Essays in Information and Incentives

Amir Habibi

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Department of Economics
University College London

March 31, 2020

Declaration

I, Amir Habibi, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Dedication

For my dad ~ you were the best.

Acknowledgements

I am very grateful to Martin Cripps and Konrad Mierendorff for all their time and encouragement and also to numerous other faculty at UCL for many useful conversations. I am also grateful to all my classmates, and in particular my friend Pete, for making my PhD an enjoyable experience. Finally, I would like to thank my parents and my wonderful girlfriend for all their love and support.

Abstract

This thesis consists of three chapters that theoretically consider different ways in which incentives can be provided through information. Chapter 1 is an information design with moral hazard problem in which a planner wants to optimally motivate a time-inconsistent agent by providing feedback. I provide conditions under which the optimal feedback takes a simple form of a cutoff. Chapter 2 and 3 consider whether or not a firm would want to choose to be transparent about pay within the organisation. Chapter 2 considers a static set-up, and Chapter 3 considers a dynamic set-up. The main finding—across the two chapters—is that as the value of retaining the best workers goes up, then transparency becomes more favourable.

Impact statement

I hope that the work I have presented in this thesis can have an impact both inside and outside academia. An increasing amount of information is being gathered in many areas of life, and the way in which we design how individuals receive feedback from this data can play an important role in providing incentives. My thesis aims to contribute to our theoretical understanding of this topic.

Chapter 1 has already been published in a well regarded academic journal, and I plan to publish the other chapters in due course—this is the first step to disseminating my outputs to the academic community. I also plan to continue presenting my work at academic departments and at academic conferences.

Beyond academia, I think that businesses and policy makers, will find Chapters 2 and 3 of interest. These chapters analyse when a firm would want to choose to be internally transparency about pay—this is clearly of interest to businesses and business schools. These chapters can also help a policy maker evaluate the impact of making pay universally transparent. The majority of countries do not have such policies, but there are some notable exceptions, for example tax records are public in Norway and state workers' salaries are public in California.

Contents

[1] A version of this chapter has been published in *Journal of Economic Behavior & Organization, Volume 169, January 2020, Pages 1-18.*

List of Figures

Introduction

This thesis consists of three chapters, connected by a common theme: how to provide incentives through information. I use theoretical models to bring new insights to informational problems encountered by individuals and organisations.

In Chapter 1, I study how a time-inconsistent agent can be incentivised by a planner who is able to commit to provide feedback on the agent's output. The leading example I consider is how a benevolent supervisor—aligned with the long-run interests of the student—would want to incentivise a time-inconsistent student who works on a project of unknown quality over two periods. The student makes unobserved effort choices that produce output that only the supervisor can observe. The supervisor is able to commit to provide feedback based on her observation. From a theoretical perspective, this is an information design problem (Kamenica and Gentzkow (2011)) but with moral hazard—meaning that the state variable is endogenously generated by the agent's effort choice. I show that the supervisor's feedback can motivate the student to take a higher effort in the first period than he would do without feedback. First, I show that the optimal mechanism always takes the simple form of recommending an action to the agent. Then, I provide sufficient conditions under which the feedback takes the simple form of a cutoff. The main normative result is that incentivising a time-inconsistent agent in this way makes all 'selves' of the agent unambiguously better off—this is not the case when a time-inconsistent agent is incentivised with monetary instruments.

In Chapters 2 and 3, I analyse when a firm would want to make pay transparent within their organisation. In order to answer this question I use a multidimensional signalling model (as in, for example, Bénabou and Tirole (2006)). The key result—that provides a clear, testable prediction—is that as the value of retaining the best workers goes up, then transparency becomes more favourable.

The intuition behind the model is as follows. Workers are unsure about their future prospects at the firm, and want to learn about this in order to decide whether or not to take an outside option. The firm, who has an information

advantage and wants to retain all workers, signals to them about their prospects through bonus payments. The firm also may face a cost shock—that affects their marginal cost of paying bonuses at that time—meaning it is possible it does not have resources to pay bonuses. In the equilibrium that I focus on, the firm only pays bonuses to the better workers when there is no cost shock—this makes the best workers more optimistic. The positive effect of transparency is that a worker becomes *less* pessimistic when he does *not* receive a bonus and observes that other workers also don't receive a bonus—this means that the firm was more likely to note be able to pay a bonus. The negative effect is when a worker does not receive a bonus but sees that another worker is paid a bonus. In this case he becomes *more* pessimistic since it is clear that the firm *was* able to pay a bonus. The firm needs to balance this trade-off when deciding whether or not to commit to transparency.

In Chapter 2, I analyse the static game—where continuation payoffs are exogenously given—which allows me to provide sufficient conditions for the existence and uniqueness of the separating equilibrium of interest. I can also derive comparative statics on the firm's value of retaining different quality workers. I consider a number of extensions of the static model that demonstrate the robustness of the comparative statics and provide some further insights. The model also provides a rational explanation for empirical findings in the relative pay literature that have thus far been explained by non-standard preferences.[2] I demonstrate this by considering the results in a prominent paper in this literature—Card et al. (2012). The explanation provided in the paper for their reduced-form empirical findings is that workers have social preferences. I show that my model—where agents are rational and have standard preferences—can provide an alternative explanation, and that this alternative explanation is important since it leads to different policy implications.

The theoretical innovation of the model, compared with the existing literature on multidimensional signalling, is that I consider how a sender would want to optimally design the informational environment in which the signalling game with multiple receivers takes place. From a technical perspective, in order to refine equilibria in my game, I introduce an appropriate intuitive criterion refinement that is appropriate for a game with multiple receivers, and I argue that it is in the spirit of the original refinement (Cho and Kreps (1987)).

In Chapter 3, I analyse a dynamic version of the model—where continuation payoffs become endogenised. This is a much richer model where workers'

[2]Relative pay is where a researcher analyses how a worker reacts to learning about a coworker's pay.

productivity changes over time. I show that an equilibrium exists in which in every stage game the principal plays the same strategy as in the equilibrium of interest in the static game—i.e. only pays bonuses to good workers when funds are available. Although I cannot obtain closed form solutions to obtain comparative statics, numerical results suggest that the comparative static results from the static model continue to hold. From a theoretical perspective the model contributes to the literature on reputations with changing types (for example, see Phelan (2006)).

Chapter 1

Motivation and Information Design[1]

1 Introduction

Imagine a supervisor who wants to motivate a student prone to procrastination to work on a long-term project. In each period the student finds work costly, but would find it worthwhile if the project is of a sufficiently high quality. To begin with both the student and supervisor are unsure of the underlying quality of the project and the prior is such that there is a conflict of interest: the supervisor would prefer the student to work on the project while the student would prefer to shirk. This conflict of interest arises because the supervisor wants to maximise the student's long term interests, while the time-inconsistent student wants to maximise from today's perspective in which future returns to effort are discounted more heavily.

Typically, in economic theory, monetary instruments are used to overcome such incentive problems. I assume that the supervisor does not have access to monetary instruments and instead can only incentivise the student by committing to provide feedback on the output she observes. I provide conditions in which feedback can be designed in such a way to incentivise effort today even in the presence of moral hazard—when the supervisor cannot observe effort. To do this it must be that exerting effort today makes it more likely that the feedback induces the student's future self to choose a more favourable level of effort from today's perspective. I provide conditions under which the feedback will take the particularly simple form of recommending working in future only if the earlier

[1] A version of this chapter has been published in *Journal of Economic Behavior & Organization, Volume 169, January 2020, Pages 1-18.*

output is above a certain level—a cutoff. The main contribution of the chapter is to formalise a mechanism under which moral hazard can be overcome in a setting where it occurs naturally, but where monetary incentives are not available. I also analyse the characteristics of the optimal mechanism, provide some testable implications and discuss welfare.

Formally, I study a persuasion game with three players. There is a planner (the supervisor) and two incarnations of a time-inconsistent agent (the student): Self 1 and Self 2 (today's self and tomorrow's self). Players have a common prior on an unknown state (the underlying quality of the project). In periods 1 and 2 the respective Self chooses an action from a binary set. The action is not observed by the planner. The action, together with the state, produces a payoff relevant output that is not observable to the agent. The agent incurs immediate costs for actions and enjoys output in the future. At the beginning of the game the planner chooses a mechanism that generates a signal conditional on Self 1's output. Self 2 is then able to learn about the state from this signal. Since the planner and Self 2 have different preferences over Self 2's choice of action, the game is similar to the standard Bayesian persuasion set up of Kamenica and Gentzkow (2011) (henceforth KG), but with the signal being conditioned on a variable that is endogenised by the third player (in this case Self 1).

First, I establish that any optimal mechanism is 'straightforward' in that the signal recommends an action to Self 2.[2] This is not a simple application of existing revelation principle results, this includes the ones in KG and Myerson (1986). The reason is that in my model the planner must commit to a mechanism before some players take private actions. In Myerson (1986) there is a different sequence of events: at the end of each stage the players privately report their private information to the mediator who then recommends actions for the next stage and so the mediator's recommendation can directly take into account the player's private actions. In the Appendix I provide an example of a three player game with the same sequence of events as my model, but where such a revelation principle result does not hold.

Then, I establish conditions under which the optimal feedback is a cutoff—where the high action is recommended if and only if output is above a given level. This includes the case when the project's quality is uniformly distributed. I provide a counter-example to show that a cutoff is not always optimal. The intuition behind this counter-example is as follows. There is an intermediate

[2]Since the set of actions is binary this means that it is without loss for there to be just two signals. Each of the two signals results in a posterior that results in Self 2 choosing one of the two different actions.

part of the distribution of quality with low density. The level of output resulting from high effort and this intermediate level of quality should not be rewarded by providing a useful signal, which in this case will be a signal recommending a high action. This is because rewarding this output makes low effort—which is more likely to result in this output—more favourable from the perspective of Self 1.[3]

Finally, I consider the welfare of the agent and show that the ability of the planner to commit to a mechanism makes both Self 1 and Self 2 better off from an ex ante point of view in expectation. This would not necessarily be the case if the agent was being motivated by monetary instruments.

In the final section of the chapter I discuss how changes in the preferences affect the results for the three player persuasion game I have analysed including, for example, a principal-agent problem where the agent is no longer time-inconsistent and the principal wants the agent to exert more effort than the agent would otherwise exert. In such a setting it is no longer without loss to use a straightforward mechanism that recommends an action. In my setting, the key thing that drives the revelation principle result is that the planner and Self 1 have the same preferences over Self 2's choice of action.

My model could also apply to other settings. For example, it might shed light on the design of messages sent by a fitness tracker phone application. Such devices are able to measure an individual's performance (for example, how many steps they take in a day) and can be programmed to send users messages conditional on the observed performance in order to encourage them to exert effort in the future.

2 Related literature

The model is related to several strands of literature. First, it fits into the growing literature on Bayesian persuasion following Kamenica and Gentzkow (2011). Second, it is related to the literature within contract theory that analyses provision of feedback to agents. Third, it is related to motivation of a time-inconsistent agent as in Bénabou and Tirole (2002).

Several recent papers also analyse a three player persuasion game where the state is endogenous. Rodina (2017) studies a two period career concerns model. However, unlike my model, the focus is on conditions for which full disclosure of output is optimal from the perspective of the principal who designs

[3]This is related to the MLRP for monotone rewards in a setting with moral hazard and monetary instruments.

the information structure. Farragut and Rodina (2017) characterise optimal disclosure policies for a school trying to induce effort from students through its grading system that affects students' employment prospects. Their analysis assumes that either effort or talent are perfectly revealed by output (or the 'score') of the agent. In my model, there will always be output levels for which there will be a signal extraction problem in the sense that multiple effort/talent pairs can generate the same output. Boleslavsky and Kim (2018) study a similar three player game with more general preferences but with a finite state space and continuous action space. This difference, and the more specific set of preferences I consider, allow me to show that a binary signal is optimal (there no is equivalent result in their paper), and highlights what specific features of my environment make this the case. The structure I impose also allows me to characterise when the optimal mechanism is a cutoff.

My model is also related to Kolotilin et al. (2017) who study a two player sender-receiver game where the receiver has private information. Before receiving a signal from the sender, the receiver is able to communicate his private information to the sender. The sender is able to commit in advance to a mechanism that sends signals conditional on this report. My setting is different for two reasons: first, the receiver (the agent in my model) does not have any private information at the start of the game, this private information is endogenous in that it is the effort choice of the receiver (the Self 1 incarnation of the agent) and this choice can be affected by the sender's (the planner's) choice of mechanism; second my receiver is actually two separate players (Self 1 and Self 2 incarnation of the agent) with different preferences.

A number of papers in contract theory analyse the role of feedback as a tool for providing incentives. Lizzeri et al. (2002) analyse feedback provision of an agent within a two period dynamic moral hazard model and so they combine information disclosure with monetary instruments. The feedback after the first period is a choice of either full revelation (feedback) or no revelation (no feedback) rather than more complicated information structures. The main result is that when rewards are designed by the principal to minimize costs along with the choice of feedback, no feedback is always optimal. Hansen (2013) studies optimal interim feedback in a two period additively-separable career concerns model. After two periods of costly effort the agent wants to induce a posterior belief of his talent to be above a given threshold. Effort and talent are assumed to be substitutes rather than complements (as in my model). This means that, unlike my model, raising the posterior belief on talent does not always induce higher effort in future, i.e. there is no demand for self confidence. The role of

feedback in this setting is to induce effort in the first period through a 'ratchet effect' rather than directly manipulating his future self's belief over the uncertain state of as in my model. This difference, along with the fact that the agent only wants to manipulate the belief of unknown talent above a certain threshold, mean that the optimal feedback rule is qualitatively very different to the one I find. In particular, to maximise the joint surplus, interim output is precisely disclosed only when period 1 output is close to the required threshold, allowing the agent to fine-tune effort in period 2. There are also a number of papers that study interim feedback in a tournament setting, for example, see Ederer (2010).

Bénabou and Tirole (2002) study a time-inconsistent agent's demand for self confidence. A person might not want to know their talent precisely since it will discourage their future self from taking action that the current self would want to take.[4] In their model, the focus is on intra-personal strategies that distort one's memory. In a later paper, Bénabou and Tirole (2004) analyse how today's (Self 1's) action can signal private information about the return on effort to a future self. In this setting memories are again—*endogenously*—distorted by the agent himself. In my model, I introduce a planner, who I think of as an outside agency, that is able to exploit the demand for self confidence—through influencing what the agent learns from his past actions—and so influence effort both today and in future. In contrast to Bénabou and Tirole (2004), what Self 2 will learn from Self 1's action is *exogenously* determined by the planner. In a similar setting, Mariotti et al. (2018) analyse optimal ex ante information disclosure in order to have a time-inconsistent agent take the most appropriate action given the state in each of two periods. Their information influences both Self 1's and Self 2's actions in the same way and so is similar to a static Bayesian persuasion model as studied by KG without moral hazard. In contrast, in my model Self 1 is incentivised to take an otherwise undesirable action in order to provide Self 2 with the information which will enable him to take the appropriate action from Self 1's perspective.

3 The model

In this section I introduce the formal model and then discuss my modelling assumptions.

[4]This observation was first made in Carrillo and Mariotti (2000).

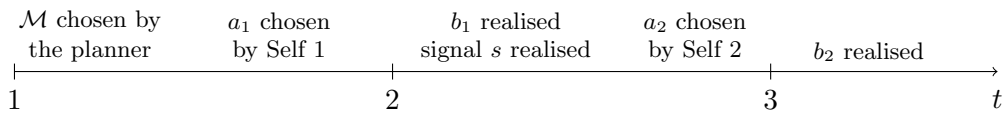

Figure 1.1: *Sequence of events.*

3.1 Set-up

There are three periods $t = 1, 2, 3$. There are two economic actors: an agent (he) and a planner (she). The agent's project has quality $\theta \in [0, 1]$ drawn from a distribution $F(\theta)$ with density $f(\theta)$. $F(\theta)$ is assumed to be continuous (i.e. there are no mass points).

Actions. In each period $t = 1, 2$ the agent chooses how much costly effort to exert $a_t \in A = \{a^H, a^L\}$, where $a^H > a^L > 0$. I will assume that $a^H = 1$ (this is without loss). Effort results in output $b_t = a_t\theta$.

The planner commits to a mechanism that sends a message to the agent at the start of period 2 depending on the output b_1. Formally a message is a function $\mathcal{M} : [0, 1] \times \{a^H, a^L\} \to S$ when the planner observes Self 1's action; and $\mathcal{M} : [0, 1] \to S$ when the planner does not observe Self 1's action. In both cases S is the space of signals (restricted to be finite for simplicity) and M denotes the set of all possible message functions. The sequence of events is depicted in Figure 1.1.

Payoffs. The agent's effort incurs an immediate cost and a long run benefit. Effort in period t incurs a cost $c_t(a_t)$, where $c_t(a^H) = c > 0$ and $c_t(a^L) = 0$. Benefits are enjoyed in period 3.[5] So the flow payoff of the agent in period t is given by

$$u_t = -c_t, \text{ for } t = 1, 2;$$
$$u_t = b_1 + b_2, \text{ for } t = 3.$$

The agent has time-inconsistent preferences, as in Laibson (1997), that are modelled by a different 'self' acting in each period. Preferences for Self 1 and 2

[5]Since there will be no long run discounting, this assumption is equivalent to benefits being consumed the period after they are produced. The key assumption is that benefits are enjoyed at some point in the future.

are given by:

$$U_1 = u_1 + \beta u_2 + \beta u_3, \tag{3.1}$$

$$U_2 = u_2 + \beta u_3, \tag{3.2}$$

where $\beta \in (0, 1)$ is the short run discount factor that captures present bias.[6]

The planner has preferences aligned with a fictitious Self 0 and so is aligned with the 'long run' interests of the agent. Formally, her preferences are given by:

$$U_0 = u_1 + u_2 + u_3. \tag{3.3}$$

Information. Players start with a common prior on θ. The agent does not see outputs until they are consumed in period 3, but does recall past actions. I analyse two cases for the planner: one in which she observes the agent's actions and another in which she does not.

Strategies. The planner's strategy is a choice of message function $\mathcal{M}$. Self 1's strategy is a function of the planner's choice of message $\sigma^1 : M \to A$. Self 2's strategy is a function of Self 1's action, the message function and the realisation of the signal. It is given by $\sigma^2 : M \times S \times A \to A$.

Equilibrium. I assume that the agent is sophisticated in the sense that Self 1 correctly anticipates the decision of Self 2. So the three players (Self 1, Self 2 and the planner) play a game of incomplete information and the relevant solution concept is perfect Bayesian equilibrium.[7] In equilibrium:

- Self 2 correctly updates his prior on θ given the choice of message function, realised signal and Self 1's action and then chooses a_2 to maximise his utility;
- Self 1 correctly anticipates Self 2's choice of a_2 and chooses a_1 to maximise his utility based on his prior on θ and the expected outcome of the signal given the choice of message function;
- The planner correctly anticipates the choice of action of Self 1 and 2 and chooses a message function to maximise her expected utility.

[6] I have taken δ, the long run discount factor, to be $\delta = 1$ to simplify the analysis.

[7] Equilibrium is also KG's 'sender preferred' perfect Bayesian equilibrium meaning that the action taken is the one preferred by the information designer which in my setting is the planner. This refinement is not important in my setting since indifference happens with zero probability in equilibrium I consider.

Assumptions. I make several assumptions to make sure that the analysis will be interesting. I maintain these assumptions throughout the chapter

Assumption 1. $\bar{c} \equiv \frac{c}{\beta(a^H - a^L)} < 1$.

This rules out the case that $a_2 = a^H$ will never be chosen by Self 2 for any posterior belief of θ.

Assumption 2. $E[\theta] \in (\beta\bar{c}, \bar{c})$.

This means from the planner's (Self 0's) perspective it is best for the agent to choose $a_1 = a^H$ but from a Self 1 perspective it is best to choose $a_1 = a^L$. This tension means that the planner wants to incentivise Self 1 to choose the action that he wouldn't otherwise take.

Assumption 3. $\mathbb{E}[\theta|\theta \geq \beta\bar{c}] < \bar{c}$.

The planner wants the agent to choose $a_t = a^H$ when $\theta \geq \beta\bar{c}$ while each self in period t wants to take $a_t = a^H$ when $\theta \geq \bar{c}$. This assumption ensures that the planner is not able to simply recommend $a_2 = a^H$ for values of θ that benefit her and have the agent follow the recommendation. For this assumption to hold it must be that β is sufficiently below 1.

3.2 Discussion of the model

A critical part of the model is that rewards are realised in the future while costs are immediate. For the settings I have in mind, such as working on an academic project, this assumption makes sense. It is also important that the output function captures a complementarity between the quality of the project and effort. This means that the marginal return on effort is greater for higher (expected) quality projects. This is also an assumption that makes sense in the settings I have in mind.[8]

I assume that the agent sees his effort but not the output. In a student-supervisor relationship the student sees how much effort he put in but might find it difficult to evaluate the quality of output. Instead it is the supervisor (planner) who has the expertise to evaluate the quality of output. In the main part of the analysis (Section 4.2) I assume that the planner cannot see the effort exerted by the agent. This is a standard moral hazard assumption that makes sense in the setting I have in mind: the supervisor cannot or finds it too costly to monitor the student at all times.[9]

[8] Both these assumptions are similar to the ones made in Bénabou and Tirole (2002).

[9] These assumptions, and the reasoning behind them, are similar to those in the literature on 'subjective evaluations' in moral hazard, see for example, MacLeod (2003).

I assume that the planner can commit to a mechanism in line with the Bayesian persuasion literature following KG. In a student-supervisor relationship there are a number of possible justifications for this. One possibility that is often used is that reputational concerns could ensure that a supervisor will provide the feedback that she has committed to provide. An alternative interpretation is that although the planner may not be able to commit to a mechanism, she may be able to commit to limit her attention to interpreting the agent's output. This could be done in such a way that it only allows her to determine if output is above or below a given threshold. Such a commitment might be made by limiting her time, or limiting her attention to only look at certain aspects of the project. This justification makes sense particularly when the optimal mechanism takes the form of a cutoff—as is the case in Propositions 4 and 5—and so highlights the importance of these results. In the other setting discussed in the introduction, a fitness tracker, the commitment assumption is much easier to implement. A mobile phone application (the planner) can be programmed in order to send messages dependent on data it collects about the agent's performance. Also, as mentioned by other authors in the literature, even when full commitment is too strong an assumption it provides a useful benchmark for what can be achieved.

An important insight from KG is that signals can be designed so that a single state can result in more than one signal. I have assumed that each outcome (equivalent to the state in KG) can only result in a single signal. With a continuous outcome space with no atoms, this assumption is without loss.[10]

Finally, I assume that the agent is sophisticated in the sense that he anticipates his time-inconsistency. There is a lot of evidence that people underestimate their self control problems. In my model the results would be qualitatively unchanged so long as the agent is not completely naive, i.e. they do understand that they have some level of time-inconsistency.

4 Analysis

In Section 4.1, I analyse the case in which the planner observes Self 1's action and so can make the signal contingent on this action. The planner's optimal mechanism is a binary signal recommending an action to Self 2 when Self 1

[10]Formally, this can be seen by generalising the linear programming problem used in proofs of Proposition 3, 4 and 5 by allowing $w(\theta)$ and $w_L(\theta)$ to be in $[0,1]$ rather than in $\{0,1\}$. The definitions of these choice variables would become and . The interior values represent mixed signals—sending different signals with positive probability for the same outcome. It is straightforward that the maximisation problem in the proof has the 'bang-bang' property meaning that these interior choices of $w(\theta)$ can never be optimal—this uses the fact that the distribution of θ has no atoms. It follows that the assumption made is without loss.

