pr(\theta = 1|y=1) = \frac{(q_{1g}p + q_{0g}(1-p))\gamma}{(q_{1g}p + q_{0g}(1-p))\gamma + (q_{1b}p + q_{0b}(1-p))(1-\gamma)}$$
$$\Pr(\theta = 1|y=0) = \frac{((1-q_{1g})p + (1-q_{0g})(1-p))\gamma}{((1-q_{1g})p + (1-q_{0g})(1-p))\gamma + ((1-q_{1b})p + (1-q_{0b})(1-p))(1-\gamma)}$$

Then, $\pi(y=1)$ is greater than $\pi(y=0)$ if and only if

$$\frac{(q_{1g}p + q_{0g}(1-p))\gamma}{(q_{1g}p + q_{0g}(1-p))\gamma + (q_{1b}p + q_{0b}(1-p))(1-\gamma)} > \gamma$$

or

$$\begin{aligned}
q_{1g}p + q_{0g}(1-p) &> (q_{1g}p + q_{0g}(1-p))\gamma + (q_{1b}p + q_{0b}(1-p))(1-\gamma) \\
q_{1g}p + q_{0g}(1-p) &> q_{1b}p + q_{0b}(1-p) \\
(q_{1g} - q_{1b})p &> (q_{0b} - q_{0g})(1-p) \\
\frac{q_{0b} - q_{0g}}{q_{1g} - q_{1b}} &< \frac{p}{1-p}
\end{aligned}$$

■

If the two states of the world are equiprobable, Proposition 2 requires that

$$q_{1g} - q_{1b} > (1 - q_{0g}) - (1 - q_{0b}).$$

That is, the difference between the probability that the good type gets the right signal and the probability that the bad type gets the right signal must be greater if $x = 1$ than if $x = 0$. Then, observing $y = 1$ raises the agent's posterior above γ while observing $y = 0$ decreases it.

If the two states have different probability, then the inequality is:

$$p(q_{1g} - q_{1b}) > (1-p)((1 - q_{0g}) - (1 - q_{0b})).$$

3 Analysis

In this section, we begin with a simple example of how revealing the agent's action generates conformism. We then analyze separately the concealed action scenario and the revealed action scenario. We conclude by identifying a necessary and sufficient condition on the primitives of the game under which action revelation is the optimal policy.

3.1 An example

Suppose that $\gamma = \frac{1}{2}$, $p = \frac{1}{2}$, $q_{0b} = q_{1b} = \frac{1}{2}$, $q_{0g} = \frac{1}{2}$, and $q_{1g} = 1$. A bad agent receives an uninformative signal. A good agent observes the state $x = 1$ with certainty and gets pure noise if the state is $x = 0$. It is easy to check that $y = 1$ is the smart realization.

This setup represents the delegated portfolio management example discussed in the introduction. The state $x = 0$ corresponds to the "boring" situation in which blue chips are the optimal investment strategy, while $x = 1$ is the "exciting" scenario in which high tech smaller firms do better. To make the point as clear as possible, everything is symmetric between 0 and 1 except the agent signal. In the boring blue-chip case, both types of agents have the same precision (a completely useless signal). In the exciting high-tech case, the good agent performs better: he observes $y = 1$ with probability 1. Going backwards, it is easy to see that $y = 1$ is the smart realization. To show this, one can either check the smartness condition in Proposition 2 or just compute the posteriors given the agent signal:

$$\begin{aligned} \Pr(\theta = g | y = 0) &= \frac{1}{3}; \\ \Pr(\theta = g | y = 1) &= \frac{3}{5}. \end{aligned}$$

This asymmetry between signal realizations creates a problem. The signal $y = 0$ is bad news for the ability of the agent, and the agent will try to conceal this information from the principal. As we shall see shortly, this leads to conformism: the agent has an incentive to act as if he had observed the smart realization $y = 1$ even when he observes $y = 0$.

We now argue that in this example the only equilibrium behavior with revealed action involves complete conformism and it is damaging to the principal. We say "argue" rather than "prove" because in this section we restrict attention to pure-strategy equilibria (separating or pooling). The next section will provide a full analysis, including semi-separating equilibria.

First, consider the revealed action scenario and suppose that there exists a separating equilibrium in which the agent plays $a = y$. The principal's belief $\pi(a, x)$ in such a separating equilibrium is:

$$\begin{aligned} \pi(1,1) &= \tfrac{2}{3} \quad \pi(1,0) = \tfrac{1}{2} \\ \pi(0,1) &= 0 \quad \pi(0,0) = \tfrac{1}{2} \end{aligned}$$

The belief when $a = 1$ dominates the one when $a = 0$, in the sense that for any realization of x, $\pi(1, x) \geq \pi(0, x)$. Then,

$$E(\pi(1, x) | y = 0) > E(\pi(0, x) | y = 0),$$

which means that the agent who observes $y = 0$ has a strict incentive to report $a = 1$. This generates a contradiction.

A similar non-existence argument applies to the perverse separating equilibrium in which $a = |1 - y|$. The only remaining pure-strategy equilibria are then pooling equilibria in which no information is revealed (either the agent always plays $a = 0$ or he always plays $a = 1$). It is

easy to check the existence of such equilibria and that the principal is indifferent among them (because $x = 1$ and $x = 0$ are equiprobable).

Thus, with revealed action, the best equilibrium for the principal is one in which her expected payoff in the current period is $\frac{1}{2}$ and her posterior is the same as her prior.

Instead, in the concealed action scenario there exists a separating equilibrium in which the agent plays $a = y$. To see this, compute the agent posterior in such an equilibrium:

$$\tilde{\pi}(1) = \frac{\frac{1}{2}\left(1+\frac{1}{2}\right)}{\frac{3}{4}+\frac{1}{2}} = \frac{3}{5} \qquad \tilde{\pi}(0) = \frac{\frac{1}{2}\left(0+\frac{1}{2}\right)}{\frac{3}{4}+\frac{1}{2}} = \frac{2}{5}.$$

The agent maximizes his expected posterior by maximizing the expected value of u. As the signal y is decision-relevant, this means that the optimal strategy is $a = y$.

In this separating equilibrium, the probability that the principal gets utility 1 in the first period is

$$\Pr(u = 1) = \frac{1}{4}\frac{1}{2} + \frac{1}{4}1 + \frac{1}{4}\frac{1}{2} + \frac{1}{4}\frac{1}{2} = \frac{5}{8}.$$

Thus, with concealed action, the principal receives an expected payoff of $\frac{5}{8}$ in the first period and she learns something about the agent type.

To sum up, by committing to keep the action concealed, the principal gets a double benefit. On the discipline side, she increases her expected payoff in the current period because the agent follows his signal. On the sorting side, she improves the precision of her posterior on her agent type. As her utility is convex in the posterior, this can only be beneficial.

3.2 Concealed action

We now begin the analysis of the game introduced in Section 2. In this section we look at what happens when the principal observes only the consequence u, which turns out to be the easier part.

The principal's posterior after observing the consequence is $\pi(u) = \Pr(\theta = g|u)$. The agent observes his signal y and maximizes $E_x[\pi(u(a,x))|y]$.

The main result is:

Proposition 3 *With concealed action, there exists a separating equilibrium.*

Proof. Consider a separating equilibrium in which $a = y$. The posterior is

$$\begin{aligned}
\tilde{\pi}(u=1) &= \frac{\gamma\left(pq_{1g} + (1-p)(1-q_{0g})\right)}{\gamma\left(pq_{1g} + (1-p)(1-q_{0g})\right) + (1-\gamma)\left(pq_{1b} + (1-p)(1-q_{0b})\right)}; \\
\tilde{\pi}(u=0) &= \frac{\gamma\left((1-p)(1-q_{1g}) + pq_{0g}\right)}{\gamma\left((1-p)(1-q_{1g}) + pq_{0g}\right) + (1-\gamma)\left((1-p)(1-q_{1b}) + pq_{0b}\right)}.
\end{aligned}$$

From (1), we see that $\tilde{\pi}(u=1) > \gamma > \tilde{\pi}(u=0)$. The agent chooses a to maximize $\Pr(u=1|a,y)$. Because of decision relevance (2), this is achieved by selecting $a=y$. ■

The analysis of the concealed action case is straightforward. There exists a separating equilibrium in which the agent follows his signal and the principal puts a higher posterior on an agent who obtains $u=1$ than on one who fails. There may be other equilibria: uninformative, perverse separating, semi-separating. But the separating equilibrium above is clearly the best from the viewpoint of the principal.

3.3 Revealed action

We now consider the case in which the principal observes the action a as well, which turns out to be the harder case because we need to deal with semi-separating equilibria.

We begin by excluding a certain class of mixed-strategy equilibria. There cannot exist an informative equilibrium in which the agent plays a strictly mixed strategy both when $y=0$ and $y=1$:

Proposition 4 *There cannot exist an informative equilibrium in which* $\alpha_0 \in (0,1)$ *and* $\alpha_1 \in (0,1)$.

Proof. Assume that there exists an equilibrium in which:

$$\alpha_0 \in (0,1), \alpha_1 \in (0,1), \alpha_0 \neq \alpha_1,$$

The agent must be indifferent between the two actions for both realizations of y :

$$\Pr(x=0|y=1)(\pi(0,0)-\pi(1,0)) = \Pr(x=1|y=1)(\pi(1,1)-\pi(0,1)), \quad (5)$$

$$\Pr(x=0|y=0)(\pi(0,0)-\pi(1,0)) = \Pr(x=1|y=0)(\pi(1,1)-\pi(0,1)). \quad (6)$$

There are two cases:

$$(\pi(0,0)-\pi(1,0))(\pi(1,1)-\pi(0,1)) \leq 0 \quad (7)$$

$$(\pi(0,0)-\pi(1,0))(\pi(1,1)-\pi(0,1)) > 0 \quad (8)$$

If (7) holds, note that in an informative equilibrium it cannot be that both $\pi(0,0)=\pi(1,0)$ and $\pi(1,1)=\pi(0,1)$. But then we have a contradiction because the two sides of (5) have different signs.

If (8) holds, subtract (6) from (5)

$$(\Pr(x=0|y=1)-\Pr(x=0|y=0))(\pi(0,0)-\pi(1,0)) \quad (9)$$
$$= (\Pr(x=1|y=1)-\Pr(x=1|y=0))(\pi(1,1)-\pi(0,1)).$$

But by assumption (1) signals are informative on x:

$$\begin{aligned} \Pr(x=0|y=1) - \Pr(x=0|y=0) &< 0; \\ \Pr(x=1|y=1) - \Pr(x=1|y=0) &> 0. \end{aligned}$$

Then, (8) creates a contradiction in (9). ■

This kind of result is common to many signalling games. If there existed an informative equilibrium in which both α_0 and α_1 are interior, the agent would always be indifferent between playing 0 or 1. But this can be true only if signals are uninformative, which contradicts our assumptions on q, or if posteriors are flat, which cannot be true in an informative equilibrium.

We now provide another result on the characterization of the equilibrium set. If there exists an informative equilibrium, then there must also exist a (non-perverse) separating equilibrium:[13]

Proposition 5 *There exists an equilibrium in which $\alpha_0 \neq \alpha_1$ if and only if there exists an equilibrium in which $\alpha_0 = 0$ and $\alpha_1 = 1$.*

Proof. We begin by expressing beliefs in terms of primitives and strategies. It is useful to make the dependence on strategies explicit (we use Π rather than π):

$$\Pi(1,x,\alpha_0,\alpha_1) = \frac{(\alpha_1 q_{xg} + \alpha_0(1-q_{xg}))\gamma}{(\alpha_1 q_{xg} + \alpha_0(1-q_{xg}))\gamma + (\alpha_1 q_{xb} + \alpha_0(1-q_{xb}))(1-\gamma)}; \tag{10}$$

$$\Pi(0,x,\alpha_0,\alpha_1) = \frac{((1-\alpha_1)q_{xg} + (1-\alpha_0)(1-q_{xg}))\gamma}{((1-\alpha_1)q_{xg} + (1-\alpha_0)(1-q_{xg}))\gamma + ((1-\alpha_1)q_{xb} + (1-\alpha_0)(1-q_{xb}))(1-\gamma)} \tag{11}$$

To simplify notation in the proof, we use the following (slightly abusive) notation for special cases of $\Pi(a,x,\alpha_0,\alpha_1)$:

$$\begin{aligned} \Pi(a,x) &\equiv \Pi(a,x,\alpha_0=0,\alpha_1=1) \\ \Pi(a,x,\alpha_1) &\equiv \Pi(a,x,\alpha_0=0,\alpha_1) \\ \Pi(a,x,\alpha_0) &\equiv \Pi(a,x,\alpha_0,\alpha_1=1) \end{aligned}$$

Throughout the proof, assume without loss of generality that $y=1$ is the smart realization. If $y=0$ is the smart realization, just switch 0 and 1 for a, x, and y.

We begin by considering perverse informative equilibria. Suppose there exists an equilibrium in which $\alpha_0 > \alpha_1$, with beliefs $\Pi(a,x,\alpha_0,\alpha_1)$. For $y \in \{0,1\}$, if a is played in equilibrium it must be that:

$$a \in \arg\max_{\tilde{a}} \sum_{x\in\{0,1\}} \Pr(x|y)\,\Pi(\tilde{a},x,\alpha_0,\alpha_1)$$

[13] A similar equilibrium characterization result is found in Ottaviani and Sorensen [28, Lemma 1]. Their setup is a special case of the present one because the agent signal y is symmetric. Using our notation, this corresponds to the restriction $q_{1\theta} = 1 - q_{0\theta}$ for $\theta \in \{b,g\}$.