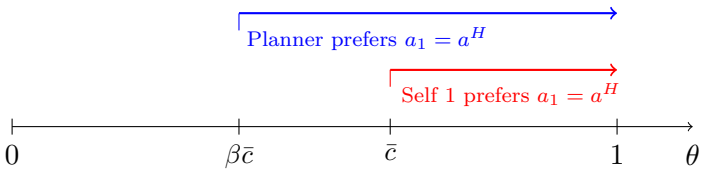

Figure 1.2: *Conflict of interest between the planner and Self 1 over Self 1's choice of action.*

takes the desired action ($a_1 = a^H$), and a single (uninformative) signal when Self 1 does not take the desired action. The recommendation of an action to Self 2 when $a_1 = a^H$ acts as a 'reward' since Self 2 takes the desired action (from the perspective of Self 1) more often. In contrast, the uninformative signal acts as a 'punishment' since it means that Self 2 will always take $a_2 = a^L$ which from the perspective of Self 1 is not necessarily optimal.

In Section 4.2, I analyse the more interesting case in which Self 1's action is not observable—this introduces moral hazard. Here the 'punishment' for deviating to $a_1 = a^L$ cannot be in the form of a completely uninformative signal. Instead, Self 1 is punished by having Self 2 being recommended the action that Self 1 wants Self 2 to take less often when $a_1 = a^L$ is chosen compared to when $a_1 = a^H$ is chosen.

4.1 Actions observable to the planner

First, I consider the incentives of Self 1 with no informative signal. In this case it is clear that Self 1 cannot influence Self 2's action. Keeping Self 2's action fixed, Self 1 gets a payoff $\beta a^H \theta - c$ from taking $a_1 = a^H$ and $\beta a^L \theta$ from taking $a_1 = a^L$. Comparing the value of these expressions means that Self 1 prefers $a_1 = a^H$ if and only if $\theta \geq \bar{c}$. Similarly the planner prefers $a_1 = a^H$ if and only if $\theta \geq \beta\bar{c}$. The conflict of interest between the planner and Self 1 over Self 1's choice of action is depicted in Figure 1.2. Note that the conflict over Self 2's action from the perspective of either the planner or Self 1 and Self 2 could be depicted in the same way.

Due to Assumption 2 ($\mathbb{E}[\theta] \in (\beta\bar{c}, \bar{c})$), the planner wants to incentivise the agent to choose $a_1 = a^H$, which Self 1 wouldn't otherwise choose, and can only do this through the use of an informative signal which will have a favourable effect on Self 2's action.

To ensure that Self 1 takes the desired action, a mechanism needs to be incentive compatible. This means that Self 1 finds it 'useful' in the sense that he benefits from Self 2 choosing $a_2 = a^H$ some of the time when Self 2 would

not make this choice without an informative signal, and loses relatively little from choosing $a_1 = a^H$ over $a_1 = a^L$. The mechanism must also deter Self 1 from deviating by providing the 'worst' possible signal. I show later that this is a completely uninformative signal. Formally, the constraint that makes Self 1 chooses $a_1 = a^H$ over $a_1 = a^L$ is

$$\beta a^H \mathbb{E}[\theta] - c + \beta \mathbb{E}_{\theta,a_2}[a_2\theta - c(a_2)|\mathcal{M}, a_1 = a^H] \geq 2\beta a^L \mathbb{E}[\theta], \tag{4.1}$$

where the second expectation on the LHS is over the expected value of θ and a_2 given the signal that will result from $a_1 = a^H$. The first part of the LHS is the expected benefit and cost of choosing $a_1 = a^H$, and the second part is the expected benefit and cost of the action that Self 2 will choose given the mechanism and Self 1's choice of action. The RHS is the payoff from $a_t = a^L$ in both periods since Self 2 will not update his prior and so will always choose $a_2 = a^L$.

If the signal when $a_1 = a^H$ is in the form of a cutoff where $s = h$ (a signal that induces a posterior such that $a_2 = a^H$ is chosen) is sent if and only if $b_1 \geq \hat{\theta}$, then the constraint can be written as[11]

$$\beta a^H \mathbb{E}[\theta] - c + \beta \int_{\hat{\theta}}^{1} (a^H\theta - c)\ dF(\theta) + \beta \int_{0}^{\hat{\theta}} (a^L\theta)\ dF(\theta) \geq 2\beta a^L \mathbb{E}[\theta]. \tag{4.2}$$

Since a_1 is known, θ can be perfectly inferred from b_1 and so a cutoff value for b_1 is equivalent to a cutoff value for θ (recall I assume that $a^H = 1$). The optimal cutoff $\hat{\theta}$ is given by θ^* and is the optimal disclosure that Self 1 would design in order to motivate Self 2 if he were free to choose the mechanism himself.[12,13] Formally, θ^* solves the equation

$$\int_{\theta^*}^{1} (\beta a^H\theta - c)\ dF(\theta) - \int_{\theta^*}^{1} (\beta a^L\theta)\ dF(\theta) = 0. \tag{4.3}$$

The optimal choice of mechanism for the planner is summarised in the following Proposition.

Proposition 1. *If $\hat{\theta} = \theta^*$ satisfies inequality 4.2, when a_1 is observable the planner can incentivise Self 1 to choose $a_1 = a^H$ and the following mechanism is optimal:*

[11] A cutoff will be the optimal mechanism from the perspective of Self 1. For details see the proof below and Appendix 1.

[12] The condition for this is derived in Appendix 1. From a technical perspective this is effectively a static persuasion game as in KG.

[13] Assumption 3, that says $\mathbb{E}[\theta|\theta \geq \beta\bar{c}] < \bar{c}$, rules out the case where $\theta^* \leq \beta\bar{c}$.

- *if* $a_1 = a^H$ *then send* $s = h$ *if* $b_1 \geq \theta^*$ *and* $s = l$ *otherwise;*
- *if* $a_1 = a^L$ *then send* $s = l$ *for all* b_1*.*

If $\hat{\theta} = \theta^*$ *does not satisfy 4.2 then the planner cannot incentivise Self 1 to choose* $a_1 = a^H$*.*

Throughout the thesis, proofs not in the main body of text are in the relevant Appendix.

In this section, with a_1 being observable to the planner, it is easy to deter Self 1 from choosing $a_1 = a^L$ by having a completely uninformative signal following this action. In the next section, when a_1 is not observable, it is still possible to deter Self 1 from choosing $a_1 = a_L$, but now it may be that Self 2 receives a favourable signal (from the perspective of Self 1) even when Self 1 chooses $a_1 = a_L$. The key to incentivising $a_1 = a_H$ will be to ensure that the favourable signal is received more often and is of more value when $a_1 = a_H$ compared to when $a_1 = a_L$.

4.2 Actions not observable to the planner

Now, I analyse what happens when the planner cannot see the action taken by Self 1. The signal the planner sends to Self 2 will be contingent only on the realisation of b_1 (the only information the planner will have other than the prior). The planner will be able to incentivise $a_1 = a^H$ since a higher action will induce a set of signals that is more likely to have Self 2 take the more favourable action (from the perspective of Self 1). I illustrate this through an example and then discuss the more general problem. I then establish conditions when the optimal mechanism is monotone partitional—i.e. takes the form of a cutoff.

Example

The project's quality is distributed as $\theta \sim U[0,1]$. The other parameter values are $a^H = 1$, $a^L = \frac{1}{2}$, $c = \frac{1}{9}$, $\beta = \frac{1}{3}$ (and so $\bar{c} \equiv \frac{c}{\beta(a^H - a^L)} = \frac{2}{3}$).

As derived in the previous section, with no informative signal, Self 1 prefers $a_1 = a^H$ if and only if $\theta \geq \bar{c}$ and the planner prefers $a_1 = a^H$ if and only if $\theta \geq \beta\bar{c}$. With the parameter values above this means that the planner would like Self 1 to choose $a_1 = a^H$ iff $\theta \in [\frac{2}{9}, 1]$ and Self 1 would choose $a_1 = a^H$ iff $\theta \in [\frac{2}{3}, 1]$. Since $\mathbb{E}[\theta] = \frac{1}{2}$ without any feedback Self 1 will choose $a_1 = a^L$, which is not the planner's preferred action—this means that Assumption 2 is satisfied.[14]

[14]Note also that it is easily verified that Assumption 3 is satisfied; and since $\bar{c} < 1$, Assumption 1 is also satisfied.

Now consider the following mechanism:

$$s = \begin{cases} h & \text{if } b_1 \in (\frac{1}{2}, 1], \\ l & \text{if } b_1 \in [0, \frac{1}{2}]. \end{cases}$$

If Self 1 chooses $a_1 = a^L$, then he cannot learn anything from the resulting signal and so enjoys no benefit from improving Self 2's action. To see this, note that if Self 1 chooses $a_1 = a^L$ then $b_1 \in [0, \frac{1}{2}]$ and so $s = l$ for any θ. Self 2 will always have the same prior and will choose $a_1 = a^L$.

If Self 1 chooses $a_1 = a^H$, then there is some potential benefit. In particular, if θ is sufficiently high, then Self 2 will update his posterior on θ and take an action that is desirable from the perspective of Self 1. To see this, first note that if $a_1 = a^H$, then $b_1 \in [0, 1]$ and so both signals $s = l, h$ are possible. If $s = h$, then $\mathbb{E}[\theta | s = h, a_1 = a^H] = \frac{3}{4}$; and if $s = l$, then $\mathbb{E}[\theta | s = l, a_1 = a^H] = \frac{1}{4}$. Notice that following $s = h$ the posterior is sufficiently high ($\geq \frac{2}{3}$) such that Self 2 will choose $a_2 = a^H$—the action that is preferred by Self 1 for the corresponding values of θ.

So Self 1 faces a tradeoff: choosing $a_1 = a^L$ is preferred if Self 2's action did not matter, but choosing $a_1 = a^H$ means that it is more likely that Self 2 takes Self 1's preferred action. It is straightforward to compute the expected payoffs in each case. Formally, Self 1 prefers $a_1 = a^H$ if:

$$\beta \mathbb{E}[\theta] a^H - c + \beta \int_0^{\frac{1}{2}} (a^L \theta) d\theta + \beta \int_{\frac{1}{2}}^1 (a^H \theta - c) d\theta \geq \beta \mathbb{E}[\theta] a^L + \beta \int_0^1 (a^L \theta) d\theta. \quad (4.4)$$

Notice that this is the same as 4.2 from the previous section but with $\hat{\theta} = \frac{1}{2}$. The reason for this is that with the given mechanism $s = h$ is sent when it must be the case that $a_1 = a^H$. With the given parameter values this inequality is satisfied.[15]

Although this mechanism incentivises $a_1 = a^H$ (an improvement on no mechanism), it is not optimal. The optimal mechanism is:

$$s = \begin{cases} h & \text{if } b_1 \in (\hat{b}, 1], \\ l & \text{if } b_1 \in [0, \hat{b}]; \end{cases} \quad (4.5)$$

where $\hat{b} \approx 0.42$. Optimality is proved later—see Corollary 1. Choosing $a_1 = a^L$ with this mechanism can lead to $s = h$. It is for intermediate values of θ that choosing $a_1 = a^H$ rather than $a_1 = a^L$ leads to $s = h$ instead of $s = l$.

[15] Calculations are in the Appendix 2.

Notice that this optimal mechanism takes a very simple form. First, there are just two signals meaning that a signal effectively recommends actions to Self 2. Second, the mechanism is also monotone partitional—i.e. it takes the form of a cutoff. The first property holds generally, but the second does not.

An example of a setting in which a non-monotone partition is optimal is as follows. $F(\theta)$ is a uniform distribution with a 'hole' in the interval $[\frac{4}{9}, \frac{1}{2}]$, formally:

$$f(\theta) = \begin{cases} \frac{18}{17} & \text{if } \theta \in [0, \frac{4}{9}] \cup (\frac{1}{2}, 1], \\ 0 & \text{otherwise.} \end{cases}$$

The other parameter values are as before: $a^H = 1$, $a^L = \frac{1}{2}$, $c = \frac{1}{9}$, $\beta = \frac{1}{3}$. It is still the case that for $\theta \geq \frac{2}{9}$ the planner wants Self 1 and 2 to take $a_t = a^H$ and that for $\theta \geq \frac{2}{3}$ Self 1 and 2 themselves want to take $a_t = a^H$; and now $\mathbb{E}[\theta] = \frac{307}{612} \approx 0.5 \in (\frac{2}{9}, \frac{2}{3})$ so Self 1 and the planner still have a conflict over a_1.[16]

The optimal mechanism is:

$$s = \begin{cases} h & \text{if } b_1 \in (\bar{b}, \frac{4}{9}] \cup (\frac{1}{2}, 1], \\ l & \text{if } b_1 \in [0, \bar{b}] \cup (\frac{4}{9}, \frac{1}{2}], \end{cases} \tag{4.6}$$

where $\bar{b} \approx 0.30$. It is straightforward to verify that Self 1 would prefer to choose $a_1 = a^H$ over $a_1 = a^L$ and that following $s = h$ Self 2 will choose $a_2 = a^H$.[17]

Beyond simply showing that a non-monotone mechanism may be optimal, there are some insights that can be taken from this example. In particular, the recommendation of $a_2 = a^H$ will not be made for a particular outcome b_1 when the low action $a_1 = a^L$ is more likely to result in that particular b_1. The example takes this to the extreme: it is not possible to achieve outcomes $b_1 \in (\frac{4}{9}, \frac{1}{2}]$ with $a_1 = a^H$. The intuition is similar to the monotone likelihood ratio property (MLRP) that drives a monotonic reward structure in a standard moral hazard problem with monetary incentives. The equivalent conditions in this environment are in Propositions 4 and 5, where I provide conditions for a monotone partitional mechanism to be optimal.

The general case

Now I consider the general problem. As illustrated in the example, to incentivise $a_1 = a^H$ an IC constraint for Self 1 needs to be satisfied. This trades off the increased likelihood of a benefit from a signal that induces Self 2 to take the desired action and the immediate cost of taking the less desirable action.

[16] Note that it is easily verified that Assumptions 1 and 3 are also satisfied.

[17] For details of why this is optimal see Appendix 3.

Let $\mathcal{B}_j = \{b_1 : \mathcal{M}(b_1) = j\}$ for $j \in S = \{h, l\}$ which is the set of outputs that result in the signal j; and let $\Theta_i^j \equiv \{\theta : a_i\theta \in \mathcal{B}_j\}$ which are the qualities θ for which the action i leads to the signal j.[18] The planner maximises over the sets $\mathcal{B}_h, \mathcal{B}_l$.[19]

The planner's problem when trying to incentivise $a_1 = a^H$ is:[20]

$$\sup_{\mathcal{B}_h, \mathcal{B}_l} \left\{ a^H \mathbb{E}[\theta] - c + \int_{\theta \in \Theta_H^h} (a^H\theta - c)\ dF(\theta) + \int_{\theta \in \Theta_H^l} (a^L\theta)\ dF(\theta) \right\},$$

subject to:

$$\beta a^H \mathbb{E}[\theta] - c + \beta \int_{\theta \in \Theta_H^h} (a^H\theta - c)\ dF(\theta) + \beta \int_{\theta \in \Theta_H^l} (a^L\theta)\ dF(\theta) \geq$$
$$\beta a^L \mathbb{E}[\theta] + \beta \int_{\theta \in \Theta_L^h} (a^H\theta - c)\ dF(\theta) + \beta \int_{\theta \in \Theta_L^l} (a^L\theta)\ dF(\theta) \quad \text{(IC)},$$

$$\mathbb{E}[\theta | \theta \in \Theta_H^h] \geq \bar{c} \quad \text{(H)},$$
$$\mathbb{E}[\theta | \theta \in \Theta_H^l] < \bar{c} \quad \text{(L)}.$$

The first constraint (IC) is the incentive compatibility constraint for Self 1. The second constraint (H) ensures that the signal $s = h$ leads to Self 2 taking the high action. The third constraint (L) ensures that the signal $s = l$ leads to Self 2 taking the low action and is implied from the constraint (H) and the prior belief (see Lemma 1). The problem can be simplified to:

[18] I have assumed that the signal is binary. I prove later that this is without loss.

[19] I assume that the sets are measurable and note that it must be that $\mathcal{B}_h \cap \mathcal{B}_l = \emptyset$ and $\mathcal{B}_h \cup \mathcal{B}_l = [0, 1]$.

[20] I use 'sup' rather than 'max' since a solution to the problem may not exist. The analysis is focused on the cases where the solution does exist, and where necessary I state the conditions under which a maximum exists.

$$\sup_{\mathcal{B}_h,\mathcal{B}_l}\left\{(a^H+a^L)\mathbb{E}[\theta]-c+\int_{\theta\in\Theta_H^h}((a^H-a^L)\theta-c)\ dF(\theta)\right\},$$

subject to:

$$\beta\mathbb{E}[\theta](a^H-a^L)-c+\beta\int_{\theta\in\Theta_H^h}((a^H-a^L)\theta-c)\ dF(\theta)$$
$$-\beta\int_{\theta\in\Theta_L^h}((a^H-a^L)\theta-c)\ dF(\theta)\geq 0\quad\text{(IC)},$$

$$\mathbb{E}[\theta|\theta\in\Theta_H^h]\geq\bar{c}\quad\text{(H)},$$
$$\mathbb{E}[\theta|\theta\in\Theta_H^l]<\bar{c}\quad\text{(L)}.$$

This formulation of the IC has the following interpretation. The first part of the LHS is the net benefit of choosing a^H over a^L (this is negative under the assumptions on $\mathbb{E}[\theta]$) and the two integrals are the benefit that taking a^H over a^L has on the signal sent to Self 2 from the perspective of Self 1.

Proposition 2. *In the planner's problem the use of binary signals is without loss of generality.*

Proposition 1 of KG states that in their two player persuasion game it is without loss to focus on 'straightforward signals'. Straightforward means that a signal recommends an action to the receiver. The result above is not a trivial application of the result in KG. The reason is that the agent induces different distributions of b_1 based on the choice of a_1.[21] In the proof I consider which signals result in different actions from Self 2. I show that if more than one signal results in the same action then the outcome would be the same if an alternative mechanism were used where these different signals were combined into just one signal. A difficulty in this environment is that such a change can have different effects to posterior beliefs on and off the equilibrium path. However, I show that

[21] Note also that it is also not possible to just use the revelation principle of Myerson (1986) to prove that a straightforward signal is without loss (as, for example, is done in Lemma 1 of Kremer et al. (2014)). The reason is that the planner must commit to a mechanism before Self 1 privately chooses an action. In a communication equilibrium with a mediator, the players report their private information to the mediator after which it recommends an action to them. So the communication equilibrium might require that for a given b_1, depending on the choice of a_1 the mediator might want to recommend different actions to Self 2 corresponding to the implied value of θ. The fact that in general more than two signals might be necessary here is similar to the result in Boleslavsky and Kim (2018) that optimal signal may induce three posterior beliefs when the state is binary. In Appendix 6 I provide an example of when reducing three signals to two is not without loss.

when these effects differ, it always results in the constraints being relaxed, which means that the alternative mechanism still results in the same outcome.

The problem can be simplified by showing that the constraint (L) is redundant. This will be useful for the next part of the analysis.

Lemma 1. *For any choice of $\mathcal{B}_h$ and $\mathcal{B}_l$ in the planner's problem, when (H) is satisfied (L) is also satisfied.*

Next I provide necessary and sufficient conditions for when it is possible to incentivise Self 1 to choose $a_1 = a^H$. To simplify notation, define

$$v(\theta) \equiv \begin{cases} (a^H - a^L)\theta - c & \text{if } \theta \in [0,1], \\ 0 & \text{otherwise;} \end{cases}$$

and

$$h(\theta) \equiv v(\theta)f(\theta) - \frac{1}{a^L}v(\theta/a^L)f(\theta/a^L).$$

$v(\theta)$ is the net benefit of choosing a^H over a^L for a given θ from the long run perspective of the agent. Notice that for small θ ($\theta < \beta\bar{c}$) $v(\theta)$ is negative, and for large θ ($\theta > \beta\bar{c}$) $v(\theta)$ is positive.

$h(\theta)$ can be interpreted as follows. For any $\theta \in \Theta_H^h$ it is the benefit to Self 1 from influencing Self 2's action when choosing $a_1 = a^H$ minus any benefit for these θ's when deviating to $a_1 = a^L$. This is weighted by the respective densities following the different choices of a_1 and multiplied by a factor $1/\beta$ throughout since the benefit is discounted by Self 1. Notice that for $\theta > a^L$ the second term—which represents the benefit from getting $s = h$ following $a_1 = a^L$—vanishes since $s = h$ will never be realised following $a_1 = a^L$. Also notice that for values of θ close to $\beta\bar{c}$ this will be negative since $v(\beta\bar{c}) = 0$ and $v(\beta\bar{c}/a^L) > 0$.

Proposition 3. *(Necessary) For it to be possible for the planner to be able to incentivise $a_1 = a^H$, $h(\theta)$ must satisfy*

$$\int_{\{\theta:h(\theta)\geq 0\}} h(\theta)d\theta \geq -\mathbb{E}[v(\theta)] + \frac{1-\beta}{\beta}c.$$

(Sufficient) The planner is able to incentivise $a_1 = a^H$ if $h(\theta)$ is such that the following inequalities are satisfied

$$\int_{\{\theta:h(\theta)\geq 0\}} h(\theta)d\theta \geq -\mathbb{E}[v(\theta)] + \frac{1-\beta}{\beta}c,$$

$$\mathbb{E}[\theta|\theta \in \{\theta : h(\theta) \geq 0\}] \geq \bar{c}.$$

The planner incentivises $a_1 = a^H$ by choosing $\mathcal{M}$ such that $\theta \in \Theta_H^h$ if and only if $h(\theta) \geq 0$.

The proof, in Appendix 1, reformulates the problem as a linear programming problem. The first statement (necessity) requires that the maximum possible benefit from choosing $a_1 = a^H$ rather than $a_1 = a^L$ derived from inducing a more favourable action from Self 2 (LHS of the inequality) is greater than the direct cost (RHS of the inequality). The second statement (sufficiency) adds the fact that a mechanism that recommends $a_2 = a^H$ whenever it is beneficial from the perspective of Self 1, must also induce a posterior sufficiently high such that Self 2 is obedient.

The previous result focused on when $a_1 = a^H$ could be incentivised. Now I turn attention to analysing what form the optimal mechanism takes. I am particularly interested in understanding when the optimal mechanism will be monotone partitional (i.e. $a_2 = a^H$ is recommended iff $b_1 \geq \hat{b}$), as I claimed was the case in the earlier example. I consider two cases:

- when the cutoff $\hat{b}$ is such that $\hat{b} \in [a^L, 1]$,
- when the cutoff $\hat{b}$ is such that $\hat{b} \in [0, a^L)$ (as in the example with a uniform distribution).

The first case will require a weaker set of assumptions to ensure that the optimal mechanism is monotone partitional. The reason for this is that in the first case Self 2 only gets the message $s = h$ when Self 1 chooses $a_1 = a^H$ (the maximum value of b_1 when $a_1 = a^L$ is $b_1 = a^L$). In effect, the resulting distribution of posteriors is similar to the posteriors resulting from the optimal mechanism when a_1 was observed in Section 4.1: only following $a_1 = a^H$ does Self 2 get an 'informative' signal.

The following Proposition provides conditions when the optimal solution is monotone partitional and the cutoff lies in the interval $\hat{b}_1 \in [a^L, 1]$. It will require one additional assumption on the distribution of θ.

Assumption 4. *$f(\theta)v(\theta)$ is increasing for $\theta \in [\beta\bar{c}, 1]$.*

Assumption 4 informally means that $f(\theta)$ is not decreasing too quickly for higher values of θ. For example, for a uniform distribution, where $f(\theta) = 1$, this assumption is satisfied. This assumption rules out distributions where the density vanishes for higher values of θ which may be of interest. However, note that the previous result (Proposition 3) does not rely on Assumption 4 and so

even for distributions where the density vanishes for high values of θ my results show when the planner can incentivise $a_1 = a^H$.

Define $\bar{\theta}$ to satisfy $\mathbb{E}[\theta|\theta \geq \bar{\theta}] = \bar{c}$.

Proposition 4. *A solution exists if*

$$\int_{\bar{\theta}}^{1} h(\theta)dF(\theta) \geq -\mathbb{E}[v(\theta)] + \frac{1-\beta}{\beta}c.$$

If a solution exists, Assumption 4 is satisfied and $\mathbb{E}[\theta|\theta \geq a^L] < \bar{c}$, then the optimal mechanism will be monotone partitional with the signal $s = h$ if $b_1 \geq \bar{\theta}$ and the signal $s = l$ otherwise.