However, if such equilibrium exists, there also exist an equilibrium in which the agent plays $\hat{\alpha}_0 = \alpha_1$ and $\hat{\alpha}_1 = \alpha_0$, and beliefs are

$$\hat{\Pi}(a, x, \hat{\alpha}_0, \hat{\alpha}_1) = \Pi(1 - a, x, \alpha_0, \alpha_1)$$

The agent's strategy is still a best response: if a is played in equilibrium

$$a \in \arg\max_{\tilde{a}} \sum_{x \in \{0,1\}} \Pr(x|y) \hat{\Pi}(\tilde{a}, x, \alpha_0, \alpha_1)$$

Thus, if there exists a perverse informative equilibrium, there exists a non-perverse informative equilibrium. The rest of the proof focuses on the existence of non-perverse informative equilibria ($\alpha_0 < \alpha$).

We begin with a result on separating equilibria. The necessary and sufficient conditions for the existence of a non-perverse separating equilibrium are:

$$\Pr(x = 1|y = 0)(\Pi(1,1) - \Pi(0,1)) \leq \Pr(x = 0|y = 0)(\Pi(0,0) - \Pi(1,0)) \quad (12)$$

$$\Pr(x = 1|y = 1)(\Pi(1,1) - \Pi(0,1)) \geq \Pr(x = 0|y = 1)(\Pi(0,0) - \Pi(1,0)) \quad (13)$$

Claim 1: The inequality (13) is always satisfied. There exists a separating equilibrium if and only if (12) holds.

Proof of Claim 1: (13) rewrites as:

$$\begin{aligned} & \Pr(x = 1, y = 1)\left(\frac{\Pr(g, y = 1, x = 1)}{\Pr(y = 1, x = 1)} - \frac{\Pr(g, y = 0, x = 1)}{\Pr(y = 0, x = 1)}\right) \\ \geq\ & \Pr(x = 0, y = 1)\left(\frac{\Pr(g, y = 0, x = 0)}{\Pr(y = 0, x = 0)} - \frac{\Pr(g, y = 1, x = 0)}{\Pr(y = 1, x = 0)}\right) \end{aligned}$$

or

$$\begin{aligned} & \Pr(g, y = 1, x = 1) + \Pr(g, y = 1, x = 0) = \Pr(g, y = 1) \\ \geq\ & \frac{\Pr(y = 1, x = 1)}{\Pr(y = 0, x = 1)} \Pr(g, y = 0, x = 1) + \frac{\Pr(y = 1, x = 0)}{\Pr(y = 0, x = 0)} \Pr(g, y = 0, x = 0). \end{aligned}$$

But

$$\begin{aligned} & \frac{\Pr(y = 1|x = 1)}{\Pr(y = 0|x = 1)} \Pr(g, y = 0, x = 1) + \frac{\Pr(y = 1|x = 0)}{\Pr(y = 0|x = 0)} \Pr(g, y = 0, x = 0) \\ \geq\ & \frac{\Pr(y = 1|x = 0)}{\Pr(y = 0|x = 0)} (\Pr(g, y = 0, x = 1) + \Pr(g, y = 0, x = 0)) \\ =\ & \frac{\Pr(x = 0|y = 1)\Pr(y = 1)}{\Pr(x = 0|y = 0)\Pr(y = 0)} (\Pr(g, y = 0, x = 1) + \Pr(g, y = 0, x = 0)) \\ \geq\ & \frac{\Pr(y = 1)}{\Pr(y = 0)} \Pr(g, y = 0), \end{aligned}$$

where the two inequalities are due to assumption (1). This shows that a sufficient condition for (13) is

$$\Pr(g, y=1) \geq \frac{\Pr(y=1)}{\Pr(y=0)} \Pr(g, y=0)$$

But this corresponds to $\Pr(g|y=1) \geq \Pr(g|y=0)$, which is equivalent to the condition that $y=1$ is smart. The claim is proven.

From Proposition 4, there cannot exist an equilibrium in which $0 < \alpha_0 < \alpha_1 < 1$. There can be two cases: either $\alpha_0 = 0$ and $\alpha_1 \in (0,1]$ or $\alpha_0 \in [0,1)$ and $\alpha_1 = 1$. Claims 2 and 3 deal with the two cases separately. Together, the claims prove that there exists an equilibrium with $\alpha_0 < \alpha_1$ only if there exists an equilibrium with $\alpha_0 = 0$ and $\alpha_1 = 1$.

Claim 2: There cannot exist an equilibrium in which $\alpha_0 = 0$ and $\alpha_1 \in (0,1)$.

Proof of Claim 2: Suppose there exists an equilibrium in which $\alpha_0 = 0$ and $\alpha_1 \in (0,1]$. It must be that

$$\Pr(x=1|y=1)\left(\Pi(1,1,\alpha_1) - \Pi(0,1,\alpha_1)\right) = \Pr(x=0|y=1)\left(\Pi(0,0,\alpha_1) - \Pi(1,0,\alpha_1)\right). \quad (14)$$

Note that $\Pi(0,x,\alpha_1) = \Pi(0,x)$ and

$$\begin{aligned}
& \Pi(0,x,\alpha_1) \\
= & \frac{\left(\Pr(y=0|g,x) + (1-\alpha_1)\Pr(y=1|g,x)\right)\Pr(g)}{\Pr(y=0|x) + (1-\alpha_1)\Pr(y=1|x)} \\
= & \frac{\frac{\Pr(y=0|g,x)\Pr(g)}{\Pr(y=0|x)}\Pr(y=0|x) + (1-\alpha_1)\frac{\Pr(y=1|g,x)\Pr(g)}{\Pr(y=1|x)}\Pr(y=1|x)}{\Pr(y=0|x) + (1-\alpha_1)\Pr(y=1|x)} \\
= & A(x,\alpha_1)\Pi(0,x) + (1 - A(x,\alpha_1))\Pi(1,x),
\end{aligned}$$

where

$$A(x,\alpha_1) \equiv \frac{\Pr(y=0|x)}{\Pr(y=0|x) + (1-\alpha_1)\Pr(y=1|x)}.$$

Condition (14) rewrites as

$$\begin{aligned}
& \Pr(x=1|y=1)\, A(1,\alpha_1)\left(\Pi(1,1) - \Pi(0,1)\right) \\
= & \Pr(x=0|y=1)\, A(0,\alpha_1)\left(\Pi(1,0) - \Pi(0,0)\right),
\end{aligned}$$

which in turn is expressed as

$$\Pr(x=1|y=1)\left(\Pi(1,1) - \Pi(0,1)\right) = \Pr(x=0|y=1)\frac{A(0,\alpha_1)}{A(1,\alpha_1)}\left(\Pi(0,0) - \Pi(1,0)\right).$$

Note that

$$\begin{aligned}\max_{\alpha_1}\frac{A(0,\alpha_1)}{A(1,\alpha_1)} &= \max_{\alpha_1}\frac{\Pr(y=0|x=0)}{\Pr(y=0|x=1)}\frac{\Pr(y=0|x=1)+(1-\alpha_1)\Pr(y=1|x=1)}{\Pr(y=0|x=0)+(1-\alpha_1)\Pr(y=1|x=0)}\\ &= \frac{\Pr(y=0|x=0)}{\Pr(y=0|x=1)}\frac{\Pr(y=0|x=1)+\Pr(y=1|x=1)}{\Pr(y=0|x=0)+\Pr(y=1|x=0)}\\ &= \frac{\Pr(y=0|x=0)}{\Pr(y=0|x=1)}.\end{aligned}$$

A necessary condition for (14) to hold is then

$$\Pr(x=1|y=1)(\Pi(1,1)-\Pi(0,1)) \leq \Pr(x=0|y=1)\frac{\Pr(y=0|x=0)}{\Pr(y=0|x=1)}(\Pi(0,0)-\Pi(1,0)).$$

This rewrites as

$$\begin{aligned}&\Pr(x=1)\Pr(y=1|x=1)\Pr(y=0|x=1)(\Pi(1,1)-\Pi(0,1))\\ \leq\ &\Pr(x=0)\Pr(y=1|x=0)\Pr(y=0|x=0)(\Pi(0,0)-\Pi(1,0));\end{aligned}$$

$$\begin{aligned}&\Pr(x=1)(\Pr(y=0|x=1)\Pr(g,y=1|x=1)-\Pr(y=1|x=1)\Pr(g,y=0|x=1))\\ \leq\ &\Pr(x=0)(\Pr(y=1|x=0)\Pr(g,y=1|x=0)-\Pr(y=0|x=0)\Pr(g,y=0|x=0));\end{aligned}$$

Because θ and x are independent,

$$\begin{aligned}&\Pr(x=1)(\Pr(y=0|x=1)\Pr(y=1|g,x=1)-\Pr(y=1|x=1)\Pr(y=0|g,x=1))\\ \leq\ &\Pr(x=0)(\Pr(y=1|x=0)\Pr(y=0|g,x=0)-\Pr(y=0|x=0)\Pr(y=1|g,x=0));\end{aligned}$$

By recalling that $\Pr(y|x)=\gamma\Pr(y|g,x)+(1-\gamma)\gamma\Pr(y|g,x)$, and with some simplification, we get

$$\begin{aligned}&\Pr(x=1)(\Pr(y=0|b,x=1)\Pr(y=1|g,x=1)-\Pr(y=1|b,x=1)\Pr(y=0|g,x=1))\\ \leq\ &\Pr(x=0)(\Pr(y=1|b,x=0)\Pr(y=0|g,x=0)-\Pr(y=0|b,x=0)\Pr(y=1|g,x=0));\end{aligned}$$

$$\begin{aligned}&p((1-q_{1b})q_{1g}-q_{1b}(1-q_{1g}))\\ \leq\ &(1-p)(q_{0b}(1-q_{0g})-(1-q_{b0})q_{0g});\end{aligned}$$

$$p(q_{1g}-q_{1b}) \leq (1-p)(q_{0b}-q_{0g});$$

$$\frac{q_{0b}-q_{0g}}{q_{1g}-q_{1b}} \geq \frac{p}{1-p},$$

which contradicts smartness.

Claim 3: If there exists an equilibrium in which $\alpha_0 \in [0,1)$ and $\alpha_1 = 1$, there exists an equilibrium in which $\alpha_0 = 0$ and $\alpha_1 = 1$.

Proof of the claim: A necessary condition for the existence of an equilibrium in which $\alpha_0 \in [0,1)$ and $\alpha_1 = 1$, is that for some $\alpha_0 \in [0,1)$,

$$\Pr(x=1|y=0)\left(\Pi(1,1,\alpha_0)-\Pi(0,1,\alpha_0)\right) \leq \Pr(x=0|y=0)\left(\Pi(0,0,\alpha_0)-\Pi(1,0,\alpha_0)\right). \quad (15)$$

We have $\Pi(0,x,\alpha_0) = \Pi(0,x)$ and

$$\begin{aligned} & \Pi(1,x,\alpha_0) \\ = & \frac{(\Pr(y=1|g,x)+\alpha_0\Pr(y=0|g,x))\Pr(g)}{(\Pr(y=1|x)+\alpha_0\Pr(y=0|x))} \\ = & \frac{\frac{\Pr(y=1|g,x)\Pr(g)}{\Pr(y=1|x)}\Pr(y=1|x)+\alpha_0\frac{\Pr(y=0|g,x)\Pr(g)}{\Pr(y=0|x)}\Pr(y=0|x)}{(\Pr(y=1|x)+\alpha_0\Pr(y=0|x))} \\ = & B(x,\alpha_0)\Pi(1,x)+(1-B(x,\alpha_0))\Pi(0,x), \end{aligned}$$

where

$$B(x,\alpha_0) = \frac{\Pr(y=1|x)}{(\Pr(y=1|x)+\alpha_0\Pr(y=0|x))}.$$

We can rewrite (15) as

$$\Pr(x=1|y=0)B(1,\alpha_0)\left(\Pi(1,1)-\Pi(0,1)\right) \leq \Pr(x=0|y=0)B(0,\alpha_0)\left(\Pi(0,0)-\Pi(1,0)\right),$$

which in turn holds only if

$$\Pr(x=1|y=0)\left(\Pi(1,1)-\Pi(0,1)\right)\min_{\alpha_0}\frac{B(1,\alpha_0)}{B(0,\alpha_0)} \leq \Pr(x=0|y=0)\left(\Pi(0,0)-\Pi(1,0)\right). \quad (16)$$

But

$$\min_{\alpha_0}\frac{B(1,\alpha_0)}{B(0,\alpha_0)} = \min_{\alpha_0}\frac{\Pr(y=1|x=1)}{\Pr(y=1|x=0)}\frac{\Pr(y=1|x=0)+\alpha_0\Pr(y=0|x=0)}{\Pr(y=1|x=1)+\alpha_0\Pr(y=0|x=1)} = 1,$$

Then (16) rewrites as (12). If (15) holds, (12) holds, and by Claim 1 there exists a n equilibrium in which $\alpha_0 = 0$ and $\alpha_1 = 1$. ■

Proposition 5 says that if the equilibrium set contains some kind of informative equilibrium then it must also contain a non-perverse separating equilibrium. This is a useful characterization because the existence conditions for semi-separating equilibria are hard to find, while the existence conditions for separating equilibria are – as we shall see – straightforward.

The proposition is arrived at in two steps. First, we show that for every perverse informative equilibrium (one in which the agent plays $\alpha_0 > \alpha_1$, i.e. knowingly chooses the wrong action),

there exists a specular non-perverse informative equilibrium. Second, if there exists a non-perverse informative equilibrium, there must also exist a separating equilibrium.

The intuition for the second step has to do with the choice of the agent who observes the non-smart realization (the "non-smart agent"). The incentive of the non-smart agent to follow his own signal depends on the proportion of non-smart agents who follow their own signal. To fix ideas, take $y = 1$ to be the smart action and suppose that $\alpha_1 = 1$. The question is how the posterior depends on the proportion of non-smart agents who pretend to be smart: α_0. Note that α_0 does not affect the posteriors when $a = 0$ because only non-smart agents play $a = 0$. Instead, the higher α_0, the less information the principal can gather about the agent's type when $a = 1$. As $\alpha_0 \to 1$, when $a = 1$ the state x provides no information about the agent's type (because x and θ are independent). In this case, a non-smart agent should certainly choose $a = 1$. This line of reasoning holds for any α_0: the higher α_0, the stronger the incentive for a non-smart agent to play $a = 1$. But then, if the agent is indifferent between $a = 0$ and $a = 1$ for some α_0, he must strictly prefer $a = 0$ for $\alpha_0 = 0$: if there exists a semi-separating equilibrium, there exists a separating equilibrium.

At an even more abstract level, the intuition is that a non-smart agent who chooses to imitate a smart agent creates a positive externality for other conformist non-smart agents, who are now better able to hide among smart agents without getting punished if they get the consequence wrong. Then, a non-smart agent is most likely to want to follow his own signal when all other non-smart agents are following their signals.

The proposition does not imply that there do not exist semi-separating equilibria. It is possible to find games in which there exist both a separating equilibrium and a semi-separating equilibrium.

Now that we know that the condition for the existence of an informative equilibrium is the same as the condition for the existence of a separating equilibrium, it is not hard to find such condition.

Proposition 6 *There exists an informative equilibrium if and only if*

$$\frac{p}{1-p}\frac{\gamma q_{0g} + (1-\gamma)\, q_{0b}}{\gamma q_{1g} + (1-\gamma)\, q_{1b}} \leq \frac{q_{0b} - q_{0g}}{q_{1g} - q_{1b}} \leq \frac{p}{1-p}\frac{\gamma\,(1-q_{0g}) + (1-\gamma)\,(1-q_{0b})}{\gamma\,(1-q_{1g}) + (1-\gamma)\,(1-q_{1b})}. \tag{17}$$

Proof. The necessary and sufficient conditions for the existence of an equilibrium in which $\alpha_0 = 0$ and $\alpha_1 = 1$ are (12) and (13). We first show that (12) holds if and only if the first inequality in (17) holds.

Note that

$$\begin{aligned}\pi(1,x)-\pi(0,x) &= \left(\frac{\Pr(y=1|g,x)}{\Pr(y=1|x)}-\frac{\Pr(y=0|g,x)}{\Pr(y=0|x)}\right)\Pr(g)\\ &= \frac{\Pr(y=1|g,x)\left(1-\Pr(y=1|x)\right)-\left(1-\Pr(y=1|g,x)\right)\Pr(y=1|x)}{\Pr(y=1|x)\Pr(y=0|x)}\Pr(g)\\ &= \frac{\Pr(y=1|g,x)-\Pr(y=1|x)}{\Pr(y=1|x)\Pr(y=0|x)}\Pr(g)\end{aligned}$$

Then,

$$\begin{aligned}\Pr(x|y=0)\left(\pi(1,x)-\pi(0,x)\right) &= \frac{\Pr(y=0|x)\Pr(x)}{\Pr(y=0)}\frac{\Pr(y=1|g,x)-\Pr(y=1|x)}{\Pr(y=1|x)\Pr(y=0|x)}\Pr(g)\\ &= \frac{\Pr(g)}{\Pr(y=0)}\Pr(x)\left(\frac{\Pr(y=1|g,x)}{\Pr(y=1|x)}-1\right)\end{aligned}$$

Then, (12) holds if and only if

$$\begin{aligned}&\frac{\Pr(g)}{\Pr(y=0)}\left(\Pr(1)\frac{\Pr(y=1|g,1)}{\Pr(y=1|1)}+\Pr(0)\frac{\Pr(y=1|g,0)}{\Pr(y=1|0)}-1\right)\leq 0\\ \Leftrightarrow\quad &\Pr(1)q_{1g}\left(q_{0g}\Pr(g)+q_{0b}\Pr(b)\right)+\Pr(0)q_{0g}\left(q_{1g}\Pr(g)+q_{1b}\Pr(b)\right)\\ &-\left(q_{0g}\Pr(g)+q_{0b}\Pr(b)\right)\left(q_{1g}\Pr(g)+q_{1b}\Pr(b)\right)\leq 0\\ \Leftrightarrow\quad &\Pr(1)q_{1g}\left(q_{0g}\Pr(g)+q_{0b}\Pr(b)\right)+\Pr(0)q_{0g}\left(q_{1g}\Pr(g)+q_{1b}\Pr(b)\right)\\ &-\left(\Pr(0)+\Pr(1)\right)\left(q_{0g}\Pr(g)+q_{0b}\Pr(b)\right)\left(q_{1g}\Pr(g)+q_{1b}\Pr(b)\right)\leq 0\\ \Leftrightarrow\quad &\Pr(1)\left(q_{1g}-\left(q_{1g}\Pr(g)+q_{1b}\Pr(b)\right)\right)\left(q_{0g}\Pr(g)+q_{0b}\Pr(b)\right)\\ &+\Pr(0)\left(q_{0g}-\left(q_{0g}\Pr(g)+q_{0b}\Pr(b)\right)\right)\left(q_{1g}\Pr(g)+q_{1b}\Pr(b)\right)\leq 0\\ \Leftrightarrow\quad &\Pr(1)\left(q_{1g}-q_{1b}\right)\Pr(b)\left(q_{0g}\Pr(g)+q_{0b}\Pr(b)\right)+\Pr(0)\left(q_{0g}-q_{0b}\right)\Pr(b)\left(q_{1g}\Pr(g)+q_{1b}\Pr(b)\right)\leq 0\\ \Leftrightarrow\quad &\Pr(1)\left(q_{1g}-q_{1b}\right)\left(q_{0g}\Pr(g)+q_{0b}\Pr(b)\right)\leq\Pr(0)\left(q_{0b}-q_{0g}\right)\left(q_{1g}\Pr(g)+q_{1b}\Pr(b)\right)\\ \Leftrightarrow\quad &\frac{q_{0b}-q_{0g}}{q_{1g}-q_{1b}}\geq\frac{p}{1-p}\frac{\Pr(b)q_{0b}+\Pr(g)q_{0g}}{\Pr(g)q_{1g}+\Pr(b)q_{1b}}\end{aligned}$$

The proof that (13) is equivalent to the second inequality in the proposition is similar to the argument above and it is omitted. ■

To understand Proposition 6, note that

$$\frac{p}{1-p}\frac{\gamma q_{0g}+(1-\gamma)q_{0b}}{\gamma q_{1g}+(1-\gamma)q_{1b}}<1\qquad\text{and}\qquad\frac{\gamma(1-q_{0g})+(1-\gamma)(1-q_{0b})}{\gamma(1-q_{1g})+(1-\gamma)(1-q_{1b})}>1.$$

We can link condition (17) with the condition for the smart signal found in Proposition 2. Both impose bounds on the term

$$\frac{q_{0b}-q_{0g}}{q_{1g}-q_{1b}},$$

which is the relative informativeness of the two y's. The smartness condition establishes which signal is more informative. The condition in Proposition 6 says whether one signal is *much* more informative than the other.

If, for instance, $y = 1$ is the smart signal, then

$$\frac{q_{0b} - q_{0g}}{q_{1g} - q_{1b}} < \frac{p}{1-p} \tag{18}$$

We can disregard the second inequality in Proposition 6 because it is implied by (18). Instead, the inequality

$$\frac{q_{0b} - q_{0g}}{q_{1g} - q_{1b}} \geq \frac{p}{1-p} \frac{\gamma q_{0g} + (1-\gamma) q_{0b}}{\gamma q_{1g} + (1-\gamma) q_{1b}} \tag{19}$$

can hold or not. If it holds, there is no informative equilibrium because $y = 1$ is "too smart" to allow for separation. If the equilibrium were informative, the agent who observes $y = 0$ would always want to pretend he observed $y = 1$. If instead the inequality (19) does not hold, separation is possible because the agent who observes $y = 0$ prefers to increase his likelihood to get $u = 1$ rather than pretend he has $y = 1$.

If we revisit the example presented earlier, we can now formally verify the result that there is no informative equilibrium. Recall that in that example $\gamma = \frac{1}{2}$, $p = \frac{1}{2}$, $q_{0b} = q_{1b} = \frac{1}{2}$, $q_{0g} = \frac{1}{2}$, and $q_{1g} = 1$. The smartness condition (18) is

$$\frac{0}{\frac{1}{2}} < 1.$$

The smart signal is $y = 1$. There exists an informative equilibrium if and only if (19) is satisfied. That is,

$$0 \geq 1 \frac{\frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{2}}{\frac{1}{2}1 + \frac{1}{2}\frac{1}{2}} = \frac{2}{3},$$

which shows that informative equilibria are impossible.

If instead the smart signal had been less smart, an informative equilibrium would have been possible. For instance, modify the example by assuming that if $x = 0$, the good type receives an informative signal: $q_{0g} = \frac{1}{6}$. The existence condition (19) becomes

$$\frac{\frac{1}{2} - \frac{1}{6}}{1 - \frac{1}{2}} \geq 1 \frac{\frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{6}}{\frac{1}{2}1 + \frac{1}{2}\frac{1}{2}},$$

that is, $\frac{2}{3} \geq \frac{4}{9}$. Indeed, one can show that, holding the other parameters constant, there exists an informative equilibrium if and only if $q_{0g} \leq \frac{1}{4}$.

3.4 When should the action be revealed?

We are now in a position to compare the expected payoff of the principal in the best equilibrium under concealed action with her expected payoff in the best equilibrium with revealed action. As

we saw in Section 2, ex ante social welfare corresponds to the expected payoff of the principal because the expected payoff of the agent is constant.

From Proposition 3, the best equilibrium with concealed action is a separating equilibrium with $a = y$.

What happens with revealed action depends on condition (17). If the condition holds, there exists a separating equilibrium with $a = y$. The agent behavior is thus the same as with concealed action but the principal gets more information. The variance of the agent posterior increases and the principal's payoff, which is convex in the posterior, goes up. Compared to concealed action, the discipline effect is the same but the sorting effect improves. Thus, the principal is better off.[14]

If instead condition (17) fails, there is no informative equilibrium and the best equilibrium is one where the agent chooses the action that corresponds to the most likely state. The discipline effect worsens because the agent disregards useful information. sorting too is affected negatively because in an informative equilibrium the posterior is equal to the prior. Thus, the principal is worse off. We summarize the argument as follows:

Proposition 7 *If (17) holds, revealing the agent's action does not affect discipline and improves sorting. If (17) fails, revealing the agent's action worsens both discipline and sorting. Hence, the principal prefers to reveal the action if and only if (17) holds.*

It may also be interesting to see if there is a tension between what is optimal ex ante and what is optimal ex post. Suppose we are in a separating equilibrium. After the agent has chosen his action, the principal always benefits from observing the action because she can use the additional information for sorting purposes. However, before the agent has chosen his action, the principal may want to commit not to observe the action ex post. In which way is the benefit of ex ante concealment connected to the incentive for ex post observation?

Suppose that the smart signal is $y = 1$. As we saw above, a policy of concealment is optimal if and only if (19) fails. The incentive to commit is simply represented by the variable c which takes value 0 if (19) holds and 1 if it fails.

[14] Proof that screening is better when a is observed:

$$\begin{aligned}
& \sum_{x,y} \Pr(x,y) v\left(\pi\left(y,x\right)\right) \\
= \; & \Pr(y=x) \sum_{x} \Pr(x) v\left(\pi\left(x,x\right)\right) + \Pr(y \neq x) \sum_{x} \Pr(x) v\left(\pi\left(1-x,x\right)\right) \\
\geq \; & \Pr(y=x) \sum_{x} \Pr(x) v\left(\tilde{\pi}\left(1\right)\right) + \Pr(y \neq x) \sum_{x} \Pr(x) v\left(\tilde{\pi}\left(0\right)\right) \\
= \; & \sum_{x,y} \Pr(x,y) v\left(\pi\left(u\left(x,y\right)\right)\right)
\end{aligned}$$

In a separating equilibrium, the benefit of observing a ex post is

$$r = \sum_{x,y} \Pr(x,y) v\left(\pi\left(y,x\right)\right) - \sum_{x,y} \Pr(x,y) v\left(\tilde{\pi}\left(u\left(y,x\right)\right)\right).$$

To make a simple comparative analysis exercise, we fix p and γ. We also hold constant the following expressions: the probability that a good agent gets the signal right:

$$\Pr\left(y = x|g\right) = pq_{1g} + (1-p)(1-q_{0g}) \equiv s;$$

and the probability of the signal given the state, non-conditional on the agent type,

$$\Pr\left(y = 1|x\right) = q_{xg}\gamma + q_{xb}\left(1-\gamma\right) \equiv q_x \qquad \text{for } x = 1, 2.$$

Note that this also implies that $\Pr(y = x)$ and $\Pr(y = x|b)$ are constant. This leaves one degree of freedom on q, which can be represented without loss of generality with movements of the ratio $\frac{q_{0b}-q_{0g}}{q_{1g}-q_{1b}}$. This degree of freedom corresponds to the relative informativeness of the two realizations of y.