The first part of the Proposition provides conditions when a solution exists (similar to Proposition 3) and the second part states that when a solution exists what additional conditions guarantee that the optimal mechanism will be monotone partitional and what the optimal mechanism will be.

The condition $\mathbb{E}[\theta|\theta \geq a^L] < \bar{c}$ means that if a mechanism was chosen with $\Theta_H^h = \{\theta : \theta \geq a^L\}$, Self 2 will not follow the recommendation from the signal $s = h$ (i.e. condition (H) will be violated). This means for a mechanism to be monotone partitional then it must have a cutoff $\hat{b} > a^L$. The intuition for why Assumption 4 must be satisfied is that if $f(\theta)$ is not decreasing too quickly, it means the that most 'useful' recommendations given to Self 2 (from the perspective of Self 1) will be the highest θ's (or equivalently b_1's), since this is where $v(\theta)$ is greatest and there is a sufficiently high mass at this point in the distribution from the density of $f(\theta)$.

The next result provides a condition when the optimal mechanism will be monotone partitional when the cutoff $\hat{b}$ might be such that $\hat{b} \in [0, a^L)$. It no longer relies on the condition $\mathbb{E}[\theta|\theta \geq a^L] < \bar{c}$ that was assumed in Proposition 4 that meant that the cutoff was such that $\hat{b} \in [a^L, 1]$, instead it imposes another condition on the distribution of θ.

Assumption 5. $\frac{v(\theta/a^L)f(\theta/a^L)}{v(\theta)f(\theta)}$ *is decreasing for* $\theta \in [\beta\bar{c}, a^L]$.

Proposition 5. *If Assumptions 4 and 5 are satisfied then the optimal mechanism (if it exists) is monotone partitional.*

When $\mathbb{E}[\theta|\theta \geq a^L] < \bar{c}$, as in Proposition 4, the cutoff will be such that $\hat{b} > a^L$ and so Assumption 5 is redundant. When $\mathbb{E}[\theta|\theta \geq a^L] > \bar{c}$, if the optimal mechanism is monotone partitional, the cutoff will be such that $\hat{b} < a^L$. Now Assumption 5 is important. The reason for this is that now the optimal mechanism will have $\theta' \in \Theta_H^h$ where $\theta' < a^L$, i.e. $s = h$ for some $\theta < a^L$. This

means that a deviation by Self 1 ($a_1 = a^L$) can lead to $s = h$. What now becomes important for ensuring that the optimal mechanism is monotone partitional is that in the region $[\beta\bar{c}, a^L)$ it is the highest values of θ that derive the greatest benefit from choosing $a_1 = a^H$ over $a_1 = a^L$ through their influence over Self 2's action. This is the case when Assumption 5 is satisfied. Note that this condition can be seen as an analogue of the MLRP in a standard moral hazard problem with monetary incentives.

Corollary 1 shows when $\theta \sim U[0,1]$, as in the example, the optimal mechanism is always monotone partitional.

Corollary 1. *$\theta \sim U[0,1]$ the optimal mechanism (when it exists) is monotone partitional.*

When Assumption 5 is not satisfied it is still possible to find the optimal mechanism. The objective function is maximised when $\mathcal{B}_h$ is chosen such that the $\theta \in \Theta_H^h$ minimise

$$\frac{v(\theta/a^L)f(\theta/a^L)}{v(\theta)f(\theta)},$$

and this should be done until one of the two constraints binds.

5 Discussion

5.1 Welfare

A difficulty of time-inconsistent preferences is in making statements about welfare. The seminal paper O'Donoghue and Rabin (1999) use a 'long run utility' as a welfare criterion and most papers continue to use this.[22] In my model this corresponds to the planner's (or 'Self 0's') preferences. This criterion is problematic since it may be that a future self is worse off as a result of the earlier self becoming better off. A stronger criterion, that is not liable to this criticism, is a Pareto improvement for all selves, which is used in, for example, Carrillo and Mariotti (2000). The following Proposition says that in my setting the use of information to incentivise the agent is a Pareto improvement over the situation in which no information can be provided to the agent.

Proposition 6. *The ability of the principal to commit to a mechanism leads to a (weak) increase in the (ex ante) expected payoff for all selves compared with the case in which there is no mechanism.*

[22] For example Galperti (2015) comments: *'as in most of the literature, this paper uses time-1 preferences to measure efficiency'* (here time-1 is in effect Self 0).

It should be noted that typically contracting with time-inconsistent individuals involves transfers or restricting the action set.[23] These often lead to situations in which a self is made worse off. For example, if someone wanted to ensure that they would carry out an exercise regime in the future they could sign a contract which imposes a large fine at a future date if they fail to exercise. The current self (who signs the contract) is better off, whereas the future self (who does the exercise but would prefer not to) is worse off.

Note that compared to a benchmark of full information revelation, the optimal mechanism will never be a Pareto improvement since Self 2 would always prefer more information. However, comparing to a benchmark where there is no feedback seems the most appropriate to measure the welfare implications of an intervention. This is comparing a situation in which someone does not have access to feedback and has access to feedback that is optimally designed from the perspective of their long run self.

5.2 Testable implications

With either observable or unobservable a_1 (Sections 4.1 and 4.2 respectively), I have provided conditions under the planner can incentivise $a_1 = a^H$. How does this condition change as the parameters of the model change? The comparative statics lead to intuitive results. In particular making $a_t = a^H$ less costly (decreasing c) or reducing the time-inconsistency problem of the agent (increasing β) both lead to the incentive constraints being relaxed which makes it easier to implement $a_1 = a^H$.[24]

Now, I compare different set-ups and their respective optimal mechanism in order to highlight the key differences and to propose some testable implications of the model. Consider the following three set-ups of the model:

1. Observable a_1;

2. Unobservable a_1 and $\mathbb{E}[\theta] \geq \bar{c}$;

3. Unobservable a_1 and $\mathbb{E}[\theta] \in (\beta\bar{c}, \bar{c})$.

Case 1 and 3 correspond to those studied in Sections 4.1 and 4.2 respectively. Case 2 represents the situation in which although a_1 is unobservable, since Self 1 will choose $a_1 = a^H$ (the desired action from the perspective of the planner) the problem is similar to when a_1 is observable. A testable implication

[23] See, for example, Amador et al. (2006) and Galperti (2015).

[24] The derivations are straightforward, and are given in Appendix 9.

of the resulting optimal mechanism is that a non-monotone partitional mechanism should only be observed when there is unobservable effort and $\mathbb{E}[\theta] \in (\beta\bar{c}, \bar{c})$ (case 3 above). Note that this is a necessary but not sufficient condition for a non-monotone partitional mechanism.

5.3 Reformulating the model without time-inconsistency

An alternative formulation of the model is with a time-consistent agent who again is unsure of the quality of the project and a planner who has preferences that are no longer aligned with the agent. This formulation could represent an ambitious supervisor and a time-consistent student, but, for example, no longer makes sense in the case of a supervisor who is benevolent in the sense that they want to maximise the student's true long run utility (since then all parties would have aligned interests).

In the reformulated version of the problem studied in Section 4.2, Self 0 still wants Self 1 and 2 to take the costly action for a greater set of θ's than they do themselves. The difference in the optimal mechanism will be with the IC constraint: Self 1 no longer wants to manipulate Self 2's beliefs but just inform Self 2 of the true state. Essentially now the planner uses the fact that Self 1 wants to inform Self 2 as precisely as possible and chooses the mechanism in a way that induces a high action. However, there is a technical problem with this formulation: Proposition 2, that means that the use of binary signals is without loss, may no longer hold. In the proof of Proposition 2 I exploit the fact that the planner and Self 1 are aligned about Self 2's choice of action.[25] In Appendix 6 I provide a set of preferences for a three player game where having a third signal is not without loss. Understanding for which preferences the use of binary signals is without loss is an interesting question that I leave for future research.

[25] In particular this means that $(a^H - a^L)\theta - c > 0 \; \forall \; \theta \in \Theta_L^h$ may not necessarily be true.

Appendix to Chapter 1

1 Optimal θ^* for Proposition 1

Below I characterise the optimal mechanism that Self 1 would choose to optimally motivate Self 2 that is used in Proposition 1.

Lemma 2. *The optimal signal will recommend an action to Self 2, i.e.* $s \in \{h, s\}$.

Proof. This follows from Proposition 1 in KG. □

Lemma 3. *The optimal signal will be cutoff so that above a certain level of* $\theta = \theta^*$ *the signal recommends taking action* $a_2 = a^H$ *with the signal* $s = h$.

Proof. This follows from Proposition 1 and the example in section 5.2 in Ivanov (2015).

□

The cutoff is pinned down by making Self 2 indifferent between actions. The condition for the agent to prefer taking action a^H over a^L on receiving the signal $s = h$ is

$$\int_{\hat{\theta}}^{1} (\beta a^H \theta - c)\ dF(\theta) \geq \int_{\hat{\theta}}^{1} (\beta a^L \theta)\ dF(\theta) \tag{1.1}$$

Lemma 4. *Let* $\hat{\theta} = \theta^*$ *solve the inequality 1.1 when it binds. If* $\mathbb{E}[\theta] < \bar{c}$ *then* $\max\{\theta^*, \beta\bar{c}\}$ *defines an optimal cutoff. If* $\mathbb{E}[\theta] \geq \bar{c}$ *then if* $\beta a^H \mathbb{E}[\theta] - c < \int_{\theta^*}^{1} (\beta a^H \theta - c)\ dF(\theta) + \int_{0}^{\theta^*} (\beta a^L)\theta\ dF(\theta)$ *then* $\max\{\theta^*, \beta\bar{c}\}$ *defines the cutoff for the optimal signal, otherwise an uninformative signal is optimal.*[26]

Proof. When $\mathbb{E}[\theta] < \bar{c}$ then without any more information $a_2 = a^L$ and so the agent can only gain from additional information that will lead to the action being changed appropriately. When $\mathbb{E}[\theta] \geq \bar{c}$ the inequality condition above ensures that remaining ignorant is not in the best interest of the planner. This

[26] An uninformative signal has $|S| = 1$ or equivalently $\mathcal{M}(b) = s$ for all b.

would be the case if the prior was sufficiently high that all selves would choose $a = a^H$ without further information and that the optimal signal would cause more harm when it is revealed that $\theta \in (\beta\bar{c}, \theta^*)$ than benefit when it is revealed that $\theta \in [0, \beta\bar{c}]$. □

2 Calculations for example in Section 4.2.1

In this section I provide calculations for the examples provided in the main text.

For the mechanism

$$s = \begin{cases} h & \text{if } b_1 \in [0, \frac{1}{2}] \\ l & \text{if } b_1 \in [\frac{1}{2}, 1] \end{cases} \tag{2.1}$$

The IC constraint given by inequality 4.4 is

$$\beta\mathbb{E}[\theta]a^H - c + \beta\int_0^{\frac{1}{2}}(a^L\theta)d\theta + \beta\int_{\frac{1}{2}}^1(a^H\theta - c)d\theta \geq \beta\mathbb{E}[\theta]a^L + \beta\int_0^1(a^L\theta)d\theta \tag{2.2}$$

The RHS

$$\beta\mathbb{E}[\theta]a^H - c + \beta\int_0^{\frac{1}{2}}(a^L\theta)d\theta + \beta\int_{\frac{1}{2}}^1(a^H\theta - c)d\theta = \frac{1}{6} - \frac{1}{9} + \frac{1}{54} + \frac{1}{8} = \frac{43}{216} \tag{2.3}$$

The LHS

$$\beta\mathbb{E}[\theta]a^L + \beta\int_0^1(a^L\theta)d\theta = \frac{1}{12} + \frac{1}{12} = \frac{36}{216} \tag{2.4}$$

and so for the parameter values the inequality is satisfied.

The proof that for $f \sim U[0,1]$ a monotone partition is optimal is in Corollary 1. As in the proof of Proposition 5 the cutoff is reduced until either (IC) or (H) binds. In this case it will be that (IC) binds first at $\hat{b}_1 \approx 0.42$. This can been seen from the calculation below for (IC)

$$\begin{aligned}\beta\mathbb{E}[\theta]a^H - c + \beta\int_0^{\hat{b}_1}(a^L\theta)d\theta + \beta\int_{\hat{b}_1}^1(a^H\theta - c)d\theta \\ - \beta\mathbb{E}[\theta]a^L - \beta\int_0^{2\hat{b}_1}(a^L\theta)d\theta - \beta\int_{2\hat{b}_1}^1(a^H\theta - c)d\theta \approx 0\end{aligned} \tag{2.5}$$

and (H)

$$\mathbb{E}[\theta|\theta \geq \hat{b}_1] > \bar{c}. \tag{2.6}$$

3 Optimality of non-monotone partitional example

Consider the mechanism given in the text

$$s = \begin{cases} h & \text{if } b_1 \in (\bar{b}, \frac{4}{9}] \cup (\frac{1}{2}, 1], \\ l & \text{if } b_1 \in [0, \bar{b}] \cup (\frac{4}{9}, \frac{1}{2}], \end{cases} \tag{3.1}$$

where $\bar{b} \approx 0.30$. To show this is optimal first it must be shown that it is both incentive compatible (i.e. induces $a_1 = a^H$) and Self 2 follows the recommendation (i.e. following $s = h$ Self 2 chooses $a_2 = a^H$). Both of these can easily be verified numerically. Note that the incentive constraint is slack, while the constraint on the recommendation to Self 2 is binding.

To see that this mechanism is optimal consider changes to send the signal $s = h$ for other values of b_1.

First, consider the effect of sending the signal $s = h$ in the interval $b_1 \in (\frac{4}{9}, \frac{1}{2}]$. If Self 1 chooses $a_1 = a^H$, this will not change the values of θ for which Self 2 gets the signal $s = h$. The only changes are when Self 1 deviates to $a_1 = a^L$. This means that Self 2 takes a more favourable action only off the equilibrium path, which strengthens the incentive constraint and leaves the objective function unchanged. Such a change in the mechanism does not benefit the principal.

Second, consider the effect of sending the signal $s = h$ when $b_1 \leq \bar{b}$. For $b_1 < \frac{2}{9}$ this means that when Self 1 chooses $a_1 = a^H$ Self 2 is recommended the action $a_2 = a^H$ for values of θ for which the planner (and Self 1) would not want Self 2 to take this action. This cannot improve the mechanism. If $s = h$ for some $b_1 \in [\frac{2}{9}, \bar{b})$ then without any other changes Self 2 would no longer follow the recommendation to choose $a_2 = a^H$ following $s = h$. It might be possible that 'swapping' some higher b_1's with b_1's in the interval $[\frac{2}{9}, \bar{b})$ might lead to the recommendation still being followed and also an increase in the planner's objective function. The formal argument for why this cannot be the case follows from a very similar argument to the ones given in Propositions 4 and 5. The intuition is that shifting higher levels of θ from those for which Self 2 is recommended $a_2 = a^H$ cannot benefit the principal. This is the case with the given distribution since the density is equal for all θ.

4 Proofs

4.1 Proof of Proposition 1

If the IC constraint (4.2) is satisfied Self 1 will choose $a_1 = a^H$. So the (uninformative) signal sent in the case that $a_1 = a^L$ does not affect Self 1's payoff (it is off the equilibrium path). The payoff of Self 1 will be the LHS of the IC constraint. The first part is the payoff from choosing $a_1 = a^H$ and the second part is from the optimal choice of mechanism to influence Self 2 to take the desired action (follows from the derivation in the Appendix 1). Adjusting the signal in the case that $a_1 = a^H$ (i.e. changing θ^*) cannot improve the payoff of the planner, since by design it is the best policy from the perspective of Self 1 (recall that the planner and Self 1 have aligned interests over Self 2's choice of action).

To show that if 4.2 is not satisfied there is no way for the planner to create better incentives for Self 1 to choose $a_1 = a^H$, it is also necessary to show that following $a_1 = a^L$ the uninformative signal is the worst punishment that the planner is able to do to Self 1. The reason that this must be the case is as follows. For an $\mathcal{M}$ to result in a lower payoff for Self 1 it must be that there is some signal $s' \in S$ that is sent with positive probability that results in Self 2 changing his action (i.e. choose $a_2 = a^H$). For this to be the case it must be that $\mathbb{E}[\theta|s = s'] \geq \bar{c}$. Since Self 1 has preferences such that $a_2 = a^H$ is preferred when $\mathbb{E}[\theta|s = s'] \geq \beta\bar{c}$, following any such signal s Self 1's payoff has improved compared with the case in which there is no informative signal. Therefore, it is not possible to construct an $\mathcal{M}$ that lowers Self 1's payoff following $a_1 = a^L$ compared to when he receives a completely uninformative signal.

4.2 Proof of Proposition 2

I begin by proving that if there is an optimal mechanism with three signal outcomes it is without loss to reduce this to a mechanism with just two signal outcomes. I then extend my argument to prove the case with more than three signal outcomes.

I will use the notation $a_2^*(s, a_1) \equiv \arg\max_{a_2} a_2(s, a_1)$. Informally this is the optimal action of Self 2 following the signal s and action a_1 for a given mechanism.[27]

[27] Note that the assumption that the equilibrium is 'sender preferred' means that that in the case of indifference the action favours the planner and Self 1 who have aligned preferences over Self 2's actions.

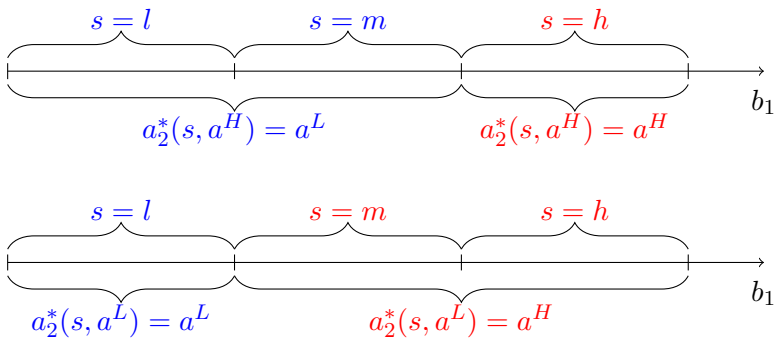

Figure 1.3: ***Actions induced on and off equilibrium path:*** *The top line represents the resulting actions on the equilibrium path, i.e. following* $a_1 = a^H$*.The bottom line represents the resulting actions off the equilibrium path, i.e. following* $a_1 = a^L$*.*

Three signal outcomes. Suppose that there is a mechanism with three signal outcomes $S = \{h, m, l\}$. Note that any optimal mechanism must have at least one signal $s \in S$ with $a_2^*(s, a^H) = a^H$. If not, incentivising $a_1 = a^H$ will not be possible since Self 2 will just take the same action as if Self 1 takes the less costly action $a_1 = a^L$ (the prior on θ is never updated). It also cannot be the case that all signals $s \in S$ result in $a_2^*(s, a) = a^H$ for any $a \in A$ since this would violate Bayesian consistency. So in an optimal mechanism with three signals, either one or two of the three possible signals must result in $a_2^*(s, a) = a^H$ for $a \in \{a^H, a^L\}$. I now show that when more than one signal results in the same action for Self 2, it is without loss to replace these signals with a single signal which in effect recommends an action to Self 2.

Begin with the case where only one signal s results in $a_2^*(s, a^H) = a^H$, without loss let this be $s = h$ and so $s = m, l$ means Self 2 chooses $a_2^*(s, a^H) = a^L$. Consider what happens after Self 1 deviates and chooses $a_1 = a^L$. Following the realisation of s the agent will have a different posterior of θ compared with what he would have if he had not deviated and chosen $a_1 = a^H$. If for all $s \in S$ $a_2^*(s, a^H) = a_2^*(s, a^L)$ then clearly it is without loss to replace l, m with a single signal l'. However it might be that $s = m, h$ means Self 2 chooses $a_2^*(s, a^L) = a^H$ and $s = l$ means Self 2 chooses $a_2^*(s, a^L) = a^L$. This scenario is depicted in Figure 1.3. In this situation combining the signals m, l will have a different impact on *and* off the equilibrium path and so it is not immediate that replacing m and l with l' is going to be without loss.

Now I show that in this case it is indeed without loss to combine the signals $s = m, l$. On the equilibrium path combining the signals m, l will have no

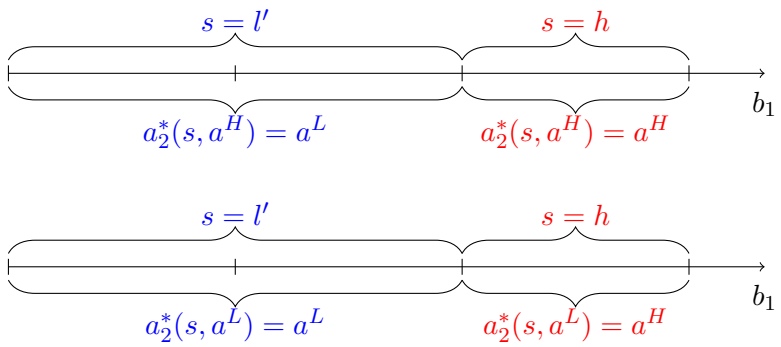

Figure 1.4: ***Actions induced on and off equilibrium path with two signals:*** *Combining the signals m, l into a single signal l'. Note that compared to Figure 1.3 in the lower line, off the equilibrium path, Self 2 has changed action for b_1 that previously resulted in $s = m$.*

effect. Off the equilibrium path the potentially problematic case is where the agent deviates to $a_1 = a^L$ and $s = m$ results in $a_2^*(m, a^L) = a^L$ rather than $a_2^*(m, a^L) = a^H$. This is depicted in Figure 1.4. To verify that this will not change the outcome of the optimisation problem we need to verify that the constraints are still satisfied and that there is no change to the objective function. The only relevant constraint to consider is the IC. This is because this is the only place where off path deviations enter into the maximisation problem. The set Θ_L^h will be smaller following the change in signals and so the IC will still be satisfied (note that this follows since what is inside the relevant integral is always positive: $(a^H - a^L)\theta - c > 0 \; \forall \; \theta \in \Theta_L^h$).

The second case I consider is where 2 signals, $s = m, h$, result in $a_2^*(s, a^H) = a^H$. In this case I show that due to the nature of the preferences it must be that off the equilibrium path Self 2 makes the same choices (i.e. $a_2^*(s, a^H) = a_2^*(s, a^L)$ for all s). This means it will be without loss to combine the signals $s = m, h$.

It cannot be that $a_2^*(s, a^L) = a^H$ for $s = l, m, h$ due to Bayesian consistency. So the only cases to consider are that $a_2^*(s, a^L) = a^H$ for either $s = h$ or for no s. These cases both mean that given a signal s, $a_1 = a^L$ leads to a more pessimistic expectation of θ compared to $a_1 = a^H$ since $a_2 = a^H$. I rule these two cases out meaning it must be that $a_2^*(s, a^L) = a^H$ for $s = m, h$.

To do this I show that it must be the case that $\mathbb{E}[\theta|\theta \in \Theta_L^s] \geq \mathbb{E}[\theta|\theta \in \Theta_H^s]$ for $s \in \{h, m, l\}$ and so it is not possible that for some $s \in S$ that $a_2^*(s, a^L) = a^L$ whilst $a_2^*(s, a^H) = a^H$. To see this first note that $\nexists \; \theta' \in \Theta_H^s$ s.t. $\theta' > \sup \Theta_L^s$ and $\nexists \; \theta'' \in \Theta_L^s$ s.t. $\theta'' < \inf \Theta_H^s$. Now define $\hat{\Theta}^s \equiv \Theta_H^s \cap \Theta_L^s$ (this is the intersect

of the two sets). Using this

$$\begin{aligned}\mathbb{E}[\theta|\theta \in \Theta_H^s] &= p_1\mathbb{E}[\theta|\theta \in \hat{\Theta}^s] + (1-p_1)\mathbb{E}[\theta|\theta \in \Theta_H^s \setminus \hat{\Theta}^s] \\ &\leq \mathbb{E}[\theta|\theta \in \hat{\Theta}^s] \\ &\leq p_2\mathbb{E}[\theta|\theta \in \hat{\Theta}^s] + (1-p_2)\mathbb{E}[\theta|\theta \in \Theta_L^s \setminus \hat{\Theta}^s] \\ &= \mathbb{E}[\theta|\theta \in \Theta_L^s];\end{aligned}$$

where $p_1 \equiv Pr[\theta \in \hat{\Theta}^s|\Theta_H^s]$ and $p_2 \equiv Pr[\theta \in \hat{\Theta}^s|\Theta_L^s]$. The equality in the first line just follows from the law of total probability. The inequality going from the first to the second line follows from the fact $\mathbb{E}[\theta|\theta \in \hat{\Theta}^s] \leq \mathbb{E}[\theta|\theta \in \Theta_H^s \setminus \hat{\Theta}^s]$ (this is due to $\nexists\ \theta' \in \Theta_H^s$ s.t. $\theta' > \sup \Theta_L^s$). In a similar way I derive the inequality going from the second to the third line. The final equality again just follows from the law of total probability.

More than three signal outcomes. When $|S| > 3$ a similar argument can be used to eliminate the need for signals until $|S| = 2$. Let $S_i^j = \{s : a_2^*(s, a_i) = a_j\}$. First note that it must be $S_H^H \subseteq S_L^H$ since following $a_1 = a^L$ and a given realisation of s Self 2 will have a more optimistic (higher) posterior of θ than when $a_1 = a^H$ and so will never take a lower action. If $S_H^H = S_L^H$ then replacing the message function so that all signals in S_H^H are replaced by a single signal $s = h$ and all signals all signals in S_L^H are replaced by a single signal $s = l$ can be done without loss. If $S_H^H \subset S_L^H$ then as discussed in the three signal case, replacing signals with $s = h$ and $s = l$ will have a different effect on and off the equilibrium path. However, as in the three signal case, replacing signals in S_H^H by a single signal $s = h$ leads to the signals that were previously in $S_L^H \setminus S_H^H$ to now lead to $a_2 = a^L$ following $a_1 = a^L$. As in the three signal case this leads to the IC being weakened and so can be done without loss of generality.

4.3 Proof of Lemma 1

Assume (H) is satisfied so $\mathbb{E}[\theta|\theta \in \Theta_H^h] \geq \bar{c}$. By Assumption 2, $\mathbb{E}[\theta] < \bar{c}$. Define $p \equiv Pr[\theta \in \Theta_H^h]$. The result follows from:

$$\begin{aligned}p\mathbb{E}[\theta|\theta \in \Theta_H^h] + (1-p)\mathbb{E}[\theta|\theta \in \Theta_H^l] &= \mathbb{E}[\theta] \\ p\mathbb{E}[\theta|\theta \in \Theta_H^h] + (1-p)\mathbb{E}[\theta|\theta \in \Theta_H^l] &< \bar{c} \\ p\bar{c} + (1-p)\mathbb{E}[\theta|\theta \in \Theta_H^l] &< \bar{c} \\ \mathbb{E}[\theta|\theta \in \Theta_H^l] &< \bar{c}.\end{aligned}$$

Where the first line just follows from the law of total probability; the second line follows from Assumption 2; the third line follows from (H); and the final line is straightforward algebra.

4.4 Proof of Proposition 3

In order to facilitate the analysis I rewrite the optimisation problem as a linear programming problem. Choosing the sets $\mathcal{B}_h$ and $\mathcal{B}_l$ is equivalent to choosing values of $\theta \in [0,1]$ to include in the set $\mathcal{B}_h$. The maximisation is now over $w(\theta)$ which is defined as

$$w(\theta) \equiv \begin{cases} 1 & \text{if } \theta \in \Theta_H^h, \\ 0 & \text{otherwise.} \end{cases}$$

The problem can be written as:

$$\sup_{\{w(\theta)\}_{\theta=0}^{\theta=1}} \left\{ \int v(\theta) w(\theta) \; dF(\theta) \right\},$$

subject to:

$$\int v(\theta) w(\theta) \; dF(\theta) - \int v(\theta) w_L(\theta) \; dF(\theta) \geq -\mathbb{E}[v(\theta)] + \frac{1-\beta}{\beta} c \quad \text{(IC)},$$

$$\frac{\int \theta w(\theta) \; dF(\theta)}{\int w(\theta) \; dF(\theta)} \geq \bar{c} \quad \text{(H)},$$

$$w_L(\theta) = w(a^L \theta) \quad \text{(W)}.$$

Note that in choosing $w(\theta)$ for all θ pins down the sets Θ_L^h, Θ_H^l and Θ_L^l (using (W)). Substituting (W) into (IC) gives

$$\int v(\theta) w(\theta) \; dF(\theta) - \int v(\theta) w(a^L \theta) \; dF(\theta) \geq -\mathbb{E}[v(\theta)] + \frac{1-\beta}{\beta} c.$$

To simplify, I change variables in the second integral on the LHS by making

the substitution $\theta' = a^L\theta$. This simplifies the inequality to

$$\int v(\theta)w(\theta)\ dF(\theta) - \frac{1}{a^L}\int v(\theta'/a^L)w(\theta')\ dF(\theta') \geq -\mathbb{E}[v(\theta)] + \frac{1-\beta}{\beta}c,$$
$$\int \left[v(\theta)f(\theta) - \frac{1}{a^L}v(\theta/a^L)f(\theta/a^L)\right] w(\theta)\ d\theta \geq -\mathbb{E}[v(\theta)] + \frac{1-\beta}{\beta}c,$$
$$\int h(\theta)w(\theta)\ d\theta \geq -\mathbb{E}[v(\theta)] + \frac{1-\beta}{\beta}c.$$

Now I derive the necessary condition. When $w(\theta)$ is chosen such that $w(\theta) = 1$ for θ if and only if $h(\theta) \geq 0$, the LHS of the IC attains its maximum possible value. Note the choice of w does not affect the RHS. So if the inequality is not satisfied for this choice of $w(\theta)$ then it will not be possible to satisfy the inequality for any other choice of $w(\theta)$ and hence a solution to the optimisation problem does not exist.

In the sufficient condition, choosing $w(\theta) = 1$ for θ if and only if $h(\theta) \geq 0$, must satisfy both inequality constraints. In particular, the second condition provided ensures that for this choice of $w(\theta)$ the inequality constraint (H) is also satisfied and therefore provides a sufficient condition for existence of a solution.

4.5 Proof of Proposition 4

First note that $\mathbb{E}[\theta|\theta \geq a^L] < \bar{c}$ means there exists a unique $\bar{\theta} \in [a^L, 1]$ that satisfies $\mathbb{E}[\theta|\theta \geq \bar{\theta}] = \bar{c}$ due to the Intermediate Value Theorem.[28] Next note that it must be that $\bar{\theta} \geq \beta\bar{c}$. To see this if $\bar{\theta} < \beta\bar{c}$ then $\mathbb{E}[\theta|\theta \geq \beta\bar{c}] \geq \bar{c}$ which violates the assumption $\mathbb{E}[\theta|\theta \geq \beta\bar{c}] < \bar{c}$. $\bar{\theta} \geq \beta\bar{c}$ implies $v(\theta) \geq 0$ for all $\theta \in [a^L, 1]$.

The planner's maximisation problem for the region $\theta \in [a^L, 1]$ is:[29]

$$\sup_{\{w(\theta)\}_{\theta=a^L}^{\theta=1}} \left\{\int w(\theta)v(\theta)f(\theta)\ d\theta\right\},$$

subject to:

$$\int w(\theta)v(\theta)f(\theta)d\theta \geq -\mathbb{E}[v(\theta)] + \frac{1-\beta}{\beta}c \quad \text{(IC)},$$

$$\mathbb{E}[\theta|\theta \in \Theta_H^h] \geq \bar{c} \quad \text{(H)}.$$

Since $v(\theta) > 0$ in this region choosing any positive value for $w(\theta)$ for any θ

[28] If F has a hole in its support–so $f(\theta) = 0$ for some interval–then it might be that there is an interval of θ which all satisfy this equation. In this case take $\bar{\theta} = \sup\{\theta' : \mathbb{E}[\theta|\theta \geq \theta'] = \bar{c}\}$.

[29] I've continued to use the notation Θ_H^h rather than $w(\theta)$ in (H) to simplify notation.

will relax (IC) so (IC) will not need to be a binding constraint. Similarly choosing any positive value for $w(\theta)$ for any θ will increase the objective function. For the constraint (H), choosing $w(\theta) = 1$ for all $\theta \in [\bar{c}, 1]$ will violate the constraint. Therefore it will be optimal to choose $w(\theta) = 1$ for all $\theta \in [\bar{c}, 1]$.

(H) must be the binding constraint. So I consider the relaxed problem for $\theta \in [a^L, \bar{c})$ without (IC). This can be written in the form:

$$\sup_{\{x_1(\theta)\}_{\theta=a^L}^{\theta=\bar{c}}} \left\{ \int x_1(\theta)\ d\theta \right\},$$

subject to:

$$\int x_1(\theta) d\theta \leq K_1 \quad \text{(H)};$$

where $K_1 > 0$ is a constant and $x_1(\theta) \equiv \theta f(\theta) w(\theta)$.

It cannot be that $w(\theta) = 1$ (or equivalently $x_1(\theta) = \theta f(\theta)$) for all $\theta \in [a^L, \bar{c}]$ because $\mathbb{E}[\theta | \theta \geq a^L] < \bar{c}$ and so (H) would be violated. So at the maximum it must be that $w(\theta) < 1$ for at least some θ. Due to Assumption 4 the objective function is maximised when $w(\theta) = 1$ (or equivalently $x_1(\theta) = \theta f(\theta)$) for higher θ. So maximum is attained when $w(\theta) = 1$ for θ down to the point (H) is binding which is by definition $\theta = \bar{\theta}$. At this point (IC) is satisfied iff $\int_{\bar{\theta}}^1 h(\theta) dF(\theta) \geq -\mathbb{E}[v(\theta)] + \frac{1-\beta}{\beta} c$.

4.6 Proof of Proposition 5

When $\mathbb{E}[\theta | \theta \geq a^L] < \bar{c}$ the result just follows from Proposition 4.

Now consider what happens when $\mathbb{E}[\theta | \theta \geq a^L] \geq \bar{c}$. Choosing to send the signal $s = h$ only for values of b_1 that are unattainable when $a_1 = a^L$ (i.e. $w(\theta) = 1$ for $\theta \in [a^L, 1]$ and $w(\theta) = 0$ otherwise) will no longer be optimal. The reason is the objective function will be increased if $w(\theta) = 1$ for some $\theta < a^L$ where $v(\theta) > 0$.

The optimal mechanism must have $w(\theta) = 1$ for $\theta \in [a^L, 1]$. Suppose that this was not the case. Since $h(\theta) \geq 0$ for all $\theta \geq a^L$, for θ where $w(\theta) = 0$, choosing $w(\theta) = 1$ will weaken (IC) and increase the objective function. Also since $\mathbb{E}[\theta | \theta \geq a^L] \geq \bar{c}$ (H) will also be weakened. This leads to a contradiction.

The relaxed problem (without (H)) and taking $w(\theta) = 1$ for $\theta \in [a^L, 1]$ as given, can be written as:

$$\sup_{\{w(\theta)\}_{\theta=0}^{\theta=a^L}} \left\{ \int v(\theta)w(\theta)f(\theta)\ d\theta \right\},$$

subject to:

$$\int_0^{a^L} w(\theta)h(\theta)d\theta \geq -\mathbb{E}[v(\theta)] + \frac{1-\beta}{\beta}c - \int_{a^L}^1 v(\theta)f(\theta)d\theta \quad \text{(IC)}.$$

Since $h(\theta)$ is now not necessarily monotonic the same arguments as before cannot be applied.[30] Define $x_2(\theta) \equiv -h(\theta)w(\theta)$. The relaxed problem becomes:

$$\sup_{\{x_2(\theta)\}_{\theta=0}^{\theta=-h(\theta)}} \left\{ \int \frac{v(\theta)f(\theta)}{-h(\theta)} x_2(\theta)\ d\theta \right\},$$

subject to:

$$\int_0^{a^L} x_2(\theta)d\theta \leq K_2 \quad \text{(IC)};$$

where the constant $K_2 \equiv -\left[\mathbb{E}[v(\theta)] + \frac{1-\beta}{\beta}c - \int_{a^L}^1 v(\theta)f(\theta)d\theta\right]$. Note that $-h(\theta) > 0$ for $\theta \in [\beta\bar{c}, a^L]$–i.e. in this region there is always a negative impact on the influence on Self 2's action from the signal sent. This means that while (IC) is slack it is optimal to choose $x_2(\theta) = -h(\theta)$ for θ where $\frac{v(\theta)f(\theta)}{-h(\theta)}$ is maximised. Since

$$\frac{v(\theta)f(\theta)}{-h(\theta)} = \frac{1}{\frac{1}{a^L}\frac{v(\theta/a^L)f(\theta/a^L)}{v(\theta)f(\theta)} - 1}$$

it follows that this is equivalent to choosing θ that minimises

$$\frac{v(\theta/a^L)f(\theta/a^L)}{v(\theta)f(\theta)}. \tag{4.1}$$

When 4.1 is decreasing for all $\theta \in [\beta\bar{c}, a^L]$ it is clear that choosing $x_2(\theta) = -h(\theta)$–or equivalently $w(\theta) = 1$–for the highest values of θ until the IC constraint binds will be optimal.

For $f(\theta) = 1$ this is simply minimised for the highest values of θ. So as before $w(\theta) = 1$ for the highest values of θ until either (IC) and (H) bind. Therefore

[30] For example in the earlier (non-monotonic) example $h(\theta)$ is increasing on $\theta \in [a^L, 1]$ but decreasing on $\theta \in [0, a^L]$.

the optimal mechanism will be a cutoff.

4.7 Proof of Proposition 6

The planner (Self 0) cannot be worse off from the mechanism since it could just design a completely uninformative signal structure and be equally well off compared to the setting in which she cannot provide any information.

Self 2 is rational in a Bayesian sense and myopically maximises her payoff based on the expected value of θ, so providing further information about the state cannot make her worse off.

The optimal choice of mechanism by the planner to induce Self 1 to choose $a_1 = a^H$ instead of $a_1 = a^L$ so Self 1 will be better off than when choosing $a_1 = a^L$. With no information Self 1 and 2 would both choose $a_t = a^L$. So to make Self 1 worse off the mechanism would have to be such that when Self 1 chose $a_1 = a^L$ it would lead to Self 2 choosing $a_2 = a^H$ when Self 1 would not want this (i.e. when $\mathbb{E}[\theta] < \beta\bar{c}$). However to induce Self 2 to choose $a_2 = a^H$ it must be that conditional on the realisation of the signal $\mathbb{E}[\theta] \geq \bar{c}$. So it cannot be that the mechanism would make Self 1 worse off. Therefore Self 1 can always achieve at least the payoff that she would achieve without any mechanism.

5 Comparative statics on the incentive constraints

Here I provide a formal argument for the comparative statics discussed in Section 5.2.

Consider the inequality 4.2 which gives the incentive constraint in the case that there is observable a_1

$$\beta a^H \mathbb{E}[\theta] - c + \beta \int_{\hat{\theta}}^{1} (a^H \theta - c)\ dF(\theta) + \beta \int_{0}^{\hat{\theta}} (a^L \theta)\ dF(\theta) \geq 2\beta a^L \mathbb{E}[\theta]. \quad (5.1)$$

This can be rewritten as

$$a^H \mathbb{E}[\theta] - c/\beta + \int_{\hat{\theta}}^{1} (a^H \theta - c)\ dF(\theta) + \int_{0}^{\hat{\theta}} (a^L \theta)\ dF(\theta) \geq 2a^L \mathbb{E}[\theta]. \quad (5.2)$$

From this it is straightforward to see that if β is increased or c is decreased it leads to an increase in the LHS and no change in the RHS which means the IC constraint is easier to satisfy.

The incentive constraint for unobservable a_1 is similar, and so I omit the formal argument.

6 Use of three signals is without loss

Here I discuss a three player game with the same actions and structure as in my model but with different preferences. This illustrates a setting in which restricting the signal space to two signals is no longer without loss of generality.

The planner and Self 2 have the same preferences as before but Self 1 now has preferences such that he would prefer Self 2 to always take the action $a_2 = a^L$. So the planner and Self 1 have a conflict over what they want Self 2 to do.

Suppose there are three signals $S = \{h, m, l\}$. Using the same notation as in the proof Proposition 2, suppose the preferences and signals were such that $a_2^*(h, a^H) = a^H$ and $a_2^*(m, a^H) = a_2^*(l, a^H) = a^L$, so that following $a_1 = a^H$ only $s = h$ results in Self 2 choosing $a_2 = a^H$; and $a_2^*(h, a^H) = a_2^*(m, a^H) = a^H$ and $a_2^*(l, a^H) = a^L$, so that following $a_1 = a^L$ both $s = m, h$ result in Self 2 choosing $a_2 = a^H$. Suppose also that the signals are such that the IC constraint is satisfied so $a_1 = a^H$ is 'on the equilibrium path'. Now consider what happens if the signals $s = m, l$ are replaced by a single signal $s = l'$. On the equilibrium path this clearly has no effect, but off the equilibrium path this makes a difference. Since following $a_1 = a^L$ it cannot be that Self 2 always chooses $a_2 = a^H$ it must be that $a_2^*(l', a_L) = a^L$. But since Self 1 wants Self 2 to always choose $a_2 = a^L$ this makes the RHS of the IC greater and may lead to it being violated. So, in this example, it will not always be without loss to reduce the signals from three down to two.

Chapter 2

Pay Transparency in Organisations—A Static Model

1 Introduction

Private organisations are increasingly committing to pay transparency—making the amount each worker is paid observable to all others within the firm. In a survey of 715 UK businesses, 18% reported an increase in disclosure of pay outcomes between 2015 and 2017, while almost none reported a decrease (CIPD (2017)).[1] There is also anecdotal evidence that pay transparency is being used in technology start-ups. For example, the start-up SumAll has committed to disclose the pay of its workforce internally—the management's rationale for transparency is given in the following quote:

> *When Dane Atkinson started social analytics platform SumAll in 2011, he too was looking for a way to attract and retain talented people. Informed by two decades of experience as a serial entrepreneur, board member, advisor and executive, he was also trying to mitigate several factors that contributed to high turnover of staff at other companies.*[2]

A benefit of committing to transparency is that it gives an employer credibility, enabling her to demonstrate that she is treating everyone in the same

[1] I focus on the impact of internal disclosure. Some organisations have even started disclosing pay externally.

[2] www.theguardian.com/business/2015/jul/10/salary-wage-glassdoor-payscale-buffer-sumall. Last accessed 27/08/2019.

way. For example, if a worker sees that not only did he not receive a bonus (or equivalently, a pay rise), but also that none of his peers received a bonus, he will infer that it is more likely that the employer was not able to pay anyone a bonus because funds were not available. On the other hand, a downside to transparency is that it enables workers to compare themselves to others. In particular, when a worker sees that he was paid less than a peer, he will infer that his employer values the peer more than him, meaning he will become discouraged and potentially leave the firm or exert less effort. The main contribution of this chapter is to propose a model to formalise this trade-off and to derive testable predictions.[3] A further contribution is to use the model to provide a rational explanation for empirical findings in the relative pay literature that have thus far been explained by non-standard preferences.

In the model a principal employs two agents. At the start of the game, the principal can commit to make bonuses either transparent or not transparent. After this, the principal privately learns the agents' match qualities with the firm (from now on I refer to this as 'productivity'). The principal knows that the agents will receive outside offers and wants to encourage them to stay at the firm. I assume that more productive agents produce a greater surplus if they stay at the firm, and the additional surplus is shared such that both the principal and the agent enjoy greater benefits. The principal also privately learns the marginal cost of paying bonuses—variation in this cost may be due to a lack of funds or the opportunity cost of investing elsewhere. The principal uses discretionary bonuses to signal to more productive agents that they have good prospects at the firm and that they will benefit from staying.

When deciding about transparency, the principal faces a trade-off. The benefit of transparency comes when both agents do not receive a bonus. Here, transparency means that agents become less pessimistic about their productivity compared to when they receive no bonus under no transparency. This is because they attribute a greater probability to the principal not being able to pay them a bonus. However, transparency comes at a cost. When an agent sees the other agent has received a bonus, but he does not receive a bonus himself, he becomes more pessimistic about his productivity than if he did not see the other agent's bonus. This is because, having seen a bonus being paid to the other agent, it is clear that the principal was able to pay bonuses. Therefore, since she chose not to pay a bonus to the agent, she must have learned that the agent was of low

[3]There are a number of other factors that may affect firms' decisions on pay transparency that I do not analyse. These include discrimination and the gender pay gap (Baker et al. (2019)), and public aversion to high pay (Mas (2017)).

productivity.

For some parameters of the model the principal will prefer transparency, and for others she will prefer no transparency. The comparative statics of the equilibrium payoffs shed light on what features of the environment makes transparency more favourable. The key result is that increasing the difference between the value of retaining high and low productivity agents makes transparency more favourable. This is consistent with transparency being used in technology start-ups where it is likely that there is a lot of heterogeneity in the productivity of workers. The intuition for this result is as follows. Recall, transparency is beneficial when both agents are not paid a bonus. When this is the case, a high productivity agent is less likely to quit compared to when he receives no bonus under no transparency. Thus, when retaining him becomes more valuable for the principal, transparency becomes more beneficial. In contrast, the downside of transparency is that a low productivity agent is more likely to quit, and so decreasing the value of retaining a low productivity agent dampens this negative effect.

A number of empirical papers have studied the behavioural effects of workers seeing the pay of their peers—referred to in the literature as 'relative pay'. Card et al. (2012) conduct an experiment within Californian state universities—where pay is publicly available—varying the salience of this information. Their reduced-form findings show that when workers learn that they are paid below the median compared to other workers in similar roles, they report lower job satisfaction and are more likely to search for another job compared to when they receive no information. In contrast, when workers learn they are paid above the median there are no equivalent positive effects. They suggest that this 'asymmetric response to the information about peer salaries' can be explained by a non-standard utility function which has a component that explicitly takes into account 'feelings arising from relative pay'. Furthermore, they argue that one implication of their finding is that pay transparency will never be beneficial for a private firm. My model provides an alternative explanation for this asymmetric response that does not rely on a non-standard utility function. It also shows that a firm might want to commit to pay transparency even when workers react to relative pay in an asymmetric way. I discuss this in more detail in Section 4.

In the final part of the chapter, I consider some extensions of the static model. In order to derive some further insights, I allow the agents to want to stay at the firm, not only if they think they are a good match with the firm, but also if they think that the firm has a greater ability to pay bonuses. This is because a firm with the ability to pay high bonuses today is more likely to

be able to pay high bonuses in the future. I incorporate this directly into the preferences of agents. In this setup, it is possible that when an agent learns that he is paid less than the other agent he sees this as 'good news'—a very high bonus for others reveals that the firm is able to pay high bonuses. What is now key, is whether the agent prefers to be of high productivity at a low paying firm, or of low productivity at a high paying firm. The comparative statics depend on this preference ordering. For example, if an agent prefers to be a low productive worker at a high paying firm rather than vice versa, my model predicts increasing the difference between the value of retaining high and low productivity agents results in transparency becoming *less* favourable.[4] Other extensions of the static model demonstrate the robustness of the results from the main trade-off. For example, I consider wage increases in place of bonuses and also a richer (continuous) set of possible bonuses—in both cases the comparative statics remain unchanged.