We show that there exists a tension between the ex ante incentive to commit not to observe a and the ex post benefit of observation:

Proposition 8 *The incentive to commit ex ante c and the benefit of observing the action ex post r are both nondecreasing in $\frac{q_{0b}-q_{0g}}{q_{1g}-q_{1b}}$.*

Proof. c depends on whether (19) holds. As the right-hand side of (19) is constant, c is nondecreasing in $\frac{q_{0b}-q_{0g}}{q_{1g}-q_{1b}}$.

Let us now consider r. As s is constant, an increase in q_{1g} must be accompanied by an increase in q_{0g}. As q_x is constant, an increase in q_{xg} is associated to a decrease in q_{xb}. Thus, a decrease in $\frac{q_{0b}-q_{0g}}{q_{1g}-q_{1b}}$ corresponds to an increase in q_{1g} and q_{0g} and a decrease in q_{1b} and q_{0b}.

Given that $a = y$, the posteriors are

$$\begin{aligned}
\pi\left(1,0\right) &= \frac{q_{0g}\gamma}{q_{0g}\gamma + q_{0b}\left(1-\gamma\right)}; \\
\pi\left(0,0\right) &= \frac{\left(1-q_{0g}\right)\gamma}{\left(1-q_{0g}\right)\gamma + \left(1-q0\right)\left(1-\gamma\right)}; \\
\pi\left(1,1\right) &= \frac{q_{1g}\gamma}{q_{1g}\gamma + q_{1b}\left(1-\gamma\right)}; \\
\pi\left(0,1\right) &= \frac{\left(1-q_{1g}\right)\gamma}{\left(1-q_{1g}\right)\gamma + \left(1-q_{1b}\right)\left(1-\gamma\right)}.
\end{aligned}$$

A decrease in $\frac{q_{0b}-q_{0g}}{q_{1g}-q_{1b}}$ generates an increase in $\pi\left(1,1\right)$ and $\pi\left(1,0\right)$ and a decrease in $\pi\left(0,1\right)$ and $\pi\left(0,0\right)$. The assumption that $\Pr(y = x|g)$ is constant means that $\tilde{\pi}\left(u\right)$ is constant as well.

The benefit of observing a when the agent plays $a = y$ is

$$\sum_{x,y} \Pr(x,y) v\left(\pi\left(y,x\right)\right) - \sum_{x,y} \Pr(x,y) v\left(\tilde{\pi}\left(u\left(y,x\right)\right)\right).$$

Given that $\tilde{\pi}\left(u\right)$ is constant, we only consider the first part, which we rewrite as

$$V = \Pr(y = x) \sum_{\tilde{x}} \Pr(\tilde{x}) v\left(\pi\left(\tilde{x},\tilde{x}\right)\right) + \Pr(y \neq x) \sum_{\tilde{x}} \Pr(\tilde{x}) v\left(\pi\left(1-\tilde{x},\tilde{x}\right)\right).$$

It is easy to check that $\pi\left(1,1\right) > \pi\left(0,0\right)$ and $\pi\left(1,0\right) > \pi\left(0,1\right)$. As v is convex, a decrease in $\frac{q_{0b}-q_{0g}}{q_{1g}-q_{1b}}$ increases both $\sum_x \Pr(x) v\left(\pi\left(x,x\right)\right)$ and $\sum_x \Pr(x) v\left(\pi\left(1-x,x\right)\right)$. ■

There is a tension between the ex ante benefit of concealed action and the ex post of revealed action. If the agent plays according to his signal, the principal can infer whether the agent received the smart signal or not. The more "smart" the smart signal is, the more useful the information. In a highly asymmetric situation, ex post the principal stands to gain a lot from knowing the agent's action. But the agent realizes this and wants to hides the fact that he receives the non-smart signal. This kills the separating equilibrium and damages the principal in terms both of discipline and sorting. Thus, if the smart signal is "very smart", the principal is better off if she can commit to concealed action.

4 Complementarity between Observing Action and Consequence

We have so far asked whether revealing the agent's action is a good idea, but we have maintained the assumption that consequences are always observed. In some cases, especially in the political arena, the principal may not be able to fully evaluate the consequences of the agent's behavior or may be able to do it with such a time lag that the information is of limited use for sorting purposes. Take for instance a large-scale public project, such as a reform of the health system. Its main provisions are observable right away, but it takes years for its effects to develop. In the medium term, the public knows the characteristics of the project that has been undertaken (the action) but cannot yet judge its success (the consequence).

This section looks at what happens when consequences are not necessarily observed. The focus will be mostly on the complementarity between transparency on action and transparency on consequence. However, first, we examine the simple case in which the consequence goes totally unobserved.

The game is as in the reduced form except that the principal observes either only a or nothing at all. If the action is observed, it is easy to see that in equilibrium the choice of action must be uncorrelated with the agent's signal (otherwise one action would have a higher posterior and all agents would choose it – contradiction). The best equilibrium for the principal is an uninformative

equilibrium in which the agent chooses the most likely action. No sorting occurs. If instead the action is not observed, the best equilibrium is one in which the agent chooses $a = y$. No sorting occurs but the first-period decision is better. Therefore, when u is not observed, the principal's expected payoff is certainly higher when a is concealed.

This observation contrasts with the result obtained in the previous section that, when the consequence is observed, revealing the action may a good idea, which seems to point to a complementarity between observing consequences and revealing actions. As we shall now see, this complementarity is indeed present in a general way.

Let $\rho_u \in [0,1]$ be the probability that u is observed and $\rho_a \in [0,1]$ be the probability that a is observed. At stage 2 there are thus four possible information scenarios according to whether the consequence and/or the action is observed. The previous section considered the cases $(\rho_u = 1, \rho_a = 1)$ and $(\rho_u = 1, \rho_a = 0)$.

To simplify matters, we restrict attention to pooling and separating equilibria. We assume that $y = 1$ is the smart signal and we look at the separating equilibrium in which $a = y$ and the pooling equilibrium in which the agent plays the most likely action. The pooling equilibrium always exists. For every pair (ρ_u, ρ_a), we ask whether the separating equilibrium exists.[15]

Proposition 9 *For every ρ_u there exists $\rho_a^*(\rho_u) \in (0,1]$ such that the game has a separating equilibrium if and only if $\rho_a \leq \rho_a^*$. The threshold ρ_a^* is nondecreasing in ρ_u.*

Proof. Suppose that the agent chooses $a = y$. Let $\pi(a,x)$, $\pi(u(a,x))$, $\pi(a)$, and γ be the posterior evaluated by the principal in the four possible information scenarios. Note that because we hold fixed the agent equilibrium strategy $(a = y)$, these posteriors do not depend on (ρ_u, ρ_a) but only on the information scenario that is realized. Given a and y, the expected posterior for the agent is

$$E(\pi|a,y) = \rho_u\rho_a E_x(\pi(a,x)|y) + \rho_u(1-\rho_a)E_x(\pi(u(a,x)|y) + (1-\rho_u)\rho_a\pi(a) + (1-\rho_u)(1-\rho_a)\gamma.$$

Note that the last two addends do not depend on x, and therefore on y. A necessary and sufficient condition for the existence of a separating equilibrium is $E(\pi|0,0) \geq E(\pi|1,0)$, which rewrites as:

$$\begin{aligned} & (1-\rho_a)\rho_u(E_x(\pi(u(0,x)|y=0) - E_x(\pi(u(1,x)|y=0)) \\ \geq\ & \rho_a(\rho_u(E_x(\pi(1,x)|y=0) - E_x(\pi(0,x)|y=0)) + (1-\rho_u)(\pi(a=1) - \pi(a=0))) \end{aligned}$$

[15] It is not clear whether the analogue of Proposition 4 can be proven for this more complex case. Although examples have not been found, one cannot exclude that there exists a semi-separating equilibrium when a separating equilibrium does not exist.

or

$$(1-\rho_a)\,\rho_u\Delta_1 \geq \rho_a\left(\rho_u\Delta_2 + (1-\rho_u)\,\Delta_3\right). \tag{20}$$

Note that Δ_1, Δ_2, and Δ_3 do not depend on (ρ_u, ρ_a). It is easy to see that $\Delta_1 > 0$ and that $\Delta_3 > 0$. By (1), we see that

$$\begin{aligned} & E_x\left(\pi(1,x)|y=0\right) - E_x\left(\pi(0,x)|y=0\right) \\ < \;& E_y E_x\left(\pi(1,x)|y\right) - E_y E_x\left(\pi(0,x)|y\right) \\ < \;& E_x(\pi(1,x)|y=1) - E_x(\pi(0,x)|y=1). \end{aligned}$$

As $E_y E_x(\pi(a,x)|y) = \pi(a)$, we have that $\Delta_3 > \Delta_2$. We rewrite (20) as

$$\frac{1-\rho_a}{\rho_a} \geq \frac{\rho_u\Delta_2 + (1-\rho_u)\,\Delta_3}{\rho_u\Delta_1}.$$

On the right-hand side, the numerator is decreasing in ρ_u and the denominator is increasing. The left-hand side is decreasing in ρ_a. ■

Proposition 9 has two parts. First, given a probability that the consequence is observed, ρ_u, there exists a threshold $\rho_a^*(\rho_u)$ such that there exists a separating equilibrium if and only if the probability of observing the action is below the threshold. For any level of transparency on consequences, ρ_u, there exists a threshold $\rho_a^*(\rho_u)$ such that there exists a separating equilibrium if and only if the level of transparency on action is within the threshold. Second, the threshold is nondecreasing in ρ_u. More transparency on consequence allows for more transparency on action without creating incentives for conformism. This is because conformism is deterred by the threat of failure. The agent may impress the principal when he chooses the action associated to the smart signal, but he is going to be punished if the action does not match the state of the world. The risk of punishment is directly proportional to the probability that the consequence is observed.

5 When the Agent Knows His Own Type

So far we have assumed that the agent does not know his own type. In this section we remove this restriction and see what happens (in one example) when the agent has some self-knowledge. The agent receives a signal about his own type, which may be more or less informative. The two extreme cases are when the agent knows his type perfectly and when he has no information about it.

The objective of this section is to probe the robustness of the results obtained with no self-knowledge. As we shall see, self-knowledge may reduce the incentive to behave in a conformist

way. Two points will be made. First, if the agent is perfectly informed about his own type, then we show that there is an informative equilibrium. The good agent follows his signal while the bad agent still behaves in a conformist manner. This is because an agent who knows that he has a more precise signal has more of an incentive make decisions according to it. This result may lead one to suspect that conformism is a non-robust feature of the case in which the agent has no self-knowledge at all. However, the second point of this section is that if the signal the agent receives about his own type is informative but weak, there still exists no informative equilibrium. Thus the results appear to be robust to some self-knowledge but not to a lot of it.[16]

We modify the baseline model as follows. Before observing the signal y about the state x, the agent receives a signal $z \in \{0,1\}$ about his own type θ. The signal z has distribution

$$\Pr(z|\theta) = \begin{cases} k & \text{if } z = \theta \\ 1-k & \text{if } z \neq \theta \end{cases}$$

with $k \in \left[\frac{1}{2}, 1\right]$. If $k = \frac{1}{2}$, the signal is uninformative and we return to the baseline model. If $k = 1$, the agent knows his own type. Also, z and x are mutually independent, and z and y are mutually independent conditional on x and θ. The rest of the game is as in the baseline model. A strategy profile is then $(\alpha_{yz})_{y\in\{0,1\},z\in\{b,g\}}$, where α_{yz} is the probability that an agent with signal y and self-knowledge z plays $a = 1$.[17]

In order to make the two main points of this section, it is sufficient to restrict attention to the example discussed in Section 3.1. Thus: $\gamma = \frac{1}{2}$, $p = \frac{1}{2}$, $q_{0b} = q_{1b} = \frac{1}{2}$, $q_{0g} = \frac{1}{2}$, and $q_{1g} = 1$.

We show that if the agent knows his own type ($k = 1$) there is a new type of equilibrium in which an agent with self-knowledge $z = g$ plays in the efficient way ($a = y$), while an agent with $z = b$ randomizes between the two actions:

[16] The question of how well individuals assess their own skills is a difficult one. Chiappori and Salanié [6] propose a test of asymmetric information, which, applied to car insurance contracts, suggests that the insured do not know more about their types than the insurers.

It may also be the case that agents display systematic biases (see Malmendier and Tate [24] for a study on CEO overconfidence, which suggests that chief executives systematically overestimate their ability). It would be interesting – but outside the scope of the present work – to allow for systematic deviations between the agent's and the principal's priors on the agent's type.

[17] The type of the agent can now take four possible values while the message space, which still corresponds to the action space, is binary. One may argue that this is an unnatural restriction and that the message space should be enriched to have the same dimension as the action space. However, trading off generality for comparability and simplicity, we choose to keep the set-up used in the rest of the paper.

Also, if the agent is allowed to send another binary cheap talk signal, it is easy to see that there cannot exist a separating equilibrium in which a reveals y and the additional cheap talk signal reveals z, because the agent would always prefer to state $z = 1$ rather than $z = 0$ ($\pi(a, x, z = 1) > \pi(a, x, z = 0)$ for any a and x, because everything else equal a higher z is good news about θ).

Proposition 10 *If $k = 1$, there exists an informative equilibrium in which an agent with $z = 1$ plays $a = y$ and an agent with $z = 0$ plays according to $\alpha_0 = \alpha_1 = \frac{1}{8}\left(1 + \sqrt{41}\right) \simeq 0.92$.*

Proof. Given $k = 1$, consider strategy $0 = \alpha_{0g} \leq \alpha_{0b} = \alpha = \alpha_{1b} \leq \alpha_{1g} = 1$. The posteriors are:

$$\begin{aligned}
\pi(1,1) &= \frac{1}{1+\alpha}; \\
\pi(0,1) &= 0; \\
\pi(1,0) &= \frac{1}{1+2\alpha}; \\
\pi(0,0) &= \frac{1}{3-2\alpha}.
\end{aligned}$$

If the agent knows his type, the conditional probability of x given y is

$$\begin{aligned}
\Pr(x = 1|y = 1, z = g) &= \frac{2}{3}; \\
\Pr(x = 1|y = 0, z = g) &= 0; \\
\Pr(x = 1|y = 1, z = b) &= \Pr(x = 1|y = 0, b) = \Pr(x = 1|b) = \frac{1}{2}.
\end{aligned}$$

In order for $\alpha \in (0,1)$, the bad agent must be indifferent between $a = 0$ and $a = 1$:

$$\begin{aligned}
&\Pr(x = 1|z = b)\,\pi(1,1) + \Pr(x = 0|z = b)\,\pi(1,0) \\
= \;&\Pr(x = 1|z = b)\,\pi(0,1) + \Pr(x = 0|z = b)\,\pi(0,0),
\end{aligned}$$

yielding

$$\frac{1}{1+\alpha} + \frac{1}{1+2\alpha} = 0 + \frac{1}{3-2\alpha},$$

with solution $\alpha = \frac{1}{8} + \frac{1}{8}\sqrt{41}$.

It is easy to verify that for an agent with g it is a best response to play $a = y$. ■

An agent who knows he is bad also knows that his signal y provides no information about the state x. He evaluates $\Pr(x)$ according to the prior and he chooses the action according to whether $E_x(\pi(a,x))$ is higher for $a = 0$ or $a = 1$. Instead, an agent who knows he is good receives an informative signal about x and he uses $\Pr(x|y)$. That is why a bad agent randomizes while a good agent follows his own signal. An increase in the probability that a bad agent chooses $a = 1$ worsens the posteriors $\pi(1,x)$ and improves the posteriors $\pi(0,x)$. Thus, there exists a probability such that the bad agent is indifferent between the two actions. However, at the same point the good agent strictly prefers to follow his own signal. This determines an informative equilibrium in which the good agent follows his own signal.

In Section 3 we saw that, when the agent does not observe his type, this example has no informative equilibrium. We now show that this is still true when the agent has limited self-knowledge.

Proposition 11 *If $k < \sqrt{7} - 2 \simeq 0.64$, there exists no informative equilibrium.*

Proof. Assume that $k \in (0,1)$. For future reference, the posteriors for a generic strategy profile are:

$$\begin{aligned}
\pi(1,1) &= \frac{2k\alpha_{1g} + 2(1-k)\alpha_{1b}}{(1+k)\alpha_{1g} + (1+k)\alpha_{1b} + (1-k)\alpha_{0g} + k\alpha_{0b}}; \\
\pi(0,1) &= \frac{2 - (2k\alpha_{1g} + 2(1-k)\alpha_{1b})}{4 - ((1+k)\alpha_{1g} + (1+k)\alpha_{1b} + (1-k)\alpha_{0g} + k\alpha_{0b})}; \\
\pi(1,0) &= \frac{k\alpha_{1g} + (1-k)\alpha_{1b} + k\alpha_{0g} + (1-k)\alpha_{0b}}{\alpha_{1g} + \alpha_{1b} + \alpha_{0g} + \alpha_{0b}}; \\
\pi(0,0) &= \frac{2 - (k\alpha_{1g} + (1-k)\alpha_{1b} + k\alpha_{0g} + (1-k)\alpha_{0b})}{4 - (\alpha_{1g} + \alpha_{1b} + \alpha_{0g} + \alpha_{0b})}.
\end{aligned}$$

One of the following must hold: $\alpha_{1b} > \alpha_{0b}$, $\alpha_{1b} = \alpha_{0b}$, or $\alpha_{1b} < \alpha_{0b}$. For every equilibrium in which $\alpha_{1b} < \alpha_{0b}$ there exists a specular equilibrium in which $\alpha_{1b} > \alpha_{0b}$. Thus we restrict attention to the first two cases. The proof proceeds by showing that in each of the two cases there exists no informative equilibrium if k is below a certain threshold.

Suppose that there exists an equilibrium in which $\alpha_{1b} > \alpha_{0b}$. It must be that

$$\Pr(x=1|y=0,z=b)(\pi(1,1) - \pi(0,1)) \leq \Pr(x=0|y=0,z=b)(\pi(0,0) - \pi(1,0)) \quad (21)$$

$$\Pr(x=1|y=1,z=b)(\pi(1,1) - \pi(0,1)) \geq \Pr(x=0|y=1,z=b)(\pi(0,0) - \pi(1,0)) \quad (22)$$

This implies that:

$$\pi(1,1) \geq \pi(0,1) \quad (23)$$

$$\pi(0,0) \geq \pi(1,0) \quad (24)$$

To see this: if (23) holds but (24) does not, (21) must be false; if (24) holds but (23) does not, (22) must be false; and if neither (23) nor (24) hold, the fact that (21) holds implies that (22) must be false. Given (21) and (22),

$$\Pr(x=1|y=0,z=g)(\pi(1,1) - \pi(0,1)) < \Pr(x=0|y=0,z=g)(\pi(0,0) - \pi(1,0)) \quad (25)$$

$$\Pr(x=1|y=1,z=g)(\pi(1,1) - \pi(0,1)) > \Pr(x=0|y=1,z=g)(\pi(0,0) - \pi(1,0)) \quad (26)$$

Thus $0 = \alpha_{0g} \leq \alpha_{0b} < \alpha_{1b} \leq \alpha_{1g} = 1$. For (25) to hold, it must be $\pi(1,0) < \pi(0,0)$, implying

$$\begin{aligned}
2k\alpha_{1g} + 2(1-k)\alpha_{1b} + 2k\alpha_{0g} + 2(1-k)\alpha_{0b} &< \alpha_{1g} + \alpha_{1b} + \alpha_{0g} + \alpha_{0b}; \\
(2k-1)\alpha_{1g} + (2k-1)\alpha_{0g} &< (2k-1)\alpha_{1b} + (2k-1)\alpha_{0b}; \\
\alpha_{1g} + \alpha_{0g} &< \alpha_{1b} + \alpha_{0b}.
\end{aligned}$$

This means that $\alpha_{1b} + \alpha_{0b} > 1$. By an argument similar to Proposition 4, it cannot be that $0 < \alpha_{0b} < \alpha_{1b} < 1$. Hence, $\alpha_{1b} = 1$. We then know that $0 = \alpha_{0g} \leq \alpha_{0b} < \alpha_{1b} = \alpha_{1g} = 1$, and the posteriors are

$$\begin{aligned}
\pi(1,1) &= \frac{2k + 2(1-k)}{2k + 2(1-k) + 1 + k\alpha_{0b}}; \\
\pi(0,1) &= 0; \\
\pi(1,0) &= \frac{k + (1-k)(1+\alpha_{0b})}{k + (1-k)(1+\alpha_{0b}) + (1-k) + k(1+\alpha_{0b})}; \\
\pi(0,0) &= \frac{2 - (k + (1-k)(1+\alpha_{0b}))}{4 - (k + (1-k)(1+\alpha_{0b}) + (1-k) + k(1+\alpha_{0b}))}.
\end{aligned}$$

We see that

$$\begin{aligned}
\Pr(x = 1 | y = 0, z = b) &= \frac{\Pr(x = 1, y = 0, z = b | \theta = g) + \Pr(x = 1, y = 0, z = b | \theta = b)}{\Pr(y = 0, z = b | \theta = g) + \Pr(y = 0, z = b | \theta = b)} \\
&= \frac{\frac{1}{2}k}{\frac{1}{2}k + \frac{1}{2}} = \frac{k}{1+k}
\end{aligned}$$

Then a necessary condition for $\alpha_{0b} < 1$ is

$$k(\pi(1,1) - \pi(0,1)) + (\pi(1,1) - \pi(0,1)) \leq 0,$$

which rewrites as

$$-4k + (6k-3)\alpha_0 + 2k^2\alpha_0^2 \geq 0,$$

which, given that $k \geq \frac{1}{2}$, cannot hold if

$$-3 + 6k + 2k^2 - 4k < 0,$$

with solution

$$k < \frac{1}{2}\left(\sqrt{7} - 1\right) \cong 0.82.$$

Thus, if $k < \frac{1}{2}\left(\sqrt{7} - 1\right)$, there is no equilibrium in which $\alpha_{1b} > \alpha_{0b}$.

Next consider an equilibrium in which $\alpha_{1b} = \alpha_{0b}$. By an argument similar to the one use above, it is easy to see that $0 \leq \alpha_{0g} \leq \alpha_{0b} = \alpha_{1b} \leq \alpha_{1g} \leq 1$. It cannot be that $0 < \alpha_{0b} = \alpha_{1b} < 1$. Either $0 = \alpha_{0g} = \alpha_{0b} = \alpha_{1b} \leq \alpha_{1g} = 1$ or $0 \leq \alpha_{0g} = \alpha_{0b} = \alpha_{1b} = \alpha_{1g} = 1$. It is easy to exclude the first case (because $a = 1$, which is smart and optimal, is too appealing to an agent with $y = 1$ and $z = b$).

We are then left with the possibility that $0 \leq \alpha_{0g} \leq \alpha_{0b} = \alpha_{1b} = \alpha_{1g} = 1$. The posteriors are:

$$\begin{aligned}
\pi(1,1) &= \frac{2k + 2(1-k)}{2k + 2(1-k) + k + (1-k) + (1-k)\alpha_{0g} + k}; \\
\pi(0,1) &= 0; \\
\pi(1,0) &= \frac{k + (1-k) + k\alpha_{0g} + (1-k)}{3 + \alpha_{0g}}; \\
\pi(0,0) &= k.
\end{aligned}$$

Note that

$$\begin{aligned}\Pr(x=1|y=0,z=g) &= \frac{\Pr(x=1,y=0,z=g|\theta=g)+\Pr(x=1,y=0,z=g|\theta=b)}{\Pr(y=0,z=g|\theta=g)+\Pr(y=0,z=g|\theta=b)} \\ &= \frac{\frac{1}{2}(1-k)}{\frac{1}{2}(1-k)+\frac{1}{2}} = \frac{1-k}{2-k}.\end{aligned}$$

A necessary condition for $\alpha_{0g} < 1$ is

$$(1-k)(\pi(1,1)-\pi(0,1))+(\pi(1,0)-\pi(0,0)) \geq 0,$$

which rewrites as

$$3-4k-k^2+\left(1-2k+k^2\right)\alpha_0 \leq 0,$$

which cannot hold if

$$3-4k-k^2 > 0,$$

with solution $k < \sqrt{7}-2 \simeq 0.64..$

As $\frac{1}{2}\left(\sqrt{7}-1\right) > \sqrt{7}-2$, the proof is complete. ■

If the agent has self-knowledge, we know that an agent who thinks he is above average has a stronger incentive to follow his signal. Still, if self-knowledge is weak, this is not enough to overcome the agent's desire to make the principal believe that he received the smart signal. The proof of this fact – even just for the example – is quite convoluted because it involves excluding all the possible cases of informative equilibria.

6 General Case

The baseline model had strong restrictions on the action space, signal space, consequence space, and state space, which were all assumed to be binary. This section considers a general setup. We shall see that some of the results proven earlier can still be proven in this more general version. In particular, we can still say a lot about the existence or non-existence of separating equilibria (what we cannot study, because of sheer complexity, is the existence of other informative equilibria). In particular we are able to give support to the two main messages of this paper. First, if one of the possible realizations of the agent signal is "too smart", then there exists no separating equilibrium. Second, transparency on action generates conformism, and it does especially if there is little transparency on consequence. In other words, we prove results that correspond, at least partially, to Propositions 6 and 9.