1.1 Related literature

From a theoretical point of view, the model is a multidimensional signalling model where the sender (principal) has multiple dimensions of private information (her own cost of paying bonuses and the agents' productivities) and the signalling (bonuses) is in a single dimension.[5] As is typical in such models, in some instances it is not possible for the receiver (agent) to attribute the signal (bonus) to the type of the sender, leading to a signal extraction problem. Multidimensional signalling models where the signalling takes place in a single dimension have been studied in other contexts. Bénabou and Tirole (2006) study a model in which signalling prosocial behaviour (e.g. giving money to charity) can be attributed to either altruism or a desire to impress others. Frankel and Kartik (2019) study a multidimensional signalling model in which the focus is on comparative statics with respect to the informativeness of the set of equilibria as the 'stakes' of the game change—the stakes can be thought of as a larger audience of receivers.[6] In all existing papers in this literature, the sender wants to induce a higher belief about the *same* dimension of the state for *all* members of the audience. In contrast, in my model different parts of the audience (agents)

[4]For this to be the case there are some additional conditions that need to be satisfied, see Proposition 10 for details.

[5]This is different to multidimensional signalling models in which there are both multiple dimensions of private information and multiple dimensions of signals available, as in, for example, Quinzii and Rochet (1985) and Engers (1987).

[6]Other papers that analyse multidimensional signalling models include Austen-Smith and Fryer (2005), Esteban and Ray (2006) and Bagwell (2007).

are interested in different dimensions of the state—their own productivity and the ability of the principal to pay a bonus (and not the productivity of other agents). A novelty of my model is the trade-off the principal faces when designing the informational environment in which the signalling game, with multiple receivers, takes place. In publicly revealing the signal sent to each agent, she potentially reveals information about the state to other agents, and depending on the state, this might be beneficial or detrimental to her.[7]

There is also some similarity between my model and the signalling model in Kamenica (2008). His model has a seller who is informed of a 'global preference parameter' that affects the preferences of *all* buyers. This means that the buyers make inferences about the value of this parameter from the product line that the seller chooses, including the products that are not designed for buyers of his own type. The global preference parameter plays a similar role to the principal's costs in my model. The main contribution of the paper is to provide a rational explanation for seemingly non-standard choice behaviour from consumers facing different product lines—this application is also somewhat similar to my rational explanation for findings in the relative pay literature.

There are a number of other theoretical models with an informed principal who uses bonuses to signal her private information to an agent. Bénabou and Tirole (2003) and Fuchs (2015) both study settings with a single agent. Having a principal who is informed about an agent's productivity is also related to the literature on subjective (or private) evaluations—see, for example, MacLeod (2003).

A different strand of theoretical models study the effect of controlling the informational environment in contests. Ederer (2010) studies whether a principal should commit to provide feedback in a dynamic contest between two agents. Halac et al. (2017) consider whether a principal should commit to disclose publicly whether a contestant has made a breakthrough or not in an innovation contest. They combine this choice of information disclosure with a choice over the distribution of prizes to find the overall optimal policy that induces the maximum level of innovation.

Cullen and Pakzad-Hurson (2018) study the equilibrium effects of pay transparency within an organisation both theoretically and empirically. However, the trade-off that they analyse is different to mine. In particular, they consider an

[7]In another related paper, Ali and Bénabou (2019) modify Bénabou and Tirole (2006) so that a planner can control the privacy (or conversely the transparency) within a society where agents must choose a costly prosocial action. The planner wants to learn the underlying state and uses transparency in order to affect the benefits of signalling for the agents and, in turn, what information she will learn from the signals they choose.

organisation with many workers with homogeneous productivity. In their model, the effect of increased transparency is that it commits the firm to negotiate more aggressively with workers in future because it does not want to be seen by other workers to pay high wages.

2 Model

In this section I begin by describing the model. Then, I discuss the modelling assumptions and their connection to the literature.

2.1 Set-up

Players. There is a principal (she) and two agents (he), indexed by $i = 1, 2$.

Information. Each agent's productivity (or firm specific match) is given by $\theta_i \in \{H, L\}$. The productivity of each agent is independently drawn with the probability of high productivity given by $\Pr[\theta_i = H] = p_0 \in (0, 1)$. The principal faces uncertainty on the marginal cost of paying bonuses given by λ. This is drawn from $\lambda \in \{1, \lambda_H\}$, where $\lambda_H \in (1, \infty) \cup \{\infty\}$ and is referred to as the 'high cost state'. The prior probability of the 'low cost state' is given by $\Pr[\lambda = 1] = q \in (0, 1)$. Each agent receives an outside option—this is drawn independently from $u_i \sim F[0, 1]$ and it is assumed that F has full support and no atoms. At the start of the game the productivity of each agent, the marginal cost of paying bonuses, and the outside option of each agent is drawn independently and unknown to all players who share a common prior.

Actions and timing.

1. The principal decides on a level of transparency. Denote this decision by $a^P \in \{N, T\}$ where N and T denote no transparency and full transparency.

2. The principal *privately* learns the productivity of the agents (θ_1, θ_2) and the marginal cost of paying a bonus (λ).

3. The principal chooses whether or not to pay each agent a bonus, $b_i \in \mathbb{R}_+$. If $a^P = N$ (no transparency) agent i only learns b_i, while if $a^P = T$ (transparency) agents learn both b_1 and b_2.

4. The agents learn their outside options u_i.[8]

[8]The analyse does not rely on this being privately learned.

5. The agents simultaneously choose whether to stay at the firm or to quit. Denote this decision by $a_i^A \in \{S, Q\}$.

6. The players receive their payoffs that are given below.

Beliefs and strategies. The principal's strategy is to choose a level of transparency $a^P \in \{N, T\}$ in the initial node. She updates her belief once she has private information about (θ_1, θ_2) and λ, and then given the choice of transparency, she chooses a distribution over bonuses

$$\sigma : \{N, T\} \times \{H, L\}^2 \times \{1, \lambda_H\} \to \Delta\left(\mathbb{R}_+^2\right).$$

Agent i updates his belief about his productivity (θ_i) and the principal's costs (λ) following his own bonus b_i, and in the case of $a^P = T$, the other agent's bonus b_j.[9] He then chooses a quitting decision formally given by

$$a_i^A : \mathbb{R}_+ \times [0,1] \to \{S, Q\},$$
$$a_i^A : \mathbb{R}_+^2 \times [0,1] \to \{S, Q\},$$

in the case of no transparency and transparency respectively.[10]

Payoffs. The principal's payoff is given by

$$V = \sum_i -\lambda b_i + \mathbb{1}[a_i^A = S] g_{\theta_i}^P,$$

where g_θ^P is the expected future surplus that the principal will earn from an agent with productivity θ. Assume that $g_H^P > g_L^P > 0$—which means that the principal wants to retain all agents, but prefers to retain agents with high productivity.

Agent i's payoff is given by

$$U_i = b_i + \mathbb{1}[a_i^A = S] g^A(\theta, \lambda) + \mathbb{1}[a_i^A = Q] u_i,$$

where $g^A(\theta, \lambda)$ is the expected future surplus if an agent of productivity θ stays at a firm with cost λ. I assume $g^A(\theta, 1) \geq g^A(\theta, \lambda_H)$ for all θ—meaning that conditional on her productivity, an agent (weakly) prefers to say at a firm that has lower costs today (and is able to pay higher bonuses). I also assume

[9] Note that agent i may also make inferences about θ_j. However, this will never be relevant for his quitting decision.

[10] I assume that in the case of indifference the agent chooses to stay and so the agent will never play a mixed strategy. This will be without loss in equilibrium due to the assumptions on F.

$g^A(H,\lambda) > g^A(L,\lambda)$ for all λ—meaning that conditional on the principal's costs, an agent prefers to stay if he has high productivity. Finally, I assume that $1 > g^A(H,1)$ and $g^A(L,\lambda_H) = 0$.

Equilibrium. The equilibrium concept is perfect Bayesian equilibrium.[11] In equilibrium, upon observing the bonuses, each agent i updates his belief about θ_i and λ given the principal's strategy using Bayes rule. They then best respond given this updated belief and their outside options. The principal chooses a strategy $a^P \in \{N, T\}$ to maximise her expected payoff for the rest of the game given her strategy b_1, b_2 (once she learns θ_1, θ_2 and λ) and the agents' best responses. After learning θ_1, θ_2 and λ, the principal chooses (a distribution over) a pair of bonuses (b_1, b_2) that maximise her expected payoff given her choice of transparency, the beliefs this induces for the agents and the agents' corresponding best responses.

2.2 Discussion of the model

In the model, the principal has a better knowledge of the agents' productivity than the agents have themselves. In many organisational settings—for example, in professional services—the principal is more experienced and can assess an agent's productivity from observing him work. This assumption is similar to the assumption of subjective (or private) evaluations, the relevance of this is discussed in the survey by Prendergast (1999), and a similar assumption has been made in many other recent papers.

Agents' productivities are independent by assumption. In other papers where agents learn about themselves through peers, correlation in agents' types are used to drive results. For example, this is the case in Battaglini et al. (2005) and Halac et al. (2017). Papers that include a contest where types are heterogenous among contestants (such as Ederer (2010)) may not have explicit correlation between peers, but still have the effect that the marginal benefit of effort depends on the type of peers. Although these assumption may be valid in some organisational settings, I have made the independence assumption in order to not obfuscate the informational trade-off I wish to analyse.

There is uncertainty on the marginal cost of paying a bonus and this is privately known by the principal. All firms will have some uncertainty on the opportunity costs of paying bonuses. This will be particularly pertinent in smaller

[11] Note that this does not pin down off-path beliefs and so in each subgame there will possibly be multiple equilibria. I discuss later which equilibria I choose to focus on among those that are possible.

organisations or start-ups that are more likely to be cash constrained or have to allocate resources into new projects (i.e. λ_H could be very large—see Sections 3 and 5.2).[12] In terms of information, more junior employees won't necessarily have good knowledge of how the firm is performing, and consequently what funds are available to pay bonuses. Even if this was public information, the management (principal) will almost certainly hold some private information about possible future investment opportunities that will affect the opportunity cost of paying bonuses today. In a different setting, Li and Matouschek (2013) make similar assumptions on the uncertainty of the marginal cost of paying bonuses.[13]

The parameters g_θ^P and $g^A(\theta, \lambda)$ are the expected future surplus received by the principal and agent when an agent with productivity θ stays at a firm with costs λ. The assumption is that both parties are better off in the future when the agent is more productive which makes sense in any organisational setting. This is a key assumption in my model that allows for a separating equilibrium—which, as will be clear in Section 3, is the equilibrium of interest. In Chapter 3, I analyse a dynamic version of the model that provides micro-foundations for these assumptions.

Agents (weakly) prefer to stay at a firm with lower costs of paying bonuses: $g^A(\theta, \lambda)$ is decreasing in λ. It is reasonable that the costs of paying bonuses are persistent across time and so, regardless of his productivity, an agent would prefer to stay at a firm which has a greater ability to pay bonuses. In Section 3, I consider the case where the agent is indifferent about the principal's costs—this represents the case where there the costs are independent over time; in Section 5.1, I consider the case where the agent does have a strict preference—which corresponds to persistence in the principal's costs.

The assumption $g_L^P > 0$ means that the principal always wants an agent to stay at the firm regardless of his productivity. This makes sense particularly if the cost of hiring a new worker is high. The assumption that $1 > g^A(H, 1)$ means that even if an agent was certain they had high productivity ($\theta_i = H$) and at a firm with low costs ($\lambda = 1$), i.e. the best outcome for the agent, it is still possible that he will quit. The assumption that $g^A(L, \lambda_H) = 0$ is a normalisation and is without loss.

In the model agents do not share information with each other about their

[12] As discussed on p.59 in Bewley (1999), the most common way that firms react to financial distress is to freeze wages or reduce bonuses or raises.

[13] They study relational contracts in which the principal privately learns the cost of paying a bonus to the agent. They motivate uncertainty on the cost of paying workers from a well known case study describing the situation Lincoln's Electric faced following financial difficulty after expanding to foreign markets—Hastings (1999).

pay. This is consistent with evidence from the field—Cullen and Perez-Truglia (2018) find in a large commercial bank that although employees have a high willingness to pay for accurate information about the salary of their peers, they are not able to report them accurately. Obviously, in my model agents would benefit from learning the other agent's bonus. But since they do not benefit from sharing their own bonus, agents do not have an incentive to do this.[14]

Finally, I have assumed that the principal has the ability to commit to full transparency or no transparency about the agent's bonuses. I do not allow the principal to commit to more general mechanisms that may partially reveal the bonus of the other agent, or reveal information about the principal's costs (λ). Commitment can be a difficult assumption to justify, particularly when the outcome is stochastic—as in Kamenica and Gentzkow (2011). However, within an organisation, full transparency or no transparency is a very easy policy to implement and commit to, since if the principal reneges deviations can be detected, and reputational costs will deter such deviations.[15] In contrast, it is more difficult for the principal to commit to a stochastic experiment about the bonus of other agents since it is harder to detect deviations. In the case of disclosing information about costs, it might be possible to commit to disclose information about their balance sheet, however, there are always outside opportunities that affect the opportunity cost of paying bonuses that are always the private information of the management (principal). In addition, even if there is a more general mechanism that is optimal, by considering only (full) transparency and no transparency, my finding that no transparency is sub-optimal in some circumstances is still valid.

3 Two bonus levels: The key trade-off

In this section I analyse a simplified version of the model to illustrate the key trade-off in the principal's transparency decision. I provide sufficient conditions for the equilibrium of interest to be uniquely selected. The key results are the comparative statics on the principal's equilibrium payoffs.

For the rest of the section, I make the following assumption:

Assumption 6. *$\lambda_H = \infty$; $g^A(\theta, \lambda) \equiv g^A_\theta$ for all θ, λ; and $b_i \in \{0, 1\}$.*

[14] If there is a small preference for privacy, agents would have a strict preference for not sharing their bonus.

[15] The assumption of commitment to a disclosure policy in an organisational setting is also made in Jehiel (2015). Here the disclosure is not about the pay of other workers, but about other unknown features of the environment—e.g. the monitoring technology.

The first part means that in the high cost state the principal will never want to pay a bonus.[16] The second part means that when the agents are deciding whether or not to stay at the firm, they value only their own productivity within the firm, and not the ability of the firm to pay bonuses. The final part restricts the set of possible bonuses that the principal can pay. The assumption that the positive level is $b_i = 1$ is without loss.[17]

These assumptions effectively mean that when $\lambda = 1$ the principal has a budget of 2 units for paying bonuses, and when $\lambda = \lambda_H$ the principal has no budget for paying bonuses. More precisely, the model could be adjusted in the following way that would not affect the results. In the description of the model, the costs of the principal (λ) are replaced by a 'bonus pool' $B \in \{0, 2\}$, that is unknown with players sharing a common prior $\Pr[B = 2] = q$. The principal then chooses bonuses (b_1, b_2) such that $b_1 + b_2 \leq B$, and her payoffs are given by

$$V = \sum_i -b_i + \mathbb{1}[a_i^A = S] g_{\theta_i}^P.$$

3.1 Pure strategy separating equilibrium

Following the principal's transparency decision (a^P) there are two subgames. I will refer to these as the *no transparency subgame* and the *transparency subgame* throughout the rest of the chapter. As is typical in signalling games, in each subgame there will potentially be multiple equilibria. I will focus on the *pure strategy separating equilibrium.*

Definition 1. *A **pure strategy separating equilibrium** has the principal pay a bonus $b_i = 1$ if and only if $\theta_i = H$ and $\lambda = 1$.*[18]

In this equilibrium, the principal uses a bonus to signal to an agent that he has high productivity. The agents have beliefs consistent with this strategy—and so they are more likely to stay at the firm following a bonus. Note that there are no off-path actions and so off-path beliefs do not need to be specified. Intuitively, such an equilibrium exists if there is a strong incentive to only pay

[16] All results will continue to hold if λ_H is finite and very large—the intuition is that it is still too costly to ever pay a bonus when $\lambda = \lambda_H$.

[17] The value of the bonus only enters the decision of the principal, and here it only matters as a ratio of g_H^P and g_L^P.

[18] Note that it is not possible to have a pure strategy equilibrium in which $b_i = 1$ iff $\theta_i = L$ in the subgame following $a^P = N$. In the transparency subgame, it is possible to have such an equilibrium but I explicitly rule out such an equilibrium since I don't find it realistic. Also note that as discussed in Section 5.3, using increases in wages in place of bonuses rules out such an equilibrium. For the rest of this subsection I will assume that this equilibrium does not occur—and note that Assumption 7 formally rules out such an equilibrium in Proposition 7.

bonuses to high productivity agents—as will be formalised later, this requires g_H^P to be large and g_L^P to be small.

Now, I describe the beliefs of the agents in this equilibrium. I start by considering the no transparency subgame. Fix the action of the principal to pay a bonus $b_i = 1$ to agent i if and only if $\theta_i = H$ and $\lambda = 1$. The updated beliefs of agent i following realisations of b_i are given by

$$\Pr[\theta_i = H | b_i = 1] = 1,$$

$$\Pr[\theta_i = H | b_i = 0] = \frac{p_0(1-q)}{1-p_0 q}.$$

Define $p_N \equiv \frac{p_0(1-q)}{1-p_0 q}$ as the belief of the agent following no bonus. These beliefs are illustrated in Figure 2.1.

Now, consider the transparency subgame. As before, fix the action of the principal to pay a bonus $b_i = 1$ to agent i if and only if $\theta_i = H$. The updated beliefs of agent i following realisations of b_i and b_j $(j \neq i)$ are given by

$$\Pr[\theta_i = H | b_i = 1] = 1 \text{ for any } b_j,$$

$$\Pr[\theta_i = H | b_i = 0, b_j = 0] = \frac{p_0(1-q)}{1-2p_0 q + p_0^2 q},$$

$$\Pr[\theta_i = H | b_i = 0, b_j = 1] = 0.$$

Define $p_T \equiv \frac{p_0(1-q)}{1-2p_0 q+p_0^2 q}$ as the belief of the agent following no bonus and having observed that the other agent also received no bonus. Notice that $p_T > p_N$. This is because under transparency, when an agent does not receive a bonus and sees the other agent also does not receive a bonus, it is more likely the the principal could not pay bonuses ($\lambda = \lambda_H$) compared to when he did not receive a bonus under no transparency. This difference will be critical when I analyse the principal's optimal choice of transparency. These beliefs are illustrated in Figure 2.2.

In the next subsection, I analyse the best responses of the agents given these equilibrium beliefs. I use these to derive sufficient conditions for the principal's best response to be the separating strategy—this provides sufficient conditions for the equilibrium to exist. Then, I provide sufficient conditions for the equilibrium to be uniquely selected by the intuitive criterion.

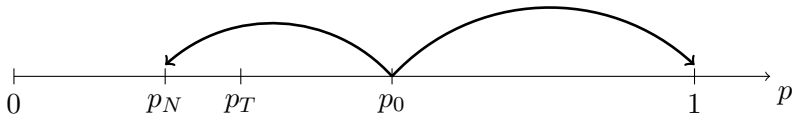

Figure 2.1: **Posterior beliefs of agent i under no transparency**. *Posterior beliefs do not depend on the other agent's bonus.*

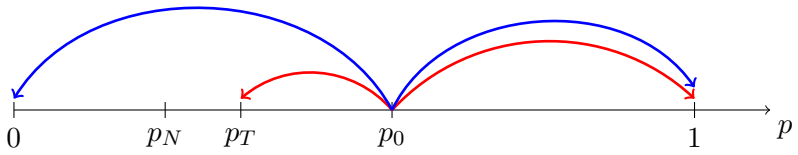

Figure 2.2: **Posterior beliefs of agent i under transparency**. *Posterior beliefs depend on the other agent's bonus: the blue line represents when $b_j = 1$ and the red line represents when $b_j = 0$.*

3.2 Existence and uniqueness of the pure strategy separating equilibrium

It is trivial that when $\lambda = \lambda_H$ the principal has no decision to make. For the rest of Section 3, to simplify notation, whenever I discuss a strategic decision for the principal, I am referring to the case when she has learned that $\lambda = 1$.

Under no transparency, agent i's best response is given by

$$a_i^A = \begin{cases} S & \text{if } \ b_i = 1 \text{ and } u_i \leq g_H^A; \text{ or } b_i = 0 \text{ and } u_i \leq p_N g_H^A, \\ Q & \text{otherwise.} \end{cases}$$

Note that the best response does not depend on the bonus payment of agent $j \neq i$ since agent i does not see this under no transparency—this means that I can consider the incentives for the principal to pay a bonus to each agent separately.

Now, I provide conditions under which this strategy from agent i induces the principal to pay a bonus $b_i = 1$ to agent i if and only if $\theta_i = H$. First, consider the case in which $\theta_i = H$. Paying a bonus is optimal if

$$\begin{aligned} &-1 + \Pr[u_i \leq g_H^A] g_H^P \geq \Pr[u_i \leq p_N g_H^A] g_H^P, \\ \iff &\left(F(g_H^A) - F(p_N g_H^A)\right) g_H^P \geq 1. \end{aligned} \tag{3.1}$$

Second, consider the case in which $\theta_i = L$. Not paying a bonus is optimal if

$$\Pr[u_i \leq p_N g_H^A] g_L^P \geq -1 + \Pr[u_i \leq g_H^A] g_L^P,$$

$$\iff 1 \geq \left(F(g_H^A) - F(p_N g_H^A)\right) g_L^P. \tag{3.2}$$

Notice that 3.1 and 3.2 will both be satisfied for any values of p_0 and q and any distribution F if g_H^P and g_L^P are respectively chosen to be arbitrarily large and small.

Under transparency, agent i's best response is given by

$$a_i^A = \begin{cases} S & \text{if } b_i = 1 \text{ and } u_i \leq g_H^A \text{; or } b_i = b_j = 0 \text{ and } u_i \leq p_T g_H^A \text{;} \\ & \text{or } b_i = 0,\ b_j = 1 \text{ and } u_i \leq 0, \\ Q & \text{otherwise.} \end{cases}$$

Note that, unlike in the no transparency case, this *does* depend on the bonus payment of agent $j \neq i$ since agent i revises his beliefs of λ (and hence his productivity θ_i) based on b_j.

As before, I provide conditions under which this strategy of agent i induces the principal to pay a bonus $b_i = 1$ to agent i if and only if $\theta_i = H$. First, consider the case in which $\theta_i = H$ and $\theta_j = L$ ($j \neq i$). Paying a bonus to agent i is optimal if

$$-1 + \Pr[u_i \leq g_H^A] g_H^P + \Pr[u_j \leq 0] g_L^P \geq \Pr[u_i \leq p_T g_H^A] g_H^P + \Pr[u_j \leq p_T g_H^A] g_L^P,$$

$$\iff \left(F(g_H^A) - F(p_T g_H^A)\right) g_H^P - F(p_T g_H^A) g_L^P \geq 1. \tag{3.3}$$

The expected payoffs from the other agent are also included in this condition since the belief of this agent is affected by the choice of bonus ($b_i = 1$ reveals that it must be that $\lambda = 1$ to agent j). The first inequality is the case in which $\theta_j = L$. The condition would be weaker if $\theta_i = \theta_j = H$ (in this case paying agent i a bonus does not affect the posterior belief of agent j).

Second, consider the case in which $\theta_i = L$ and $\theta_j = H$. Not paying a bonus to agent i is optimal if

$$\Pr[u_j \leq 0] g_L^P \geq -1 + \Pr[u_i \leq g_H^A] g_L^P,$$

$$\iff 1 \geq F(g_H^A) g_L^P. \tag{3.4}$$

As before, the condition would be weaker if $\theta_i = \theta_j = L$.

Notice that 3.3 and 3.4 will again both be satisfied for any values of p_0 and

q and distribution F if g_H^P and g_L^P are respectively chosen to be arbitrarily large and small.

Now, I provide sufficient conditions on the parameters for the pure strategy separating equilibrium to be the unique equilibrium that is selected. In order to discuss mixed strategy equilibria, it will be helpful to introduce some new notation. Recall that given $\lambda = 1$, the strategy of the principal is a mapping from a pair of productivities (θ_1, θ_2) to (a distribution over) a pair of bonuses (b_1, b_2)

$$\{H, L\}^2 \to \Delta\left(\{0, 1\}^2\right).$$

Let

$$\sigma_{\theta_i\theta_j}^{b_i b_j} \equiv \Pr[b_1 = b_i, b_2 = b_j | \theta_1 = \theta_i, \theta_2 = \theta_j]$$

denote the (mixed) strategy of the principal and let σ be the vector of the principal's entire strategy. I restrict attention to *symmetric equilibria* so that each agent is treated in the same way (in expectation) given their productivity.