Let $u \in U$, $a \in A$, $x \in X$, $y \in Y$ where U, A, X, and Y are finite sets. Additionally, U is real-valued. For analytical tractability, we still assume that $\theta \in \{b,g\}$. For every x, let

$p(x) = \Pr(x)$. For every x, y, and θ, let $q_\theta(y|x) = \Pr(y|x,\theta)$. The distributions p and q are assumed to have full support. The consequence is given by $u = \omega(a,x)$. Also define

$$q(y|x) = q_g(y|x)\gamma + q_b(y|x)(1-\gamma).$$

We make two additional assumptions

A1 *(Decision Value of Signal)* There exists $a^* : Y \to A$ such that, for all a, y, u,

$$\sum_{x:\omega(a^*(y),x)\geq u} q(y|x)\,p(x) \geq \sum_{x:\omega(a,x)\geq u} q(y|x)\,p(x)$$

Assumption A1 says that there exists a decision function a^* which is optimal in a strong sense: for every u, the probability of obtaining at least u is higher if the agent use a^* than if he uses any other decision function. The assumption implies that a^* is optimal in the usual sense: for every y,

$$a^*(y) \in \arg\max_a \sum_x \omega(a,x)\Pr(x|y).$$

A2 *(Sorting Value of Signal)* The decision function a^* also satisfies, for every $u' \geq u$ and y,

$$\frac{\sum_{x:\omega(a^*(y),x)=u'} q_g(y|x)\,p(x)}{\sum_{x:\omega(a^*(y),x)=u'} q_b(y|x)\,p(x)} \geq \frac{\sum_{x:\omega(a^*(y),x)=u} q_g(y|x)\,p(x)}{\sum_{x:\omega(a^*(y),x)=u} q_b(y|x)\,p(x)}$$

This second assumption says that, in an equilibrium in which the agent uses a^*, the ratio between the probability that a good agent plays a certain a and obtains a certain u and a bad agent plays the same a and obtains the same u is increasing in u. It is akin to the monotone likelihood ratio condition in moral hazard. It guarantees that a higher type of agent who implements the optimal decision rule a^* is more likely to produce a good consequence.

We now come to what the principal observes. Let $\rho_u \in [0,1]$ be the probability that u is observed and $\rho_a \in [0,1]$ be the probability that a is observed. The two events are independent (so, for instance, the probability that the principal sees a but not u is $(1-\rho_u)\rho_a$).

The following result parallels Proposition 9. A separating equilibrium is more likely to exist when there is more transparency on consequence and less transparency on action:

Proposition 12 *Take $\rho'_u \leq \rho_u$ and $\rho'_a \geq \rho_a$. Under A1 and A2, there exists an equilibrium in which the agent uses a^* under (ρ'_u, ρ'_a) only if there exists an equilibrium in which the agent uses a^* under (ρ_u, ρ_a).*[18]

[18] Given that this section is about generality, one may ask how much Proposition 12 depends on the assumption that the agent's utility is linear in the posterior. Suppose instead that the agent's payoff is given by some nonde-

Proof. For future reference, note that

$$\Pr(u|a,y) = \frac{\sum_{x:\omega(a,x)=u} q(y|x)p(x)}{\sum_x q(y|x)p(x)}.$$

Suppose the agent uses a^*. The posteriors in the four possible information scenarios are:

$$\begin{aligned}
\pi(a,u) &= \begin{cases} \frac{\sum_{y:a^*(y)=a}\sum_{x:\omega(a^*(y),x)=u} q_g(y|x)p(x)\gamma}{\sum_{y:a^*(y)=a}\sum_{x:\omega(a^*(y),x)=u} q(y|x)p(x)} & \text{if } a \text{ is played in equilibrium} \\ \text{unrestricted} & \text{otherwise} \end{cases} \\
\pi(u) &= \frac{\sum_y \sum_{x:\omega(a^*(y),x)=u} q_g(y|x)\, p(x)\, \gamma}{\sum_y \sum_{x:\omega(a^*(y),x)=u} q(y|x)\, p(x)} \\
\pi(a) &= \begin{cases} \frac{\sum_{y:a^*(y)=a}\sum_x q_g(y|x)p(x)\gamma}{\sum_{y:a^*(y)=a}\sum_x q(y|x)p(x)} & \text{if } a \text{ is played in equilibrium} \\ \text{unrestricted} & \text{otherwise} \end{cases} \\
\pi(\emptyset) &= \gamma
\end{aligned}$$

Given (ρ_u, ρ_a), there exists an equilibrium in which the agent uses a^* if and only if, for every y,

$$a^*(y) \in \arg\max_a \sum_u \Pr(u|a,y)\left(\rho_u\rho_a\pi(a,u) + \rho_u(1-\rho_a)\pi(u)\right) + (1-\rho_u)\rho_a\pi(a) + (1-\rho_u)(1-\rho_a)\pi(\emptyset),$$

Claim 1: For every a and y,

$$\sum_u \left(\Pr(u|a^*(y),y) - \Pr(u|a,y)\right)\pi(u) \geq 0.$$

Proof of Claim 1: Note that

$$\Pr(U \geq u|a,y) = \frac{\sum_{x:\omega(a,x)\geq u} q(y|x)p(x)}{\sum_x q(y|x)p(x)}$$

By A1,

$$\frac{\Pr(U \geq u|a^*(y),y)}{\Pr(U \geq u|a,y)} = \frac{\sum_{x:\omega(a^*(y),x)\geq u} q(y|x)p(x)}{\sum_{x:\omega(a,x)\geq u} q(y|x)p(x)} \geq 1.$$

Thus, $\Pr(\cdot|a^*(y),y)$ first–order stochastically dominates $\Pr(\cdot|a,y)$.

creasing function of the posterior: $w(\pi)$. It turns out that the part of the proposition about action revelation is still true. One can prove (see the appendix) that, for any ρ_u and for any $\rho'_a > \rho_a$, there exists an equilibrium in which the agent uses a^* under (ρ_u, ρ'_a) only if there exists an equilibrium in which the agent uses a^* under (ρ_u, ρ_a). Instead the strategy used to prove the other part of the proposition – the one about consequence revelation – does not extend beyond a linear w. It is not clear whether this failure can be fixed by using another line of proof or it is actually due to a failure of this second part for nonlinear utility functions.

By A2, for every y,

$$\frac{\sum_{x:\omega(a^*(y),x)=u} q_g\left(y|x\right)p\left(x\right)}{\sum_{x:\omega(a^*(y),x)=u} q_b\left(y|x\right)p\left(x\right)}$$

is nondecreasing in u. Hence, it easy to see that

$$\frac{\sum_{x:\omega(a^*(y),x)=u} q_g\left(y|x\right)p\left(x\right)}{\sum_{x:\omega(a^*(y),x)=u} q\left(y|x\right)p\left(x\right)}$$

is also nondecreasing in u. This means that

$$\pi\left(u\right)=\frac{\sum_y\sum_{x:\omega(a^*(y),x)=u} q_g\left(y|x\right)p\left(x\right)\gamma}{\sum_y\sum_{x:\omega(a^*(y),x)=u} q\left(y|x\right)p\left(x\right)}$$

is also nondecreasing in u.

The proof of the claim is completed by a standard argument combining first–order stochastic dominance and monotonicity.

Claim 2: Suppose $\rho_a'>\rho_a$. There exists an equilibrium in which the agent uses a^* under (ρ_u,ρ_a') only if there exists an equilibrium in which the agent uses a^* under (ρ_u,ρ_a).

Proof of Claim 2: Suppose not. Then, there exists a y and an $a\notin a^*\left(y\right)$ such that

$$\begin{aligned}&\sum_u \Pr\left(u|a^*\left(y\right),y\right)\left(\rho_u\rho_a\pi\left(a^*\left(y\right),u\right)+\rho_u\left(1-\rho_a\right)\pi\left(u\right)\right)+\left(1-\rho_u\right)\rho_a\pi\left(a^*\left(y\right)\right) \quad (27)\\ <\ &\sum_u \Pr\left(u|a,y\right)\left(\rho_u\rho_a\pi\left(a,u\right)+\rho_u\left(1-\rho_a\right)\pi\left(u\right)\right)+\left(1-\rho_u\right)\rho_a\pi\left(a\right);\end{aligned}$$

and

$$\begin{aligned}&\sum_u \Pr\left(u|a^*\left(y\right),y\right)\left(\rho_u\rho_a'\pi\left(a^*\left(y\right),u\right)+\rho_u\left(1-\rho_a'\right)\pi\left(u\right)\right)+\left(1-\rho_u\right)\rho_a'\pi\left(a^*\left(y\right)\right) \quad (28)\\ \geq\ &\sum_u \Pr\left(u|a,y\right)\left(\rho_u\rho_a'\pi\left(a,u\right)+\rho_u\left(1-\rho_a'\right)\pi\left(u\right)\right)+\left(1-\rho_u\right)\rho_a'\pi\left(a\right).\end{aligned}$$

Subtracting (27) from (28),

$$\begin{aligned}&\rho_u\sum_u\left(\Pr\left(u|a^*\left(y\right),y\right)-\Pr\left(u|a,y\right)\right)\pi\left(u\right) \quad (29)\\ &-\left(\rho_u\left(\sum_u\left(\Pr\left(u|a^*\left(y\right),y\right)\pi\left(a^*\left(y\right),u\right)-\Pr\left(u|a,y\right)\pi\left(a,u\right)\right)\right)+\left(1-\rho_u\right)\left(\pi\left(a^*\left(y\right)\right)-\pi\left(a\right)\right)\right)\\ <\ &0\end{aligned}$$

However, by Claim 1, in order for (27) to hold, it must be that

$$\rho_u\sum_u\left(\Pr\left(u|a^*\left(y\right),y\right)\pi\left(a^*\left(y\right),u\right)-\Pr\left(u|a,y\right)\pi\left(a,u\right)\right)+\left(1-\rho_u\right)\left(\pi\left(a^*\left(y\right)\right)-\pi\left(a\right)\right)<0$$

But the right-hand side of (29) is positive – contradiction.

Claim 3: Suppose $\rho'_u < \rho_u$. There exists an equilibrium in which the agent uses a^* under (ρ'_u, ρ_a) only if there exists an equilibrium in which the agent uses a^* under (ρ_u, ρ_a).

Proof of Claim 3: Suppose not. Then, there exists a y and an $a \notin a^*(y)$ such that (27) holds and

$$\begin{aligned} & \sum_u \Pr(u|a^*(y), y) \left(\rho'_u \rho_a \pi(a^*(y), u) + \rho'_u (1-\rho_a)\pi(u)\right) + \left(1-\rho'_u\right)\rho_a \pi(a^*(y)) \quad (30) \\ \geq & \sum_u \Pr(u|a, y) \left(\rho'_u \rho_a \pi(a, u) + \rho'_u (1-\rho_a)\pi(u)\right) + \left(1-\rho'_u\right)\rho_a \pi(a). \end{aligned}$$

implying

$$\begin{aligned} & \sum_u \Pr(u|a^*(y), y) \left(\rho_a \pi(a^*(y), u) + (1-\rho_a)\pi(u)\right) - \Pr(u|a, y) \left(\rho_a \pi(a, u) + (1-\rho_a)\pi(u)\right) \quad (31) \\ < & \rho_a (\pi(a^*(y)) - \pi(a)) \end{aligned}$$

If $\pi(a^*(y)) \geq \pi(a)$, (30) implies that

$$\sum_u \Pr(u|a^*(y), y) \left(\rho_a \pi(a^*(y), u) + (1-\rho_a)\pi(u)\right) - \Pr(u|a, y) \left(\rho_a \pi(a, u) + (1-\rho_a)\pi(u)\right) > 0,$$

and there is a contradiction in (31) because the left-hand side is positive and the right-hand side is negative.

If $\pi(a^*(y)) < \pi(a)$, we find a contradiction as follows. Note that

$$\sum_u \Pr(u|a, y)\pi(a, u) = \begin{cases} \pi(a) & \text{if } a \text{ is played in equilibrium (ie } \exists y, a \in a^*(y)\text{)} \\ \text{unrestricted} & \text{otherwise} \end{cases}$$

Hence, if a is played in equilibrium,

$$\sum_u \left(\Pr(u|a^*(y), y)\pi(a^*(y), u) - \Pr(u|a, y)\pi(a, u)\right) = \pi(a^*(y)) - \pi(a) > 0,$$

If a is played in equilibrium, then we can always set $\pi(a, u) = 0$ for all u's and

$$\sum_u \left(\Pr(u|a^*(y), y)\pi(a^*(y), u) - \Pr(u|a, y)\pi(a, u)\right) \geq \pi(a^*(y)) - \pi(a) > 0.$$

In both cases, the combination of

$$\sum_u \left(\Pr(u|a^*(y), y)\pi(a^*(y), u) - \Pr(u|a, y)\pi(a, u)\right) > 0$$

and Claim 1 shows that the left-hand side of (27) is positive, which generates a contradiction.

Claims 2 and 3 together prove the proposition. ■

By Proposition 12, if there exists an equilibrium in which the agent uses the optimal decision rule a^* with given ρ_u and ρ_a, the same equilibrium exists also if we take a higher ρ_u and a lower ρ_a.

The intuition behind the result is straightforward. Assumption A1 guarantees that the decision rule a^* is optimal. Assumption A2 implies that, if the agent plays a^*, a good agent is more likely to get a high consequence than a bad agent. If only the consequence is observed (the extreme case in which $\rho_u = 1$ and $\rho_a = 0$), the agent's incentives are aligned with the principal's objective function. A better consequence is good news for the agent's type. Observing the action can only upset this incentive alignment. The contribution of the proposition lies in showing that this comparative statics result also holds for any ρ_u and ρ_a are interior. An increase in ρ_u and a decrease in ρ_a can only make the agent less willing to follow the optimal decision rule a^*.