Definition 2. *A **symmetric equilibrium** restricts σ so that*

$$\sigma_{\theta_i\theta_j}^{b_i b_j} = \sigma_{\theta_j\theta_i}^{b_j b_i},$$

for any realisations of θ_i, θ_j, and choice of b_i and b_j.

Assumption 7. *Restrict attention to symmetric equilibria.*

I also restrict equilibria so that if one agent is of high productivity and the other agent is of low productivity, the principal will never only pay a bonus to the low productivity agent.[19]

Assumption 8. *Restrict equilibria such that in any equilibrium it must be that $\sigma_{HL}^{01} = \sigma_{LH}^{10} = 0$.*

In order to provide the uniqueness result below I also need to make some parameter restrictions. Define $p_H \equiv \frac{1}{2-p_0}$.

Assumption 9. *For all (p_1, p_2) where $\left\{(p_1, p_2) \in [0, 1]^2 : p_1 + p_2 \geq 1\right\}$ it must be that*

$$F(p_1 g_H^A) + F(p_2 g_H^A) > 2/g_H^P + 2F(p_0 g_H^A), \tag{3.5}$$

$$F(p_1 g_H^A) + F(p_2 g_H^A) > -1/g_H^P + 2F(p_H g_H^A). \tag{3.6}$$

[19] As already noted, there exists an equilibrium in pure strategies with the high type receiving no bonus and the low type receiving a bonus which was explicitly ruled out. This assumption is stronger, and obviously rules this equilibrium out.

To demonstrate when this is satisfied, consider the case in which F is uniform—a sufficient condition is now:

$$\frac{2}{1-p_0} < g_H^P g_H^A < \frac{2-p_0}{p_0}. \tag{3.7}$$

To show that parameters can be found to satisfy both these inequalities consider the following. For any p_0, for a given g_H^A, g_H^P can be chosen to be sufficiently large such that the left hand inequality on 3.7 is satisfied. For any g_H^A and g_H^P, p_0 can then be made sufficiently small such that the right hand inequality on 3.7 is satisfied.

Note that this is not a limit result—for fixed g_H^P and g_H^A, any p_0 below a threshold the assumption will be satisfied. When F is not uniform, the inequalities 3.5 and 3.6 become non-linear and so it is not possible to characterise a sufficient condition for the assumption as above. However, the intuition is as before, and the assumption will be satisfied for sufficiently small p_0.

The intuitive criterion is the standard refinement in signalling games to discipline off-path beliefs. However, note that in the transparency subgame the intuitive criterion needs to be adapted since there are multiple 'receivers' (agents). I provide a formal definition of the 'multi-receiver' intuitive criterion adapted for my setting in Appendix 2. It is very much in the spirit of the original definition in Cho and Kreps (1987)—off path actions should only be attributed to 'types' that could possibly want to make this deviation. There are several features of my environment that allow this definition to be adapted in a natural way, these are that:

1. Agent i's payoff does not depend on agent j's action;

2. Agents' productivities are independent;

3. Agent i's payoff does not depend on agent j's productivity θ_j.

I discuss why each of these features is important in Appendix 2.

Proposition 7. *In both subgames, if g_H^P is sufficiently large and g_L^P is sufficiently small, then the pure strategy separating equilibrium exists. Assume that Assumptions 7, 8 and 9 are satisfied, then:*

i) in the no transparency subgame, if g_H^P is sufficiently large and g_L^P is sufficiently small, then the unique equilibrium satisfying the intuitive criterion is the pure strategy separating equilibrium;

ii) in the transparency subgame, if g_H^P is sufficiently large, g_L^P is sufficiently small, then the unique equilibrium satisfying the multi-receiver intuitive criterion is the pure strategy separating equilibrium.

Proofs are omitted from the main part of the text and can be found in the Appendix. Note that this result is not a limit result—it does not require g_H^P and g_L^P to be arbitrarily large and small.

As already discussed, the existence of a pure strategy separating equilibrium requires g_H^P and g_L^P to be sufficiently large and small respectively. When this equilibrium exists, in the no transparency, the intuitive criterion rules out a pooling equilibrium where no bonus is ever paid. The intuition is that when $\theta_i = L$ and g_L^P is sufficiently small, the principal would never want to pay a bonus regardless of the (off-path) belief that this induces. This means that agent i attributes off-path action $b_i = 1$ to $\theta_i = H$, meaning that when $\theta_i = H$ the principal would benefit from a deviation—and so the equilibrium fails the intuitive criterion. In the no transparency subgame there is only one possible mixed strategy equilibrium: the principal mixes—i.e. pays a bonus and no bonus with positive probability in the same state—only when $\theta_i = H$ (and $\lambda = 1$). When g_H^P becomes sufficiently large, this equilibrium no longer exists. Intuitively, this is because the principal will want to retain a high productivity worker so much, so when she can, she will want to pay him a bonus to induce a higher belief. Things are more complex in the transparency subgame since there are many more combinations of mixed strategies to rule out and possible equilibria with off-path actions to refine. However, the intuition is similar, when g_H^P becomes sufficiently large and g_L^P becomes sufficiently small, then 'higher' types no longer want to mix and pool with lower types.[20] The restrictions on the parameters are required in order to apply the intuitive criterion in this subgame. To see why, consider a pooling equilibrium at $b = (0, 0)$. In such an equilibrium, only types (H, H), (H, L) and (L, H) can possibly benefit from a deviation to $b = (1, 1)$—there are no beliefs for which type (L, L) can benefit from this deviation. The parameter restrictions ensure that type (H, H) must *always* benefit from a deviation to $b = (1, 1)$ given any belief that is concentrated on these three types (and not type (L, L))—this means that the equilibrium fails the (multi-receiver) intuitive criterion.

[20] Here 'types' refer to the productivity of the two agents that the principal observes. 'Higher' types means that the productivity of agent 1 and 2 are both weakly higher.

3.3 Optimal choice of transparency for the principal

Now, I return to the economic question of interest—the optimal choice of transparency for the principal. I compare the expected payoff of the principal under no transparency and transparency assuming that in both subgames a pure strategy separating equilibrium exists and is selected. Define $\mathbb{E}V(a^P)$ as the expected value of the principal in the subgame following $a^P \in \{N, T\}$. The following expression captures the difference in payoff between transparency and no transparency:

$$\begin{aligned}\frac{1}{2}D_{TN} &\equiv \frac{1}{2}\left(\mathbb{E}V(T) - \mathbb{E}V(N)\right) \\ &= \left((1-q)(p_0 g_H^P + (1-p_0)g_L^P) + q(1-p_0)^2 g_L^P\right)\left(F(p_T g_H^A) - F(p_N g_H^A)\right) \\ &\quad - qp_0(1-p_0)g_L^P F(p_N g_H^A).\end{aligned} \tag{3.8}$$

The first line of the expression is positive. It represents the benefit of transparency from not discouraging agents as much when they receive no bonus and they see the other agent also received no bonus. The first part (multiplied by $(1-q)$, the probability of the high cost state) is when the principal cannot pay bonuses meaning agents will always be less discouraged under transparency. The second part (multiplied by q) is when the principal can pay a bonus but chooses not to pay either agent a bonus—again agents are less discouraged under transparency. The benefit in both cases is the difference between: the probability of an agent staying under transparency when neither he nor the other agent received a bonus, and the probability of the agent staying under no transparency when he did not receive a bonus.

The second line of the expression is negative. It represents the cost of transparency from when the principal can pay bonuses, but only pays a bonus to one agent causing more discouragement for the other agent who did not receive a bonus under transparency.

Depending on the parameters of the model, either the positive or negative parts of this expression will dominate.

Proposition 8. *Assume that in both subgames the pure strategy separating equilibrium exists and is selected. For some parameters $(p_0, q, g_H^P, g_L^P, g_H^A, F(\cdot))$, no transparency ($a^P = N$) is optimal, and for others transparency ($a^P = T$) is optimal.*

Since the parameter space is very large, it is difficult in general to characterise when transparency is better than no transparency. Transparency is always

optimal when F is uniformly distributed.[21] For no transparency to be optimal it must be that $F(p_T g_H^A)$ and $F(p_N g_H^A)$ are close—minimising the benefits of transparency. This happens, for example, when F is concave.[22]

To understand the implications of the model and the testable predictions it generates, I will consider comparative statics on D_{TN}. In reality, when a firm considers whether or not to commit to transparency, there are other factors to consider beyond my stylised model. So, considering how D_{TN} varies, helps uncover what drives a firm to want to be transparent in the type of environment that fits my model. I consider two comparative statics. The first is what happens when g_H^P, the benefit of retaining a high productivity agent for the principal, goes up, while everything else remains unchanged. The second is what happens when $g_H^P - g_L^P$, the difference between the benefit of retaining a high and low productivity agent, goes up, while keeping the ex ante expected future benefit constant. In effect, the latter considers different mean preserving spreads of retaining agents of different productivities. Define: $\Delta \equiv g_H^P - g_L^P$ and $\bar{g}^P \equiv p_0 g_H^P + (1 - p_0) g_L^P$.

Proposition 9. *Assume in both subgames the pure strategy separating equilibrium exists and is selected, then:*

1. *increasing g_H^P while keeping all other parameters $(g_L^P, g_H^A, p_0, q, F(\cdot))$ constant leads to an increase in D_{TN};*

2. *increasing Δ while keeping all the parameters $g_H^A, p_0, q, F(\cdot)$, and $\bar{g}^P$, constant leads to an increase in D_{TN}, if and only if*

$$F(p_N g_H^A) - (1 - p_0) F(p_T g_H^A) > 0. \tag{3.9}$$

The intuition behind part (1) of the result is as follows. Transparency maximises the probability of retaining workers with high productivity. The reason for this is if the firm is able to pay a bonus (i.e. $\lambda = 1$), then transparency doesn't make a difference for retention of high productivity workers. However, when the firm is not able to pay a bonus (i.e. $\lambda = \lambda_H$), the high productivity workers become less discouraged when there is transparency since they will see that the other agent also has not been paid a bonus. So as retaining high productivity workers becomes a greater priority (i.e. g_H^P goes up), transparency becomes more beneficial.

[21] The formal result—Corollary 2—and proof are in Appendix 1.

[22] For example, no transparency is optimal when $F(x) = x^{1/4}; q = 0.5; p_0 = 0.5, g_H^P = 11; g_L^P = 1.2; g_H^A = 0.4$.

For part (2) of the result, for D_{TN} to be decreasing in Δ it must be the case that F is increasing very quickly between $p_N g_H^A$ and $p_T g_H^A$. The economic interpretation of this condition is that it is very likely that the agent will receive a wage offer in this particular interval. For most distributions this is not the case, for example, for a uniform distribution it must always be the case that D_{TN} is increasing in Δ.[23]

When the principal is not able to pay a bonus ($\lambda = \lambda_H$), a change in Δ while keeping $\bar{g}^P$ constant will have no effect on D_{TN}. The reason for this is that the expected productivity of an agent (given $\lambda = \lambda_H$) remains unchanged and the principal will not pay a bonus regardless of the value of g_H^P and g_L^P. Also, note that when both agents have high productivity and the principal can pay a bonus ($\lambda = 1$), the choice of transparency makes no difference.

Where a change in Δ does make a difference is when the principal is able to pay a bonus ($\lambda = 1$) and when either both agents have low productivity or only one agent has low productivity. In the case when one agent has high productivity and the other low productivity, transparency is detrimental since it discourages the worker with low productivity (it has no effect on the worker with high productivity). So when Δ is increased while keeping $\bar{g}^P$ constant, it means that g_L^P falls and so the negative effects of transparency are reduced (so there is a positive effect on D_{TN}). In the case when both agents have low productivity, transparency is beneficial because both agents will attribute a greater probability to the possibility that the principal could not pay them a bonus (and that they could be of high productivity). So now when Δ is increased while keeping $\bar{g}^P$ constant, it reduces this benefit of transparency. The size of this reduction in benefit is proportional to difference in the likelihood of the agent leaving under transparency and no transparency $F(p_T g_H^A) - F(p_N g_H^A)$. So when this is relatively small, the increase in D_{TN} from when one agent has high productivity and the other low, outweighs the decrease in D_{TN} from when both agents have low productivity.

In terms of testable predictions, part (1) of Proposition 9 suggests transparency is more likely to occur in industries where firms have a very high value for the most productive workers. Part (2) of the result has a similar prediction (so long as the necessary condition 3.9 is satisfied). These features match the motivating examples of technology start-ups where high level of heterogeneity in the productivity of workers or worker-firm matches are likely. They also match

[23]The formal result—Corollary 3—and proof are in Appendix 1. This continues to hold for other distributions as well, for example, the distribution $F(x) = x^{1/4}$ and parameters in footnote 22.

other features of the model such as few verifiable measures of output as well as uncertainty on available funds. To summarise, the reason that transparency is favoured is because the firm can retain high productivity workers more often when the firm is not able to pay them a bonus—this captures the rationale of the entrepreneur in the introduction.

4 Application of the model to empirical work on relative pay: Comparison to results in Card et al. (2012)

As discussed in the introduction, Card et al. (2012) empirically find an 'asymmetric' response when workers learn about their peers' pay. They suggest their finding is due to non-standard preferences—whereas, in my model, there can be a similar asymmetric response, but driven by workers rationally reacting to their informational environment. Furthermore, Card et al. (2012) conclude that the preferences that induce the asymmetric response mean that a private firm would never choose to be transparent about pay. In contrast, my model can have transparency being optimal and an asymmetric response from workers.

Card et al. (2012) assume a relative income model where job satisfaction $S(\cdot, \cdot)$ is given by:

$$S(w, I) = u(w) + v(w - \mathbb{E}[m|I]) + e. \tag{4.1}$$

The arguments are w, the worker's wage, and I, the agent's information set (what agents learn about their peers' wages). The components of the utility function are as follows: $u(\cdot)$ is the utility from his own pay, e is random taste variation, and $v(\cdot)$ are what they call 'feelings arising from relative pay comparisons'. As in my model, they consider two different information sets, the first is I_0, where the worker only sees his own wage, and the second is I_1, where the worker sees the wages of all his peers. The median wage is the reference point and is given by m. It is assumed that when $I = I_0$ the median is given by $m = w$, it is also assumed, without loss, that $v(0) = 0$ (these imply that $v(w - \mathbb{E}[m|I_0]) = 0$ for any w). They test for the concavity of $v(\cdot)$ by assuming a piecewise linear functional form. They find that the slope is decreasing for $w < 0$ and flat for $w \geq 0$. This means that $v(w) < 0$ for $w < 0$ and $v(w) = 0$ for $w \geq 0$, which is what they describe as an 'asymmetric' impact of relative pay. They suggest that these findings are in line preferences that capture 'inequity aversion'—Fehr and Schmidt (1999).

My model, in Section 3, generates a similar 'asymmetric response'.[24] Intuitively, this can be seen by comparing the reaction of agents who are paid and not paid a bonus. First, consider an agent who receives a bonus. The agent is either at or above the median. Regardless, the agent's likelihood of staying at the firm is the same—he learns for sure that he has high productivity. Now consider an agent who does not receive a bonus. The agent is either at or below the median. How he compares to the other agent now matters. If he is at the median—meaning the other agent was also not paid a bonus—he is more optimistic about his productivity, and is therefore more likely to stay. Combining these two reactions generates an asymmetric response.

To formalise this intuition, I calculate $v(\cdot)$ by comparing the expected utility of the agent under transparency and no transparency for different realisations of the state $(\theta_1, \theta_2, \lambda)$. In effect, this is as if data is being generated from exogenous variation in a^P in different states of the world—which is exactly how Card et al. (2012) estimate $v(\cdot)$ in their paper. $S(w, I)$ is the agent's (expected) utility in my model. From 4.1, it follows that

$$S(w, I_1) - S(w, I_0) = v(w - \mathbb{E}[m|I_1]) - v(w - \mathbb{E}[m|I_0]). \tag{4.2}$$

$S(w, I_1) - S(w, I_0)$ is known from the agents' utilities. Given the bonuses, $\mathbb{E}[m|I_1]$ can easily be calculated. $\mathbb{E}[m|I_0]$ can also be calculated—note that, given his own bonus, the agent can rationally compute the expected bonus of the other agent.[25]

First, I calculate the expected median when the agent does not see the bonus of the other agent:

$$\mathbb{E}[m|I_0, b_i = 1] = \frac{1 + \mathbb{E}[b_j|I_0, b_i = 1]}{2} = \frac{1 + p_0}{2};$$

$$\mathbb{E}[m|I_0, b_i = 0] = \frac{\mathbb{E}[b_j|I_0, b_i = 0]}{2} = \frac{qp_0(1 - p_0)}{2(1 - qp_0)}.$$

Now I consider the four cases for the bonuses paid to agent i and j.

Case 1: $b_i = 1$ and $b_j = 1$.

[24] The purpose of this section is to show how similar results are generated within my model, and so there is no need to consider a more complex set up.

[25] This is different to the assumption in Card et al. (2012), where it is assumed that $\mathbb{E}[m|I_0] = m$ for any wage. Clearly in my model, where agents update rationally from their prior in the game played, this does not hold. However, if I was to impose this on agents' beliefs, the model still generates asymmetric updating—the intuition described above does not change.

Under no transparency, agent i believes he is above the median. Whereas, under transparency, he learns that he is paid the median. The likelihood of him remaining at the firm is the same in both cases, and so the expected utility remains unchanged. This means that he does not have additional utility (or 'satisfaction') from being above the median. Formally, $S(1, I_1) - S(1, I_0) = 0$. Now $v(\cdot)$ can be calculated using 4.2

$$\begin{aligned} v(w - \mathbb{E}[m|I_1]) - v(w - \mathbb{E}[m|I_0]) &= 0, \\ v(0) - v((1 - p_0)/2) &= 0. \end{aligned}$$

Assume, without loss, that $v(0) = 0$. This means $v((1 - p_0)/2) = v(0) = 0$.

Case 2: $b_i = 1$ and $b_j = 0$.

Under transparency, agent i learns that he is paid even further above the median, which again does not change the likelihood of him remaining at the firm. A similar calculation to the one above shows $v(1/2) = v((1 - p_0)/2) = 0$.

Case 3: $b_i = 0$ and $b_j = 0$.

Under no transparency, agent i believes he is paid below the median. Whereas, under transparency, agent i learns that he is actually paid the median. This means that[26]

$$\begin{aligned} v(0) - v(-qp_0(1 - p_0)/2(1 - qp_0)) &= S(0, I_1) - S(0, I_0), \\ \implies v(-qp_0(1 - p_0)/2(1 - qp_0)) &= \frac{1}{2}(g_H^A)^2(p_N^2 - p_T^2) < 0. \end{aligned}$$

Case 4: $b_i = 0$ and $b_j = 1$.

Under transparency, agent i learns that he is paid even further below the median, and this decreases the likelihood of him remaining at the firm. A similar calculation to the one above shows

$$v(-1/2) - v(-qp_0(1 - p_0)/2(1 - qp_0)) = -\frac{1}{2}(p_N g_H^A)^2,$$

and so $v(-1/2) = -\frac{1}{2}(p_T g_H^A)^2$.

[26]The calculation has been done with $F(x) = x$ to ease exposition. The final equality holds for any F, so the results for both this case and the later cases continue to hold.

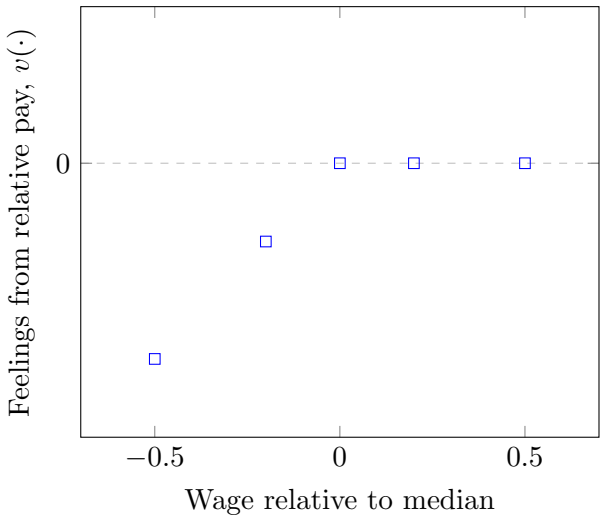

Figure 2.3: *'Feelings arising from relative pay' generated from my model.*

The values of $v(\cdot)$ are plotted in Figure 2.3. So, as in Card et al. (2012) the agent updates 'asymmetrically' in the sense that learning that he received a bonus below the median gives him lower feelings arising from relative pay, while learning that he received a bonus above the median has no effect.

5 Extensions

In this section I return to the static model from Section 3 and consider a number of natural extensions. These demonstrate the robustness of results from this simple model, and also provide some other economic insights.

5.1 Agents having preferences on the principal's costs

In this subsection, I assume when agents make their decision to stay or quit the firm, they not only value their own productivity type (θ_i), but also the ability of the principal to pay a bonus (λ). This is incorporated directly into their preferences. The motivation for this is that a firm with the ability to pay high bonuses today is more likely to be able to pay high bonuses in the future. To highlight the novel effect this has, I allow the principal to pay three different bonuses and to have the ability to pay bonuses in both the low and high cost states. Formally, for the rest of the section, I make the following assumption:

Assumption 10. $\lambda_H \in (1, \infty)$*;* $g^A(\theta, 1) > g^A(\theta, \lambda_H)$ *for all* θ*; and* $b_i \in \{0, 1, 2\}$*.*

There are now four possible productivity/principal cost pairs. To simplify notation, let $\eta = (\theta_i, \lambda)$ denote a generic pair and define $\eta_0 \equiv (L, \lambda_H)$, $\eta_1 \equiv (H, \lambda_H)$, $\eta_2 \equiv (L, 1)$, and $\eta_3 \equiv (H, 1)$. Note that the assumptions I have made, do not specify whether an agent prefers η_1 to η_2 or vice versa. This preference ordering will be important in the results that follow.

I assume that the parameters are such that in both subgames there is an equilibrium where the principal pays:

- $b_i = 0$ when $\eta = \eta_0$;
- $b_i = 1$ when $\eta = \eta_1$ or η_2;
- $b_i = 2$ when $\eta = \eta_3$.

Assumption 10 plays a critical role in the existence of this separating equilibrium. First, by allowing the principal to pay three distinct bonus levels, the only possibility for the agent to be unsure of his type is when he receives a bonus $b_i = 1$, where he is unsure whether he is type η_1 or η_2—this would not be possible with only two bonus levels. Second, if the high costs was not finite, it would not be possible for the principal to pay a bonus to type η_1 and for there to be pooling of types η_1 and η_2.

I omit the conditions in which the equilibrium exists in each respective subgame from the main text since the intuition and derivations are similar to Section 3. However, note that unlike before, this equilibrium may not be unique. Details of the sufficient conditions for existence can be found in Appendix 3.

Now, I formalise the agents' beliefs in the equilibrium described above in order to provide comparative statics. Under no transparency, the posterior probability of agent i being of high productivity given that he receives a bonus $b_i = 1$ is given by

$$\Pr[\theta_i = H | b_i = 1; a^P = N] = \frac{(1-q)p_0}{(1-q)p_0 + q(1-p_0)} \equiv \hat{p}_N.$$

Similarly, under transparency, the posterior probability of agent i being of high productivity given that he receives a bonus $b_i = 1$ and he sees the other agent receives a bonus $b_j = 1$ is given by

$$\Pr[\theta_i = H | b_i = b_j = 1; a^P = T] = \frac{(1-q)p_0^2}{(1-q)p_0^2 + q(1-p_0)^2} \equiv \hat{p}_T.$$

To simplify notation denote $F_k \equiv F(g^A(\eta_k))$, $F_{\hat{p}_N} \equiv F(\hat{p}_N g^A(\eta_1) + (1-\hat{p}_N) g^A(\eta_2))$ and $F_{\hat{p}_T} \equiv F(\hat{p}_T g^A(\eta_1) + (1-\hat{p}_T) g^A(\eta_2))$.

The difference in payoff between the two choices for the principal is

$$\begin{aligned}\frac{1}{2}D'_{TN} &\equiv \frac{1}{2}(\mathbb{E}V(T) - \mathbb{E}V(N)) \\ &= p_0(1-p_0)\left((1-q)(F_1 - F_{\hat{p}_N})g_H^P + q(F_2 - F_{\hat{p}_N})g_L^P\right) \\ &\quad + (F_{\hat{p}_T} - F_{\hat{p}_N})\left(p_0^2(1-q)g_H^P + (1-p_0)^2 q g_L^P\right). \quad (5.1)\end{aligned}$$

The first line of the expression is the difference in payoff when one agent has high productivity and the other has low productivity. There are two cases. The first is when the principal has high costs ($\lambda = \lambda_H$). Here, the difference is in the belief of the high productivity agent (who is paid $b_i = 1$) and transparency reveals to him that he is type $\eta_1 = (H, \lambda_H)$. The second case is when the principal has low costs ($\lambda = 1$). In this case the difference in payoff comes from the change in beliefs of the agent with low productivity. The value of the two cases is given by the first and second parts inside the square brackets. The sign of $F_2 - F_1$, or equivalently whether an agent prefers to be of high productivity in a firm with high costs or of low productivity in a firm with low costs, determines the sign of each component. The two expressions on the first line always go in opposite directions.

On the second line, the expression $(F_{\hat{p}_T} - F_{\hat{p}_N})$ is the difference in the probability that the agent quits when $b_1 = b_2 = 1$ under transparency and no transparency. When agent i is paid $b_i = 1$ and sees the other agent is paid $b_j = 1$, this can be 'good news' or 'bad news'.[27] In particular, the sign of $\hat{p}_T - \hat{p}_N$ depends on p_0: $\hat{p}_T > \hat{p}_N$ if and only if $p_0 > 1/2$. The sign of $F_{\hat{p}_T} - F_{\hat{p}_N}$ depends both on the sign of $\hat{p}_T - \hat{p}_N$ and also the agents' preferences over η_1 and η_2. So this part of the expression is positive if and only if $p_0 > 1/2$ and $\eta_1 \succ \eta_2$ or $p_0 < 1/2$ and $\eta_2 \succ \eta_1$.[28] The intuition is as follows. If it is likely the agent has high productivity ($p_0 > 1/2$) and the agent prefers to be of high productivity at a high cost firm ($\eta_1 \succ \eta_2$), then when he is paid $b_i = 1$ and learns that the other agent is also paid $b_j = 1$ it makes it more likely that he is of high productivity. Since he prefers to be of high productivity at a high cost firm, transparency is beneficial in this case—$F_{\hat{p}_T} > F_{\hat{p}_N}$.

A difference between the predictions of this section comapred to Section 3, is that here when an agent learns that he is paid less than the other agent it is possible that this is good news, while in the previous section this was always bad

[27] Note that this is in contrast to Section 3 where when agent i saw that $b_j = 0$ (and $b_i = 0$) this was always good news since it made it more likely that he was of high productivity. This was because $p_T > p_N$ for all parameter values.

[28] Where $A \succ B$ is a preference relation—meaning the agent prefers A over B.

news. The reason for this difference is as follows. When an agent learns that the other agent is paid a higher bonus, it can increase the likelihood that the firm is able to pay high bonuses, which is good news only when $g^A(\cdot, 1) > g^A(\cdot, \lambda_H)$—as is the case here, and not in Section 3.[29] A number of empirical papers find that workers are more likely to leave a firm after they learn they are paid less than their peers.[30] The model in Section 3 unambiguously makes this prediction, whereas the model in this section may lead to the opposite prediction.

Turning to comparative statics, I again evaluate what happens when the difference in productivity is increased while keeping the ex ante expected productivity unchanged. The results can be derived from this calculation:

$$\frac{1}{2}\left.\frac{\partial D'_{TN}}{\partial \Delta}\right|_{\bar{g}^P} = p_0(1-p_0)\left((1-q)(1-p_0)(F_1 - F_{\hat{p}_N}) - qp_0(F_2 - F_{\hat{p}_N})\right) + (F_{\hat{p}_T} - F_{\hat{p}_N})p_0(1-p_0)(p_0 - q). \quad (5.2)$$

Proposition 10. *Assume in both subgames an equilibrium in which the principal pays agent i $b_i = 0$ if $\eta = \eta_0$, $b_i = 1$ if $\eta = \eta_1$ or η_2, and $b_i = 2$ if $\eta = \eta_3$ exists and is selected.*

1. *If $p_0 = 1/2$, increasing Δ while keeping all the parameters $g_H^A, p_0, q, F(\cdot)$, and $\bar{g}^P$, constant leads to an increase in D'_{TN}, if and only if $g^A(H, \lambda_H) > g^A(L, 1)$;*

2. *If F is uniform, increasing Δ while keeping all the parameters $g_H^A, p_0, q, F(\cdot)$, and $\bar{g}^P$, constant leads to an increase in D'_{TN}, if and only if $g^A(H, \lambda_H) > g^A(L, 1)$;*

3. *If $p_0 > \max\left\{q, \frac{1}{2}\right\}$ and $g^A(H, \lambda_H) > g^A(L, 1)$, increasing Δ while keeping all the parameters $g_H^A, p_0, q, F(\cdot)$, and $\bar{g}^P$, constant leads to an increase in D'_{TN};*

4. *If $q > p_0 > \frac{1}{2}$ or $\frac{1}{2} > p_0 > q$ and $g^A(L, 1) > g^A(H, \lambda_H)$, increasing Δ while keeping all the parameters $g_H^A, p_0, q, F(\cdot)$, and $\bar{g}^P$, constant leads to a decrease in D'_{TN}.*

[29] These two effects combine the effects in the two separate models discussed in Sections 2.1 and 2.2 of Card et al. (2010).

[30] For example, Card et al. (2012) which was discussed earlier in the chapter. Rege and Solli (2014) find evidence that workers do indeed leave a firm if they learn they are paid less than their peers. They study the effects of transparency of pay in all of Norway following the public release of tax returns in 2001. They find that workers who are relatively low earners (compared to their peers) become more likely to quit following the information shock. Similar evidence is found in Cullen and Perez-Truglia (2018) who find that an increase in perceived peer salary makes it more likely that the workers leaves the firm.

The conditions in part (1) and (2) above allow for clear predictions in the direction of the comparative static for any preference ordering over the cross types. When these conditions are not satisfied the comparative static is ambiguous since the first or second part of the expression 5.2 may dominate. However, as in part (3) and (4), additional restrictions can provide sufficient conditions for the comparative static to be in a certain direction for a given preference ordering over the cross types.

Assuming that the conditions in part (1) or (2) are satisfied, the interpretation of the results is as follows. Increasing the value of retaining high productivity agents makes transparency more favourable if and only if an agent prefers to be of high productivity at a high cost firm rather than of low productivity at a low cost firm. These comparative statics line up with those in Section 3. Consider the case where the agent prefers to be of high productivity at the high cost firm compared to of low productivity at the low cost firm. Now, his preferences are in line with what was assumed in Section 3—where the agent only had preferences over his own productivity—and so it makes sense that the comparative statics go in the same direction.

The intuition for part (1) is most informative. Assume that the agent prefers to be of high productivity at the high cost firm (i.e. $\eta_1 \succ \eta_2$). When $b_i = 1$, agent i is unsure whether he is of high or low productivity (η_1 or η_2). With transparency, he also learns the bonus of the other agent which, when $b_j \neq 1$, reveals whether he is of high or low productivity. When the agent learns he is of low productivity (and the principal has low costs), he becomes discouraged compared to when there is no transparency and he did not learn about his productivity (since $F_2 - F_{\hat{p}_N} < 0$). On the other hand, when he is of high productivity (and the principal has high costs), he becomes encouraged (since $F_1 - F_{\hat{p}_N} > 0$). Increasing Δ makes the high productivity agents more valuable for the principal meaning that the encouragement—when the agent learns he is of high productivity—is more valuable. Note that when $p_0 = 1/2$ the difference in agents' beliefs between transparency and no transparency when $b_i = b_j = 1$ is zero. The intuition for the result will continue to hold if $p_0 \approx 1/2$ and this difference is relatively small (i.e. $\hat{p}_T \approx \hat{p}_N$).

5.2 Continuous choice of bonus

Now I consider the setting in Section 3 with Assumption 6, and relax the restriction on the possible bonuses the principal can pay (so $b_i \in \mathbb{R}_+$). With continuous bonuses, transparency will always be preferred to no transparency assuming that in both subgames the pure strategy separating equilibrium that survives the in-

tuitive criterion is selected—this is in effect the separating equilibrium where the principal pays the lowest possible bonuses to ensure separation. The result is driven by the fact that there is no longer a limitation on the actions of the principal (the set of bonuses it can pay). This means that now the incentive constraints bind, while with discrete bonuses they were not binding and they were tighter under transparency compared to no transparency. However, as discussed in the next subsection 5.3, if wage increases are used in place of bonuses it is again possible for both transparency or no transparency to be optimal for the principal, even when wage increases are continuous. With continuous bonuses, comparative statics in line with Proposition 9 can still be derived, demonstrating that the results in Section 3 are robust. I leave the formal analysis and a more detailed discussion to Appendix 4.

Another natural question is what would happen if (θ_i, θ_j) and λ were also continuous? In particular, would it be possible for no transparency and transparency to be optimal for different parameter values? And would the comparative statics continue to yield similar insights? Bénabou and Tirole (2006) analyse a multidimensional signalling model with continuous states and signals within a linear-quadratic-normal framework. However, they make a simplifying assumption in their payoffs—that the benefit of inducing higher beliefs is independent of the state (see footnote 9 in their paper). In contrast in my payoffs, inducing higher beliefs is more beneficial when the state (productivity) is higher—meaning that introducing continuous states with my payoffs would pose technical challenges.

5.3 Wage increase in place of bonuses

Thus far, the principal has paid bonuses to the agents. This enabled me to isolate the signalling effects of bonuses since they do not enter the agents' decision problems. However, in reality, as discussed in the survey Prendergast (1999), the majority of employers don't use bonuses, but instead increase the wage of their employees.

In this section I discuss how the model can be reformulated to have the principal increasing the wage of agents rather than paying bonuses. The difference is that in order to realise an increased wage, the agent must stay at the firm, whereas with a bonus the worker can receive the bonus and still quit the firm—this means the wage increase has an *incentive effect* as well as the signalling effect that has already been seen. Similarly, the principal only incurs the cost if the agent actually stays. For simplicity, I consider the model of Section 5.2 with

$\lambda_H = \infty$ and $g^A(\theta, \lambda) \equiv g^A_\theta$. The payoff of the prinicipal and agent i are now

$$V = \sum_i \mathbb{1}[a^A_i = S](-\lambda w_i + g^P_{\theta_i}),$$
$$U_i = \mathbb{1}[a^A_i = S](g^A_{\theta_i} + w_i) + \mathbb{1}[a^A_i = Q]u_i,$$

where $w_i \in \mathbb{R}_+$ is the wage increase offered to agent i (this takes the place of b_i in the original formulation).[31]

With wage increases in place of bonuses, the analysis changes slightly. In particular, in the no transparency subgame, assuming that the principal increases the wage by $w_i = w$ iff $\theta_i = H$, the best response of agent i is

$$a^A_i = \begin{cases} S & \text{if } \ w_i = w \text{ and } u_i \leq w + g^A_H; \text{ or } w_i = 0 \text{ and } u_i \leq p_N g^A_H, \\ Q & \text{otherwise.} \end{cases}$$

There is a similar change in the transparency subgame. In the no transparency subgame, the incentive compatibility constraints for the principal become

$$\left(F(w + g^A_H) - F(p_N g^A_H)\right) g^P_H \geq F(w + g^A_H) \times w,$$
$$F(w + g^A_H) \times w \geq \left(F(g^A_H) - F(p_N g^A_H)\right) g^P_L,$$

These are analogous to 3.1 and 3.2 in Section 3. There is a similar change in the transparency subgame. As before, if g^P_H and g^P_L are sufficiently high and low, these inequalities will be satisfied. Note that the right hand side of the first inequality and the left hand side of the second inequality, $F(w + g^A_H) \times w$, is strictly increasing in w, meaning that for given parameter values there will be a unique w that makes each inequality bind.

When $w_i \in \{0, 1\}$—analogous to the set up in Section 3—it is straightforward that the comparative static result will remain unchanged. When wage

[31] An interpretation of this reformulated model is as follows. There are discrete time periods $t = 0, 1, 2$, and without loss it is assumed there is no discounting. At time $t = 0$ the two agents start working for the principal and are paid a wage normalised to 0. At the start of period 0 the principal commits to a transparency decision $a^P \in \{N, T\}$. The principal observes the agents working in period $t = 0$ and at the end of the period learns their productivities, θ_i's. The principal then chooses whether or not to increase each agent i's wage for the next period by w_i. At this point the agent also learns his outside option u_i. If the agent chooses to take his outside option he receives the outside option for the following period. If the agent chooses to stay at the firm he gets the increased wage and also the future surplus that depends on his type, θ_i. This future surplus can be thought of as expected future wage increases after period $t = 2$.

increases are continuous ($w_i \in \mathbb{R}_+$) there is added complexity and so it is not always the case—as in Proposition 12 (in Appendix 4)—that transparency always leads to a higher payoff for the principal. As discussed in the previous subsection, with continuous bonuses, the principal can reduce bonuses under transparency and leave the signalling incentives unchanged. In contrast, with continuous wage increases, reducing wages can leave the signalling incentives unchanged but reduce the incentive effects. It is difficult to provide a characterisation of when either transparency or no transparency will be preferred, however there are examples when either will be preferred.[32]

5.4 Correlation in agent's productivity and outside options and agents receiving informative signals

A simplification of the model of Section 3 is that the agents' outside options are not correlated with their productivities. In reality, a productive agent is likely to receive better outside options from other firms. In Appendix 5, I show that the results in Section 3 do not qualitatively change when the outside option has full support for both levels of productivity. This holds regardless of whether or not the productivity and the outside option are positively or negatively correlated.[33] Note that a correlation between the agents' outside option and their productivity is equivalent to agents receiving an informative (but not fully informative) signal about their productivity. This means that if agents were to receive an informative signal about their productivity (before making their stay/quit decision) then the results also remain qualitatively unchanged.

5.5 No bonus as perfect bad news

In the pure strategy separating equilibrium of Section 3, receiving a bonus is 'perfect good news'—it reveals to the agent that he is certainly of high productivity. In contrast, in the case of no transparency, receiving no bonus leaves the agent uncertain whether or not he is of high productivity. There are parameters for which it becomes possible to have an equilibrium where no bonus acts as a 'perfect bad news' signal. More specifically, the pure strategy separating equilibrium is such that a bonus is always paid to the agent, unless he is of low productivity and the firm has high costs. The key finding is that the comparative

[32]For parameters $\{F(x) = x; g_H^A = .5; g_H^P = .4; g_L^P = .1; p_0 = .4; q = .4\}$, the wage increase under no transparency is 0.0751623, and under transparency are 0.0834506 and 0.0402955 when both or only one agent receives the increase. The expected difference in payoff between no transparency and transparency is 0.00301823.

[33]Clearly positive correlation is the more realistic case.

static result of Section 3 remain unchanged—increasing how much the principal values retaining high productivity agents makes transparency more favourable. I leave the formal analysis and a more detailed discussion to Appendix 6.

6 Discussion and conclusion

6.1 The role of commitment

As already argued, commitment to transparency (or no transparency) of pay is plausible in many organisational settings. However, to illustrate the importance of the commitment assumption in the model, I consider what happens if the principal cannot commit to transparency. The result is reminiscent of the unravelling result in Milgrom (1981): the principal is forced to be transparent in all cases since not being transparent means that the principal is choosing to hide bad news. In this case 'bad news' is that there are bonuses available and that the principal chose not to pay one of the agents a bonus.

To relax commitment, I consider the set up in Section 3 but where the timing is changed so that the principal chooses the level of transparency after learning (θ_1, θ_2) and λ.[34] The incentive of the principal to signal to each agent is unchanged, so paying a bonus ($b_i = 1$) is only worthwhile if the agent is of high productivity ($\theta_i = H$) and the bonus signals this to them. I also assume that following any choice of a^P the principal chooses $b_i = 1$ if and only if $\theta_i = H$, and so as before, the pure strategy separating equilibrium is selected. Finally, I assume in the case of indifference the principal chooses $a^P = T$ over $a^P = N$.[35]

Proposition 11. *In the game without commitment, if the principal plays a pure strategy separating equilibrium in every subgame following a^P, then in any equilibrium the principal chooses transparency ($a^P = T$) for any realisation of θ_1, θ_2 and λ with probability 1.*

6.2 Concluding remarks

In this chapter, I propose a theory that allows me to analyse what features of a firm make pay transparency more favourable. The key result—which provides a clear, testable prediction—is that increasing the difference between the value of retaining high and low productivity workers makes transparency more favourable.

[34] Formally (1) and (2) in the 'Actions and timing section' are switched.

[35] This rules out an equilibrium in which the principal always plays $a^P = T$ except when $\lambda = 1, \theta_1 = \theta_2 = H$.

Throughout the chapter it has been assumed that the principal can only commit to two extreme information structures—transparency or no transparency—on the information agents get about the bonus of the other agent. I also do not allow the principal to commit to disclose anything about her costs (λ). As already discussed, I believe that these assumptions are reasonable based on what firms are actually able to commit to in reality. Nonetheless, from a theoretical point of view, it may be interesting to investigate more general mechanisms and to characterise the optimal mechanism in future work. Such mechanisms could allow for 'experiments' that disclose a signal about the other agent's bonus as in Kamenica and Gentzkow (2011); they can also allow for an experiment that discloses a signal about the principal's costs. Intuitively, consider the choice the principal would want to make about committing to disclose information about her costs. If the value of retaining high productivity workers (g_H^P) is sufficiently high, she would want to commit to fully disclose her costs. The reason is that, in doing so, the high productivity agents never become pessimistic—either they learn they are high productivity when a bonus can be paid, or they learn nothing when a bonus cannot be paid—and when g_H^P is high, maximising the belief of high productivity agents is the principal's priority. There are also other questions of interest that could be answered by allowing for a more general class of mechanisms for disclosing information about bonuses. For example, if from an ex ante point of view, one of the two agents was more likely to have high productivity, should the principal commit to disclose more or less information about the realised bonus of the 'good' or the 'bad' agent?

The model has also abstracted from how a decision to commit to transparency affects the composition of workers within a firm. In a well known empirical study, Lazear (2000) finds that performance pay has a significant effect on the sorting into a firm—the introduction of performance pay means that the firm's workforce becomes more productive. A natural (theoretical) question is: what effect does transparency have on the sorting of workers into firms? The theoretical model in Cullen and Pakzad-Hurson (2018) addresses this question within a simple framework with homogeneous workers, but it is not clear what will happen with heterogeneous workers within the signalling framework I have developed.

Appendix to Chapter 2

1 Proofs

1.1 Proof of Proposition 7

Existence simply follows from inequalities 3.1, 3.2, 3.3, 3.4 and the discussion following them.

For uniqueness, I start by considering (the simpler case) of no transparency. First, using the intuitive criterion, I rule out equilibria where there is an off-path choice of bonus. Then, I show that there can only be a single mixed strategy equilibrium, and when g_H^P is sufficiently large and g_L^P is sufficiently small this does cannot exist. In the case of transparency, however, there are many types of equilibria to rule out. I start this section of the proof by outlining the key steps to ruling out other possible equilibria and then fill in the details.

No Transparency.

Consider each agent i separately (this can be done since the payoffs of the principal is separable across agents). There are two types of equilibria where either $b_i = 0$ or $b_i = 1$ are played with zero probability when $\lambda = 1$.

First, there cannot be an equilibrium with pooling on $b_i = 1$ since type[36] $\theta_i = L$ will always have an incentive to deviate to $b_i = 0$. The agent's beliefs following $b_i = 0$ is p_0 (note that this belief if pinned down since it is 'on-path' when $\lambda = \lambda_H$) and means such a deviation benefits the principal.

Second, the intuitive criterion can rule out an equilibrium with pooling on $b_i = 0$. To support such an equilibrium, it must be that following the off-path action $b_i = 1$ agent i's belief of his productivity is sufficiently pessimistic that neither type, in particular type $\theta_i = H$, will have an incentive to deviate.[37] The principal's (expected) equilibrium payoff for type θ_i is

[36] In this context 'type' is the productivity of agent i that the principal has observed. Also recall that when I discuss a strategic decision by the principal it is assumed that $\lambda = 1$.

[37] I use the textbook notation in Fudenberg and Tirole (1991).

$$V^*(\theta_i) = F(p_0 g_H^A) g_{\theta_i}^P. \tag{1.1}$$

Now consider the maximum possible payoff for the principal following a deviation to $b_i = 1$ and a best response from the agent given any beliefs that the agent can have. For each type, the maximum payoff is obtained when the agent is as optimistic as possible—i.e. when agent i's belief is that following $b_i = 1$ he has productivity $\theta_i = H$ with probability one. In this case, the (expected) payoff for type θ_i is from a deviation that induces the most optimistic belief is

$$\mathbb{E}\left[V(\theta_i)\right] = -1 + F(g_H^A) g_{\theta_i}^P. \tag{1.2}$$

From 3.2 type $\theta_i = L$ does strictly worse from this deviation (because $p_N < p_0$). If g_H^P is sufficiently high, type $\theta_i = H$ does benefit from this deviation (because $p_0 < 1$). This means that following a deviation to $b_i = 1$ agent i assigns probability 0 to type $\theta_i = L$. Such an equilibrium will fail the intuitive criterion since—given that a deviation must be by type $\theta_i = H$—type $\theta_i = H$ gets a strictly higher payoff from deviating.

I now rule out all possible mixed strategy equilibria—note that such equilibria never have off-path actions and so the intuitive criterion has no bite.[38] I begin with the following Lemma.

Lemma 5. *There cannot be an equilibrium where $b_i = 1$ is played with positive probability when $\theta_i = L$ and $b_i = 0$ is played with positive probability when $\theta_i = H$.*

Proof. This can be shown by contradiction. Suppose such an equilibrium existed, it would be the case that the following two inequalities must be satisfied

$$\begin{aligned} F(\bar{p}_0 g_H^A) g_H^P &\geq -1 + F(\bar{p}_1 g_H^A) g_H^P, \\ -1 + F(\bar{p}_1 g_H^A) g_L^P &\geq F(\bar{p}_0 g_H^A) g_L^P, \end{aligned}$$

where $\bar{p}_0 = \Pr[\theta_i = H | b_i = 0; \sigma]$ and $\bar{p}_1 = \Pr[\theta_i = H | b_i = 1; \sigma]$ and σ is the principal's strategy in the candidate equilibrium. Rearranging these inequalities gives

$$1/g_H^P \geq F(\bar{p}_1 g_H^A) - F(\bar{p}_0 g_H^A) \geq 1/g_L^P,$$

which leads to a contradiction since $g_H^P > g_L^P > 0$. □

[38] Also note that any possible mixed strategy equilibria must have the same mixing for both types under the Assumption 7.

Next, I show that when conditions 3.1 and 3.2 are satisfied—so that a pure strategy separating equilibrium exists—I show there is at most one mixed strategy equilibrium. In this equilibrium only one type mixes and so it can be described as a 'partially separating equilibrium'. Then, I provide a condition such that the pure strategy separating equilibrium still exists and the mixed strategy equilibrium does not exist.

Lemma 6. *When 3.1 and 3.2 are satisfied there is at most one mixed strategy equilibrium. In this equilibrium the principal plays $b_i = 0$ with probability 1 when $\theta_i = L$ and $\lambda = 1$ and mixes between $b_i = 0$ and $b_i = 1$ when $\theta_i = H$ and $\lambda = 1$. If*

$$\left(F(g_H^A) - F(p_0 g_H^A)\right) g_H^P > 1,$$

this equilibrium does not exist.