If we are willing to make a mild assumption on the optimal decision rule, we get a full characterization of the existence region:

Corollary 13 *Suppose that A1 and A2 hold and that, for some y' and y'', $a^*(y') \neq a^*(y'')$. Then, for every ρ_u there exists a $\rho_a^*(\rho_u)$ such that an equilibrium in which the agent uses a^* exists if and only if $\rho_a \leq \rho_a^*(\rho_u)$. Moreover, $\rho_a^*(1) > 0$, $\rho_a^*(0) = 0$, and ρ_a^* is nondecreasing in ρ_u.*

Proof. Given that for some y' and y'', $a^*(y') \neq a^*(y'')$, there cannot exist an equilibrium in which the agent plays a^* when $\rho_u = 0$ and $\rho_a > 0$. However, there exists one when $\rho_u = \rho_a = 0$. Thus, $\rho_a^*(0) = 0$. When $\rho_u = 1$ and $\rho_a = 0$, it is easy to check that A1 and A2 imply existence. Hence, $\rho_a^*(1) > 0$. The rest of the corollary is immediate given Proposition 12. ■

There are two possible cases. First, for every ρ_u and ρ_a there exists an equilibrium in which the agent plays a^*. Second, there are two regions: one containing $(\rho_u = 1, \rho_u = 0)$ in which the agent plays a^*, the other containing $(\rho_u = 0, \rho_u = 1)$ in which there is no equilibrium in which the agent plays a^*. The two regions are divided by $\rho_a^*(\rho_u)$ which is nondecreasing in ρ_u.

Lastly, we look at smartness. The notion of smart realization becomes more complicated in the general setup but it can still be applied. In the simpler model used in the rest of the paper, we saw that if one realization of the agent signal y is exceedingly smart, in the sense that it is a very good signal for the agent's type, then there may not be a separating equilibrium because an agent would always pretend to have observed the smart realization. As we shall see, this line of reasoning is still valid in the general case.

We keep A1 and A2, and, to simplify things, we assume that the optimal decision rule a^* is such that, for every y, $a^*(y')$ is a singleton and $a^*(y') \neq a^*(y'')$ whenever $y' \neq y''$.

Given $y'', y' \in Y$, we say that realization y'' is *uniformly more smart* than realization y' if, for all $x, \tilde{x} \in X$,

$$\frac{q_g(y''|x)}{q_b(y''|x)} > \frac{q_g(y'|\tilde{x})}{q_b(y'|\tilde{x})}. \tag{32}$$

Uniform smartness is a strong condition because it imposes an inequality on likelihood ratios even when the likelihood ratios refer to different states of the world. We can then show a partial analogous to Proposition 6:

Proposition 14 *Suppose $\rho_a = \rho_u = 1$. If there exists $y'', y' \in Y$ such that y'' is uniformly more smart than y', then there exists no equilibrium in which the agent plays a^*.*

Proof. From the proof of Proposition 12, we have that

$$\pi(a^*(y), u) = \frac{\sum_{x:\omega(a^*(y),x)=u} q_g(y|x)\, p(x)\, \gamma}{\sum_{x:\omega(a^*(y),x)=u} q(y|x)\, p(x)}$$

If y'' is smart with respect to y', the definition of smartness (32) implies that for any $u, \tilde{u} \in U$

$$\frac{\sum_{x:\omega(a^*(y''),x)=u} q_g(y''|x)\, p(x)}{\sum_{x:\omega(a^*(y''),x)=u} q(y''|x)\, p(x)} > \frac{\sum_{x:\omega(a^*(y'),x)=\tilde{u}} q_g(y'|x)\, p(x)}{\sum_{x:\omega(a^*(y'),x)=\tilde{u}} q(y'|x)\, p(x)}.$$

Therefore, $\pi(y'', u) > \pi(y', \tilde{u})$ for any u and $\tilde{u}$. But this, in turn, means that

$$\sum_u \Pr\left(u|a^*\left(y''\right), y'\right) \pi\left(a^*\left(y''\right), u\right) > \sum_u \Pr\left(u|a^*\left(y'\right), y'\right) \pi\left(a^*\left(y'\right), u\right),$$

which shows that there exists no equilibrium in which the agent plays a^*. ■

This result is easily understood. We are looking for an equilibrium in the revealed action scenario in which the agent's action fully reveals the agent's signal. The principal knows the agent signal y and the consequence u. But the uniform smartness condition implies that observing y'' and any u is better news about the agent type than observing y' and any $\tilde{u}$. The agent then prefers $a^*(y'')$ over $a^*(y')$ no matter what distributions $a^*(y'')$ and $a^*(y')$ induce over the consequence u.

Proposition 14 differs from Proposition 6, which was proven for the baseline model, in two respects. First, in the baseline model we provided a condition for the existence of any informative equilibrium, while here we can only say something about separating equilibria. This prevents us from using Proposition 14 to draw welfare conclusions. Second, Proposition 6 provided a necessary and sufficient condition, while uniform smartness is just a sufficient condition.

To elaborate on this point, we provide an example in which uniform smartness fails. Suppose that $X = A = Y = \{1, 2, 3\}$ and $U = \{0, 1\}$. Also, $p(x) = \frac{1}{3}$ for all x and $\gamma = \frac{1}{2}$. The signal

distribution is

$q_b(y\|x)$	$x=1$	$x=2$	$x=3$
$y=1$	.5	.5	.5
$y=2$	.2	.2	.2
$y=3$	.3	.3	.3

$q_g(y\|x)$	$x=1$	$x=2$	$x=3$
$y=1$	.5	.0	.0
$y=2$	.3	.8	.3
$y=3$	.2	.2	.8

which gives likelihood ratios:

$\frac{q_g(y\|x)}{q_b(y\|x)}$	$x=1$	$x=2$	$x=3$
$y=1$	1	0	0
$y=2$	$\frac{3}{2}$	4	$\frac{3}{2}$
$y=3$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{8}{3}$

Hence $y=2$ is uniformly more smart than $y=1$, but $y=2$ and $y=3$ cannot be ranked in terms of uniform smartness.

The consequence function is assumed to be:

$\omega(a,x)$	$x=1$	$x=2$	$x=3$
$a=1$	1	0	0
$a=2$	0	1	0
$a=3$	0	0	1

By combining $q_b(y|x)$ and $q_g(y|x)$, we get:

$q(y\|x)$	$x=1$	$x=2$	$x=3$
$y=1$	.50	.25	.25
$y=2$	.25	.50	.25
$y=3$	.25	.25	.50

By combining $q(y|x)$ and $\omega(a,x)$, the optimal decision function is

$$a^*(y) = \begin{cases} 1 & \text{if } y=1 \\ 2 & \text{if } y=2 \\ 3 & \text{if } y=3 \end{cases}$$

Given a^*, one can easily check that A1 and A2 are satisfied.

If the agent plays according to a^*, the posteriors are

$\pi(a,u)$	$u=0$	$u=1$
$a=1$	0	$\frac{1}{2}$
$a=2$	$\frac{3}{5}$	$\frac{4}{5}$
$a=3$	$\frac{2}{5}$	$\frac{8}{11}$

For every u and $\tilde{u}$, $\pi(2,u)$ dominates $\pi(1,\tilde{u})$. The agent prefers to play $y=2$ rather than $y=1$. Also not that $E(\pi(2,u)|y=3) > E(\pi(2,u)|y=2)$. So in this example, even if uniform smartness fails, we are able to show that with revealed action there is no separating equilibrium.

7 Conclusion

This paper has identified a set of circumstances under which committing to concealing a certain kind of information can make the principal better off. First, the principal and the agent must be unable to sign long-term contracts. Second, the agent should be an expert, in the sense that his career depends on how able he is perceived to understand the state of the world. Third, the information about the agent's behavior should be separable into a part that is directly utility-relevant for the principal and a part that is not.[19] If these conditions are met, then revealing the non-directly utility-relevant signal may make the agent behave in a more conformist way, which worsens both discipline and sorting.

Are the theoretical results obtained in this paper useful for understanding existing institutional arrangements? Let us re-consider one by one the examples of non-transparent institutions that we listed in the introduction.

In delegated portfolio management, there have been proposals to increase the frequency with which mutual funds are required to disclose their portfolio composition, which in the US is now six months. The Investment Company Institute (the fund managers' association) [39] rejects proposals for increasing the frequency of portfolio disclosure arguing that an increased frequency risks hurting investors because "[it] would focus undue attention on individual portfolio securities and could encourage a short-term investment perspective." The Institute also argues that there does not seem to be much demand by investors for more information on portfolio holdings. It is easy to use the framework developed here to back up the Institute's argument. The action of a fund manager is his investment strategy. The consequence is return to investors. Returns are observable but also volatile. In the long term they are a reliable signal of the fund manager's quality but in the short term they contain a lot of variance. If the action is observable in the short term, there is a risk that fund managers will behave in a conformist way, ignoring their private investment and following the strategy that is a priori a better signal of their competence.

There is an interesting link between this paper and Lakonishok et al. [20]. They compare returns for the equity-invested portion of mutual funds and pension funds in United States. Their evidence suggests that pension funds underperform mutual funds. This is a surprising finding because pension funds are typically monitored by professional investors with large stakes (the

[19] In view of the results on self-knowledge, we might add the requirement that the agent is not much better than the principal in judging his own ability ex ante.

treasury division of the company that sponsors the pension plan), while mutual funds are held by a very large number of individuals who presumably exert less monitoring effort. One of the hypotheses that Lakonishok et al. advance is that the ability of pension fund investors to monitor their funds closely actually creates an agency problem. The present paper makes this possibility more precise. Mutual fund investors typically choose funds only based on yearly returns, while pension fund investors selects funds only after they communicate directly with fund managers who explain their investment strategy. This may create an incentive for conformism in pension fund managers, which decreases their expected return.

Moving on to corporate governance, shareholders receive information about the management of their firm from the accounting reports that the firm makes. Clearly, accounting involves a great deal of aggregation both across time and across areas. Accounting research has been active on the issue of the optimal degree of disaggregation. One point that is particularly debated, both among researchers and policy-makers, is whether a firm should provide disaggregated data about its productive segments (*segment disclosure*) on a quarterly basis or just on a yearly basis (Leuz and Verrecchia [22]). Currently, in the US there is no legal requirement for quarterly segment disclosure: some firms follow a disclosure policy and others do not. Evidence on whether segment disclosure improves firm performance is inconclusive (Botosan and Harris [3]). Without quarterly segment disclosure, shareholders still have information about short term consequences (from quarterly aggregated reports). What they have difficulty with is inferring the strategy that the firm is following, especially with regard to resource allocation across productive areas. Segment disclosure can then be seen as an improvement in transparency over action. Thus, the present theory provides an additional angle to evaluate the optimality of segment disclosure.[20]

In politics, the idea that more information about non-directly utility-relevant information may induce the agent to behave in a suboptimal way because of career concerns has been articulated in several contexts. In its famous 1974 ruling related to the Watergate case (US *vs.* Nixon), the US Supreme Court uses the following argument to defend the principle behind executive privilege: "Human experience teaches us that those who expect public dissemination of their remarks may well temper candor with a concern for appearances and for their own interest to the detriment of the decision-making process." Britain's Open Government code of practice uses a similar rationale when it provides that "internal discussion and advice can only be withheld where disclosure of the information *in question* would be harmful to the frankness and candour of future discussions." (Campaign for Freedom Information [5, p. 3]).

More precise implications can be extracted from Proposition 9. The optimal degree of action revelation is increasing in the degree of consequence revelation. We should expect transparency

[20] Most existing work in accounting theory predicts that firms should adopt transparency policies, but see Nagar [27] for a reason why risk averse managers may want to limit disclosure.

on decisions to go hand in hand with transparency on consequences. In particular, an action, or the intention to take an action, should not be revealed *before* the consequences of the action are observed. Indeed, Frankel [16] reports that all the 30-plus countries that have adopted an open government code allow for some form of short-term secrecy while the decision process is still ongoing. For instance, Sweden, the country with the oldest and, perhaps the most forceful, freedom of information act, does not recognize the right for citizens to obtain information about a public decision until that decision is implemented. Working papers and internal recommendations that lead to a decision are released only when voters begin to have a chance to form an opinion on the consequence of the decision in question.[21]

The result on complementarity has also another implication. If, for exogenous reasons, citizens are less likely to observe the consequence, optimal institutional design dictates less transparency with regards to action. This may help explain why EU-level bodies are less transparent than the corresponding institutions at the national level. The meetings of the highest legislative body of each EU country are usually public, while, as we saw earlier, the Council of the European Union meets behind closed doors. There is no doubt that Europeans find it easier to evaluate the consequences of policy in areas that are typically under national jurisdiction (health, pensions, education, transports, etc.) rather than areas mainly under EU control (harmonization policy, competition policy, agricultural subsidies, etc.). According to our results, the exogenous differential of information on payoff-relevant observables (laws) is optimally associated to a differential of information on non-payoff-relevant observables (positions during meetings). A Council in which debates were public would risk to give its members so strong an incentive to conform to citizens' expectations that its meetings would lose their information aggregating function.[22]

It is important to stress that this paper also identifies circumstances in which information revelation *is* the optimal policy. In particular, supplying information on the consequence of the agent's action is unambiguously good for the principal. This is true in a direct sense, because it improves discipline and sorting, and – as the section on complementarities showed – in an indirect sense, because it allows for more information on action, which in turn improves sorting. Most of the recent corporate scandals involved distorted profit reporting. As profit is a consequence, nothing in the present paper lends support to the accounting policy choices, such as the expensing

[21] A historical example of this transparency policy is the US Constitutional Convention. George Mason refers to the secrecy of the Convention meetings as "a proper precaution" because it averted "mistakes and misrepresentations until the business shall have been completed, when the whole may have a very different complexion from that in which the several parts might in their first shape appear if submitted to the public eye" (Farrand [14, 3:28,32])

[22] The view that keeping Council meetings secret is desirable is often found in the writings of scholars of European politics. For instance, Calleo [4, p. 270-271] states that "Whether making Council debates more open is, of course, debatable. Discrete decision making, dominated by expert advisers, has its advantages, especially in periods of prolonged economic difficulty."

of CEO options, that can lead to a less precise measure of firm profits.

We conclude by pointing to three possible extensions. First, as we argued in the Related Literature section, there are two ways of modeling cheap talk with career concerns: the expert model which is used here and the biased advisor model of Morris [26]. It would be interesting to know to what extent the results presented here carry over to the biased advisor model. Obviously, in the biased advisor model one must assume that the advisor knows his own type. One can envision situations in which observing only the consequence induces good advisor to follow their signals and bad advisors to act in a biased way. When the action is observed as well, good and bad advisors may pool on the "politically correct" action, which leads to a breakdown of both discipline and sorting. The question is under what circumstances this situation is more likely to arise.

Second, this paper has considered action revelation and consequence revelation. We have allowed for randomization, but we have not allowed for systematic biases. For instance, one action could be revealed with a higher probability than another action. It would be interesting to know whether the principal benefits from the introduction of such asymmetric information structures.[23]

Finally, here, information revelation is exogenous. Instead, in many agency relationships information is generated endogenously by the players. Lobbies and media gather and distribute intelligence on government policy. Shareholders and financial analysts question company management. The information that is available in equilibrium depends on the incentive of players to create transparency. Prendergast [32] considers the role of consumer monitoring, and the possible biases that it may induce. If consumers' interests are distant from the principal's goals, the principal may want to restrict the customers' ability to complain about the agent's performance.

References

[1] Christopher Avery and Margaret M. Meyer. Designing hiring and propotion procedures when evaluators are biased. Working paper, 1999.

[2] Timothy Besley and Robin Burgess. The political economy of government responsiveness: Theory and evidence from India. Working paper, 2001.

[23] Leaver [21] develops an expert model in which the agent knows his type and she considers the possibility that only one of the two actions is observed with positive probability. The model, which is applied to a regulatory setting, shows that a regulated industry can control the behavior of its career-motivated regulator by creating a biased information structure.

[3] Christine A. Botosan and Mary S. Harris. Motivations for a change in disclosure frequency and its consequences: An examination of voluntary quarterly segment disclosures. *Journal of Accounting Research* 38(2): 329–353, 2000.

[4] David P. Calleo. *Rethinking Europe's Future*. Princeton University Press, 2001.

[5] The Campaign for Freedom of Information. *Freedom of Information: Key Issues*. 1997 (available on www.cfoi.org.uk/pdf/keyissues.pdf).

[6] Pierre-André Chiappori and Bernard Salanié. Testing for asymmetric information in insurance markets. *Journal of Political Economy* 108(1): 56–78, 2000.

[7] Vincent Crawford and Joel Sobel. Strategic information transmission. *Econometrica* 50: 1431–1451, 1982.

[8] Jacques Crémer. Arm's length relationships. *Quarterly Journal of Economics* 110(2): 275–295, 1995.

[9] Sanjiv R. Das and Rangarajan K. Sundaram. On the regulation of fee structures in mutual funds. *Mathematical Modelling in Finance Vol III*. Courant Institute of Mathematical, forthcoming.

[10] Mathias Dewatripont, Ian Jewitt, and Jean Tirole. The economics of career concerns, Part I: Comparing information structures. *Review of Economic Studies* 66(1): 183–198, 1999.

[11] Alexander Dyck and Luigi Zingales. Why are private benefits of control so large in certain countries and what effects does this have on their financial development? Working paper, 2001.

[12] Jeffrey Ely, Drew Fudenberg, and David K. Levine. When is reputation bad? Mimeo, 2002.

[13] Jeffrey Ely and Juuso Välimäki. Bad reputation. Mimeo, 2001.

[14] Max Farrand (ed.). *The Records of the Federal Convention of 1787*. Yale University Press, 1967.

[15] John Fingleton and Michael Raith. Career concerns for bargainers. Working paper, October 2001.

[16] Maurice Frankel. Freedom of information: Some international characteristics. Working paper, The Campaign for Freedom of Information, 2001 (available on www.cfoi.org.uk/pdf/amsterdam.pdf).

[17] Robert Gibbons and Kevin J. Murphy. Optimal incentive contracts in the presence of career concerns: Theory and evidence. *Journal of Political Economy* 100(3): 468–505, 1992.

[18] Bengt Holmström. Moral hazard and observability. *Bell Journal of Economics* 10: 74–91, 1979.

[19] Bengt Holmström. Managerial incentive problems: A dynamic perspective. *Review of Economic Studies* 66(1): 169–182, 1999.

[20] Josef Lakonishok, Andrei Shleifer, and Robert W. Vishny. The structure and performance of the money management industry. *Brookings Papers on Economic Activity* Vol 1992: 339–379, 1992.

[21] Clare Leaver. Bureaucratic minimal squawk: Theory and evidence. Working paper, University College London, February 2002.

[22] Christian Leuz and Robert E. Verrecchia. The economic consequences of increased disclosure. *Journal of Accounting Research* 38(supplement): 91–124, 2000.

[23] Gilat Levy. Strategic consultation in the presence of career concerns. STICERD Discussion Paper TE/00/404, London School of Economics, 2000.

[24] Ulrike Malmendier and Geoffrey Tate. CEO overconfidence and corporate investment. Working paper, Harvard University, 2002.

[25] Eric Maskin and Jean Tirole. The politician and the judge: Accountability in government. Working paper, Toulouse University, 2001.

[26] Stephen Morris. Political correctness. *Journal of Political Economy*, forthcoming.

[27] Venky Nagar. The role of the manager's human capital in discretionary disclosure. *Journal of Accounting Research* 37(supplement): 167–185, 1999.

[28] Marco Ottaviani and Peter Sørensen. Information aggregation in debate: Who should speak first? *Journal of Public Economics* 81: 393–421, 2001.

[29] Marco Ottaviani and Peter Sørensen. Professional advice. Working paper, September 2001.

[30] Motty Perry and Larry Samuelson. Open- versus close-door negotiations. *RAND Journal of Economics* 25(2): 348–59, 1995.

[31] Torsten Persson and Guido Tabellini. *Political Economics*. MIT Press, 2002.

[32] Canice Prendergast. Consumers and agency problems. NBER Working Paper w8445, 2001.

[33] Canice Prendergast. A theory of "Yes Men". *American Economic Review* 83(4): 757–770, 1993.

[34] Canice Prendergast and Lars Stole. Impetuous youngsters and jaded oldtimers. *Journal of Political Economy* 104: 1105–34, 1996.

[35] Mark J. Rozell. *Executive Privilege: The Dilemma of Secrecy and Democratic Accountability.* Johns Hopkins University Press, 1994.

[36] Daniel Seidmann. Imperfect delegation and the norm of consensus. Newcastle University, 2002.

[37] David Scharfstein and Jeremy Stein. Herd behavior and investment. *American Economic Review* 80: 465–479, 1990.

[38] Russell B. Stevenson, Jr. *Corporations and Information: Secrecy, Access, and Disclosure.* Johns Hopkins University Press, 1980.

[39] Craig S. Tyle. Letter to the SEC on the frequency of mutual fund portfolio holdings disclosure. Investment Company Institute, 2001 (http://www.ici.org/port_holdings_com.html).

[40] Jeffrey Zwiebel. Corporate conservatism and relative compensation. *Journal of Political Economy* 103(1): 1–25: 1995.

8 Appendix: Nonlinear Utility

We refer to the General Case developed in Section 6. The only difference is that we now assume that the agent's payoff is given by $w(\pi(I))$ where $\pi(I)$ is the posterior on the agent's type given the available information I.

Proposition 15 *Take any ρ_u and $\rho'_a \geq \rho_a$. Under A1 and A2, there exists an equilibrium in which the agent uses a^* under (ρ_u, ρ'_a) only if there exists an equilibrium in which the agent uses a^* under (ρ_u, ρ_a).*

Proof. Suppose the agent uses a^*. Using Bayes' rule, the posteriors are still

$$\begin{aligned}
\pi(a,u) &= \begin{cases} \frac{\sum_{y:a^*(y)=a}\sum_{x:\omega(a^*(y),x)=u} q_g(y|x)p(x)\gamma}{\sum_{y:a^*(y)=a}\sum_{x:\omega(a^*(y),x)=u} q(y|x)p(x)} & \text{if } a \text{ is played in equilibrium} \\ \text{unrestricted} & \text{otherwise} \end{cases} \\
\pi(u) &= \frac{\sum_y \sum_{x:\omega(a^*(y),x)=u} q_g(y|x)\, p(x)\, \gamma}{\sum_y \sum_{x:\omega(a^*(y),x)=u} q(y|x)\, p(x)} \\
\pi(a) &= \begin{cases} \frac{\sum_{y:a^*(y)=a}\sum_x q_g(y|x)p(x)\gamma}{\sum_{y:a^*(y)=a}\sum_x q(y|x)p(x)} & \text{if } a \text{ is played in equilibrium} \\ \text{unrestricted} & \text{otherwise} \end{cases} \\
\pi(\emptyset) &= \gamma
\end{aligned}$$

Given (ρ_u, ρ_a), there exists an equilibrium in which the agent uses a^* if and only if, for every y,

$$\begin{aligned}
a^*(y) \in \arg\max_a \sum_u \Pr(u|a,y) \left(\rho_u\rho_a w(\pi(a,u)) + \rho_u(1-\rho_a) w(\pi(u))\right) \\
+ (1-\rho_u)\rho_a w(\pi(a)) + (1-\rho_u)(1-\rho_a) w(\pi(\emptyset)),
\end{aligned}$$

Claim 1: For every a and y,

$$\sum_u \left(\Pr(u|a^*(y),y) - \Pr(u|a,y)\right) w(\pi(u)) \geq 0$$

Proof of Claim 1: As before, we can show that

$$\pi(u) = \frac{\sum_y \sum_{x:\omega(a^*(y),x)=u} q_g(y|x)\, p(x)\, \gamma}{\sum_y \sum_{x:\omega(a^*(y),x)=u} q(y|x)\, p(x)}$$

is also nondecreasing in u. Hence $w(\pi(u))$ is nondecreasing in u. The claim is then proven through a standard argument combining first-order-stochastic-dominance and monotonicity.

Suppose the statement of the proposition is false. Then, there exists an a and a y such that

$$\begin{aligned}
&\sum_u \Pr(u|a^*(y),y) \left(\rho_u\rho_a w(\pi(a^*(y),u)) + \rho_u(1-\rho_a) w(\pi(u))\right) + (1-\rho_u)\rho_a w(\pi(a^*(y))) \\
<\ &\sum_u \Pr(u|a^*(y),y) \left(\rho_u\rho_a w(\pi(a^*(y),u)) + \rho_u(1-\rho_a) w(\pi(u))\right) + (1-\rho_u)\rho_a w\pi((a^*(y)))
\end{aligned} \tag{3}$$

but

$$\begin{aligned}
&\sum_u \Pr(u|a^*(y),y) \left(\rho_u\rho'_a w(\pi(a^*(y),u)) + \rho_u\left(1-\rho'_a\right) w(\pi(u))\right) + (1-\rho_u)\rho'_a w(\pi(a^*(y))) \\
\geq\ &\sum_u \Pr(u|a,y) \left(\rho_u\rho'_a w(\pi(a,u)) + \rho_u\left(1-\rho'_a\right) w(\pi(u))\right) + (1-\rho_u)\rho'_a w(\pi(a))
\end{aligned}$$

implying

$$\begin{aligned} & \left(\rho'_a - \rho_a\right) \rho_u \sum_u \left(\Pr\left(u|a^*\left(y\right), y\right) - \Pr\left(u|a, y\right)\right) w\left(\pi\left(u\right)\right) \qquad (34) \\ & + \left(\rho_a - \rho'_a\right) \left(\rho_u \sum_u \left(\Pr\left(u|a^*\left(y\right), y\right) w\left(\pi\left(a^*\left(y\right), u\right)\right) - \Pr\left(u|a, y\right) w\left(\pi\left(a, u\right)\right)\right) + \right. \\ & \left. \left(1 - \rho_u\right) \left(w\left(\pi\left(a^*\left(y\right)\right)\right) - w\left(\pi\left(a\right)\right)\right)\right) \\ < \; & 0 \end{aligned}$$

However, by Claim 1, in order for (33) to hold, it must be that

$$\rho_u \sum_u \left(\Pr\left(u|a^*\left(y\right), y\right) w\left(\pi\left(a^*\left(y\right), u\right)\right) - \Pr\left(u|a, y\right) w\left(\pi\left(a, u\right)\right)\right) + \left(1 - \rho_u\right) \left(w\left(\pi\left(a^*\left(y\right)\right)\right) - w\left(\pi\left(a\right)\right)\right) < 0$$

As $\rho'_a - \rho_a \geq 0$, the right-hand side of (34) is positive – contradiction. ■