Intuitively, this final condition is required for this result because if the payoff from inducing a high posterior (g_H^P) is sufficiently high, it cannot be the case that the principal would ever be indifferent between paying a bonus (and inducing a high posterior) and not paying a bonus (and inducing a lower posterior).

Proof. From Lemma 5 there cannot be an equilibrium in which both types both mix over $b_i = 0$ and $b_i = 1$. First consider if type L mixes and type H plays $b_i = 1$ with probability 1. Type L's indifference condition is given by

$$F(\bar{p}_1 g_H^A) - F(\bar{p}_0 g_H^A) = \frac{1}{g_L^P},$$

where $\bar{p}_1$ and $\bar{p}_0$ are defined as before. Since $\bar{p}_1 < 1$ and $\bar{p}_0 > p_N$ it follows that 3.2, the incentive constraint that ensures that the low type does not want to pay a bonus, is violated.

So the only possible mixed strategy equilibria have type H mixing and type L playing $b_i = 0$ with probability 1. In such an equilibrium type H's indifference condition is given by

$$F(\bar{p}_1 g_H^A) - F(\bar{p}_0 g_H^A) = \frac{1}{g_H^P}. \tag{1.3}$$

Note that in such an equilibrium it is the case that $\bar{p}_1 = 1$. Furthermore, note that 3.1 and 3.2 can also be satisfied under such an equilibrium. The equilibrium is pinned down by $\bar{p}_0$. Inverting 1.3 gives

$$\bar{p}_0 = \frac{1}{g_H^A} F^{-1}\left[\frac{1}{g_H^P} + F(g_H^A)\right]. \tag{1.4}$$

Calculating $\bar{p}_0$ by Bayes rule gives

$$\bar{p}_0 = \frac{p_0(1 - q\sigma_H)}{1 - qp_0\sigma_H}, \tag{1.5}$$

where $\sigma_H \equiv \Pr[b_i|\theta_i = H; \sigma]$. Since 1.5 decreases in σ_H when $p_0 \in (0, 1)$, it must be that there is a unique σ_H that solves 1.4 and 1.5. It follows that there is a unique mixed strategy equilibrium.

Furthermore since $p_0 > \bar{p}_0$ when $\sigma_H \in (0, 1)$,

$$F(g_H^A) - F(p_0 g_H^A) < F(\bar{p}_1 g_H^A) - F(\bar{p}_0 g_H^A) = \frac{1}{g_H^P},$$

where the equality follows from 1.3. Therefore if g_H^P is sufficiently large and

$$F(g_H^A) - F(p_0 g_H^A) > \frac{1}{g_H^P},$$

this leads to a contradiction—so a mixed strategy equilibrium does not exist. Note that this condition can be consistent with 3.1 and 3.2—since it essentially requires g_H^P to be sufficiently large—and so a pure strategy separating equilibrium still exists. □

Transparency.

Recall the notation introduced in the main text to denote mixed strategies in the transparency subgame:

$$\sigma_{\theta_i\theta_j}^{b_i b_j} \equiv \Pr[b_1 = b_i, b_2 = b_j | \theta_1 = \theta_i, \theta_2 = \theta_j].$$

There are many possible strategies in the case of transparency. In order to make the exposition more clear, I graphically illustrate the strategies that I consider for the principal (when $\lambda = 1$) as in Figure 2.4.

The proof here will rule out all possible 'diagonal' mappings on the diagram above such as the bold line in Figure 2.4 from (H, H) to $(1, 0)$ which would represent σ_{HH}^{10}. There are two possible types of strategies that must be ruled out in any equilibria. The first is where there is a bonus pair that is not played (when $\lambda = 1$); the second is where there is mixing by one type from the principal.[39] The first possibility occurs in the mapping in Figure 2.4; the second occurs in the mapping in Figure 2.5—where types (H, H) and (L, L) mix over $(1, 1)$ and $(0, 0)$ as depicted by the bold lines. Before making the formal arguments, I

[39] In this context 'type' refers to the pair of productivities for the agents when $\lambda = 1$. I will also write the strategy of the principal as a pair (b_i, b_j) to simplify notation.

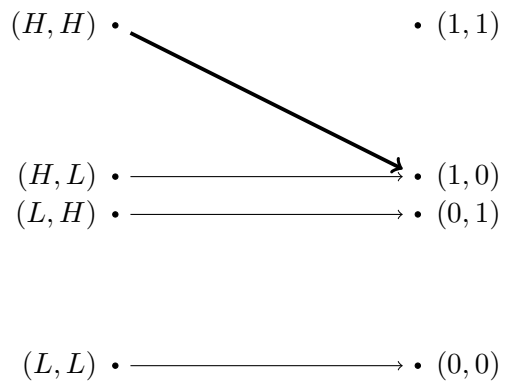

Figure 2.4: ***Example of a strategy resulting in an off-path action:*** *On the left hand side is the productivity of the agents* (θ_1, θ_2)*—or 'type' of the principal; on the right hand side is the bonus* (b_1, b_2) *that the principal pays. The bonus pair* $(1,1)$ *is never played with this set of strategies.*

outline the steps in the proof, then I provide the details. Note that each step builds on previous steps, and throughout I assume that I can make g_H^P and g_L^P sufficiently large and small as required, and when other parametric restrictions in Assumption 9 are required, I note that this is the case.

Step 1: Rule out equilibria where types (H,H) and (L,L) 'cross over', so it is not possible to have an equilibrium with both $\sigma_{LL}^{11} > 0$ and $\sigma_{HH}^{00} > 0$ (as, for example, in Figure 2.5). 'Cross over' refers to the diagonals crossing each other as in Figure 2.5.

Step 2: Rule out equilibria where two other types 'cross over', e.g. it is not possible for both $\sigma_{HL}^{11} > 0$ and $\sigma_{HH}^{10} > 0$ to be the case in an equilibrium.

Step 3: Show that in any equilibrium it must be that $\sigma_{LL}^{00} = 1$.

Step 4: Rule out equilibria where $\sigma_{\theta_i\theta_j}^{11} = 0$ for all (θ_i, θ_j) (as, for example, in Figure 2.4) with the intuitive criterion. Note that this requires the restrictions on parameters from Assumption 9. Combining Steps 1, 2 and 4 means that it must be that $\sigma_{HH}^{11} > 0$.

Step 5: Show that in an equilibrium where $\sigma_{HH}^{11} > 0$ it must be that $\sigma_{HH}^{11} = 1$.

Step 6: Show that in an equilibrium where $\sigma_{HH}^{11} = 1$ and $\sigma_{LL}^{00} = 1$ it must be that $\sigma_{HL}^{10} > 0$. This step requires ruling out equilibria where $\sigma_{\theta_i\theta_j}^{10} = 0$ for all (θ_i, θ_j) with the intuitive criterion. Note that this requires the restrictions on parameters from Assumption 9.

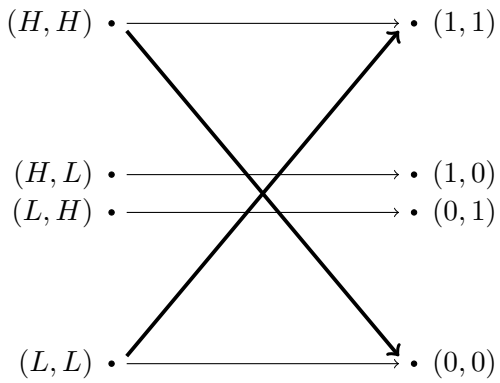

Figure 2.5: ***Example of a strategy with mixing:*** *Types* (H, H) *and* (L, L) *mix by paying bonuses* $(1, 1)$, $(0, 0)$ *with positive probability.*

Step 7: Show that in an equilibrium where $\sigma_{HH}^{11} = 1$, $\sigma_{LL}^{00} = 1$ and $\sigma_{HL}^{10} > 0$ it must be that $\sigma_{HL}^{11} = 0$.

Step 8: Show that in an equilibrium where $\sigma_{HH}^{11} = 1$, $\sigma_{LL}^{00} = 1$ and $\sigma_{HL}^{10} > 0$ it must be that $\sigma_{HL}^{00} = 0$. Combining Steps 6, 7 and 8 means that it must be that $\sigma_{HL}^{10} = \sigma_{LH}^{01} = 1$.

Now I turn to the formal argument to fill in the details on the steps above. Define the posterior probabilities of agent i, given the principal's strategy σ, following bonuses b_i and b_j as

$$\bar{p}_{b_i b_j} \equiv \Pr[\theta_i = H | b_i, b_j; \sigma].$$

Step 1.

A similar result to Lemma 5 can be derived—this rules out equilibria with 'cross overs' described above and illustrated in Figure 2.5.

Lemma 7. *There is no equilibrium with* $\sigma_{HH}^{00} > 0$ *and* $\sigma_{LL}^{11} > 0$.

Proof. If $\sigma_{HH}^{00} > 0$, type (H, H) must (weakly) prefer to pay bonuses $(0, 0)$ to $(1, 1)$. Similarly, if $\sigma_{LL}^{11} > 0$, type (L, L) must (weakly) prefer to pay bonuses $(1, 1)$ to $(0, 0)$. These inequalities are

$$\begin{aligned} 2F(\bar{p}_{00} g_H^A) g_H^P &\geq -2 + 2F(\bar{p}_{11} g_H^A) g_H^P, \\ -2 + 2F(\bar{p}_{11} g_H^A) g_L^P &\geq 2F(\bar{p}_{00} g_H^A) g_L^P. \end{aligned}$$

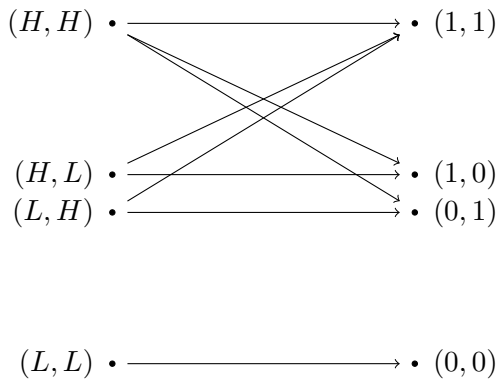

Figure 2.6: *Mixing by both type* (H,H) *and types* (H,L) *and* (L,H).

Rearranging and combining these inequalities gives

$$1/g_H^P \geq F(\bar{p}_{11}g_H^A) - F(\bar{p}_{00}g_H^A) \geq 1/g_L^P.$$

Since $g_H^P > g_L^P > 0$ this leads to a contradiction. □

Step 2.

Now, I show that mixing across the other types, as, for example, in Figure 2.6, can also be ruled out. I start with a Lemma similar to Lemma 7.

Lemma 8. *There is no equilibrium with both* $\sigma_{HH}^{10} = \sigma_{HH}^{01} > 0$ *and* $\sigma_{HL}^{11} = \sigma_{LH}^{11} > 0$*; there is no equilibrium with both* $\sigma_{HL}^{00} = \sigma_{LH}^{00} > 0$ *and* $\sigma_{LL}^{10} = \sigma_{LL}^{01} > 0$.

The proof is very similar to the proof of Lemma 7 and is omitted.

Step 3.

Now I show that in any equilibrium it must be $\sigma_{LL}^{00} = 1$. The intuition for this is that if g_L^P is sufficiently small there is no possible benefit that makes incurring the cost of a bonus worthwhile for type (L,L).

Lemma 9. *In any equilibrium it must be that* $\sigma_{LL}^{00} = 1$.

Proof. This is shown by contradiction. Suppose that there was an equilibrium with $\sigma_{LL}^{00} < 1$ and in this equilibrium type (L,L) plays $(b_1', b_2') \neq (0,0)$ with positive probability and this action induces a posterior belief for the two agents of $(p_1', p_2') \in [0,1)^2$. Also suppose that choosing $b = (0,0)$ induces belief $(\bar{p}_{00}, \bar{p}_{00})$—note that this is pinned down since even if $\sigma_{LL}^{00} = 0$ is not off-path. It must be

that (b_1', b_2') is weakly preferred to $b = (0,0)$ by type (L, L)

$$-(b_1' + b_2') + F(p_1' g_H^A) g_L^P + F(p_2' g_H^A) g_L^P \geq F(\bar{p}_{00} g_H^A) g_L^P + F(\bar{p}_{00} g_H^A) g_L^P.$$

Rearranging this gives

$$F(p_1' g_H^A) - F(\bar{p}_{00} g_H^A) + F(p_2' g_H^A) - F(\bar{p}_{00} g_H^A) \geq (b_1' + b_2')/g_L^P.$$

Since $F(p_1' g_H^A) - F(\bar{p}_{00} g_H^A) + F(p_2' g_H^A) - F(\bar{p}_{00} g_H^A) \leq 2$, for sufficiently small g_L^P this leads to a contradiction for any p_1', p_2'. □

Step 4.

Now I use the multi-receiver intuitive criterion refinement to rule out equilibria where $\sigma^{11}_{\theta_i \theta_j} = 0$ for all (θ_i, θ_j).

Lemma 10. *Under the multi-receiver intuitive criterion refinement there cannot be an equilibrium with $\sigma^{11}_{\theta_i \theta_j} = 0$ for all (θ_i, θ_j).*

Proof. In order to apply the intuitive criterion refinement, the equilibrium payoffs of each type of principal need to be determined. However, there are multiple equilibria in which $\sigma^{11}_{\theta_i \theta_j} = 0$ for all (θ_i, θ_j). Since in any equilibrium it must be that $\sigma^{00}_{LL} = 1$, only the actions of types (H, H), (H, L) and (L, H) need to be determined. By Lemma 8, there are only three possible equilibria:

1. $\sigma^{00}_{LL} = \sigma^{00}_{LH} = \sigma^{00}_{HL} = 1$ and $\sigma^{10}_{HH} = \sigma^{01}_{HH} = 1/2$;
2. $\sigma^{10}_{HH} = \sigma^{01}_{HH} = 1/2$, $\sigma^{00}_{LL} = 1$ and $\sigma^{10}_{HL} = \sigma^{01}_{LH} \in (0, 1]$;
3. $\sigma^{00}_{\theta_i \theta_j} = 1$ for all (θ_i, θ_j).

(1) cannot be an equilibrium for large g_H^P since type (H, L) and (L, H) will have an incentive to deviate to $(1, 0)$ or $(0, 1)$ and induce a belief of $(1, 1)$.

The intuitive criterion can rule out putative equilibrium (2). The argument is as follows. The equilibrium payoffs of the types are:[40]

$$\begin{aligned} V^*(H, H) &= -1 + \left(F(\bar{p}_{10} g_H^A) + F(\bar{p}_{01} g_H^A)\right) g_H^P, \\ V^*(H, L) &= -1 + F(\bar{p}_{10} g_H^A) g_H^P + F(\bar{p}_{10} g_H^A) g_L^P, \\ V^*(L, L) &= 2F(\bar{p}_{00} g_H^A) g_L^P. \end{aligned}$$

However, note that the strategies above imply that $\bar{p}_{10} = 1$ while $\bar{p}_{01}, \bar{p}_{00} < 1$. Consider a deviation to the off-path action $b = (1, 1)$. The most favourable

[40] Note I omit type (L, H) since the analysis is identical to that of type (H, L).

belief that this could induce a principal of any type is $(1,1)$. However, if g_L^P is sufficiently small then neither types (L,L) nor (H,L) will benefit from this deviation. If g_H^P is sufficiently large then type (H,H) will benefit, and thus be the only type that does not get deleted at the first round of iterations. Now with only type (H,H) remaining, the deviation to $b=(1,1)$ will clearly give this type a greater payoff than her equilibrium payoff and thus this equilibrium will fail the (multi-receiver) intuitive criterion.

An equilibrium (3) can exist with off-path beliefs following $b \neq (0,0)$ that are the same as the on-path belief p_0 for both agents. The intuitive criterion can rule out such an equilibrium under Assumption 9. In such an equilibrium, the equilibrium payoffs of the types are:

$$\begin{aligned} V^*(H,H) &= 2F(p_0 g_H^A) g_H^P, \\ V^*(H,L) &= F(p_0 g_H^A)(g_H^P + g_L^P), \\ V^*(L,L) &= 2F(p_0 g_H^A) g_L^P. \end{aligned}$$

Consider a deviation to the off-path action $b=(1,1)$. Clearly if g_H^P is sufficiently large and g_L^P is sufficiently small then under the most favourable belief, $(1,1)$, type (H,H), (H,L) and (L,H) all benefit from such a deviation, and type (L,L) does not benefit from such a deviation. This means that the deviation cannot be made by type (L,L) and this type is eliminated.

So the remaining types are (H,H), (H,L) and (L,H) and so the possible beliefs that can be induced by a deviation to $b=(1,1)$ are (p_1,p_2) where $\left\{(p_1,p_2) \in [0,1]^2 : p_1 + p_2 \geq 1\right\}$. In order to show that the equilibrium fails the intuitive criterion, I show that type (H,H) always benefits from this deviation for any belief that this induces within the set above. More formally, following the definition in Appendix 2, for $b=(1,1)$, $J(b)=(L,L)$ and $\theta'=(H,H)$, I show that

$$V^*(\theta') < \min_{(\bar{u}_1^A, \bar{u}_2^A) \in (\mathrm{BR}_1(\Theta \setminus J(b), b), \mathrm{BR}_2(\Theta \setminus J(b), b))} V(b, \bar{u}_1^A, \bar{u}_2^A, \theta').$$

Note that the best responses of the agents correspond directly to the beliefs that are induced by deviation. The inequality above can be verified by checking that the following holds for all beliefs (p_1,p_2)

$$-2 + \left(F(p_1 g_H^A) + F(p_2 g_H^A)\right) g_H^P > 2F(p_0 g_H^A) g_H^P.$$

Rearranging this give the first inequality in Assumption 9, and so this is satisfied

by assumption.

□

By Lemma 10 it must be that $\sigma^{11}_{\theta_i\theta_j} > 0$ for some (θ_i, θ_j), and by Lemmas 7 and 8, it cannot be the case that $\sigma^{11}_{\theta_i\theta_j} > 0$ for some $(\theta_i, \theta_j) \neq (H, H)$ and $\sigma^{11}_{\theta_i\theta_j} = 0$ for $(\theta_i, \theta_j) = (H, H)$. Therefore it must be that $\sigma^{11}_{HH} > 0$.

Step 5.

Now I show that if g^P_H is sufficiently large, $\sigma^{11}_{HH} > 0$ implies that $\sigma^{11}_{HH} = 1$.

Lemma 11. *If g^P_H is sufficiently large then in any equilibrium where $\sigma^{11}_{HH} > 0$ it must be that $\sigma^{11}_{HH} = 1$.*

Proof. If g^P_H is sufficiently large it has already been shown that $\sigma^{00}_{HH} = 0$. So if $\sigma^{11}_{HH} < 1$ it must be that $\sigma^{10}_{HH} = \sigma^{01}_{HH} > 0$. The indifference condition for type (H, H) is

$$-1 + \left(F(\bar{p}_{10} g^A_H) + F(\bar{p}_{01} g^A_H)\right) g^P_H = -2 + 2F(\bar{p}_{11} g^A_H) g^P_H.$$

Rearranging this gives

$$\begin{aligned} 1/g^P_H &= 2F(g^A_H) - \left(F(\bar{p}_{10} g^A_H) + F(\bar{p}_{01} g^A_H)\right), \\ &= \epsilon > 0. \end{aligned}$$

The final inequality follows since $\bar{p}_{01} < 1$ unless $\sigma^{01}_{LH} = 0$. However it must be that $\sigma^{01}_{LH} > 0$, because if not then $\bar{p}_{11} = \bar{p}_{01}$ and so type (H, H) strictly benefits from playing $(0, 1)$ meaning that $\sigma^{11}_{HH} = 0$. □

So now I have restricted equilibria to those that take the form as in Figure 2.7.

Step 6.

Since it has already been established that in any equilibria $\sigma^{11}_{HH} = \sigma^{00}_{LL} = 1$, it is now only required to consider the strategy of types (H, L) and (L, H). Now I rule out the other possible type of equilibria with 'off-path' actions—where $\sigma^{10}_{\theta_i\theta_j} = \sigma^{01}_{\theta_i\theta_j} = 0$ for all (θ_i, θ_j).

Lemma 12. *The multi-receiver intuitive criterion rules out equilibria with $\sigma^{11}_{HH} = \sigma^{00}_{LL} = 1$ and $\sigma^{10}_{\theta_i\theta_j} = \sigma^{01}_{\theta_i\theta_j} = 0$ for all (θ_i, θ_j).*

Proof. Type (H, L) must either play $(0, 0)$ or $(1, 1)$ in such an equilibrium. I first show that if g^P_H is sufficiently large it must be the case that $\sigma^{00}_{HL} = 0$. To see this

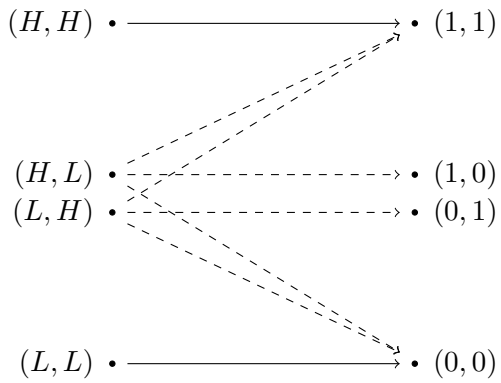

Figure 2.7: *Type* (H,H) *plays* $\sigma_{HH}^{11}=1$ *and types* (H,L) *and* (L,H) *mix* $\sigma_{HL}^{11},\sigma_{HL}^{10},\sigma_{HL}^{00}\in[0,1]$.

consider an equilibria where $\sigma_{HL}^{00}>0$. It must be that the principal (weakly) prefers to play $(0,0)$ over $(1,1)$, and so

$$F(\bar{p}_{00}g_H^A)(g_H^P+g_L^P)\geq -2+F(\bar{p}_{11}g_H^A)(g_H^P+g_L^P).$$

It must be the case that $\bar{p}_{00}<\bar{p}_{11}$, and so if g_H^P is sufficiently large, this inequality is not satisfied.

So now consider an equilibrium with $\sigma_{HH}^{11}=\sigma_{LL}^{00}=\sigma_{HL}^{11}=1$, the equilibrium payoffs are:

$$\begin{aligned}V^*(H,H)&=-2+2F(p_Hg_H^A)g_H^P,\\ V^*(H,L)&=-2+F(p_Hg_H^A)(g_H^P+g_L^P),\\ V^*(L,L)&=2F(p_Tg_H^A)g_L^P.\end{aligned}$$

Recall that p_H is defined as $p_H\equiv\frac{1}{2-p_0}$. Now consider a deviation to $b=(1,0)$. As in the proof of Lemma 10, first type (L,L) can be eliminated if g_L^P is sufficiently small. The possible beliefs that can be induced by a deviation from the remaining types are (p_1,p_2) where $\left\{(p_1,p_2)\in[0,1]^2:p_1+p_2\geq 1\right\}$. As before, I show that given type (L,L) has been eliminated, type (H,H) will want to deviate for any belief that is induced.

To verify that type (H,H) always benefits from the deviation the payoffs in equilibrium must be strictly less than the payoff of a deviation that induces belief (p_1,p_2), this is given by the following inequality

$$-1+\left(F(p_1g_H^A)+F(p_2g_H^A)\right)g_H^P>-2+2F(p_Hg_H^A)g_H^P.$$

Rearranging this give the second inequality in Assumption 9, and so this is satisfied by assumption.

□

Now I rule out type (H, L) and (L, H) playing strategies $b \neq (1, 0)$ and $b \neq (0, 1)$.

Step 7.

Lemma 13. *If 3.4 is satisfied then in any equilibrium where* $\sigma_{HL}^{10} = \sigma_{LH}^{01} > 0$ *and* $\sigma_{HH}^{11} = 1$ *it must be that* $\sigma_{HL}^{11} = \sigma_{LH}^{11} = 0$.

Proof. If $\sigma_{HL}^{11} = \sigma_{LH}^{11} > 0$ and $\sigma_{HL}^{10} = \sigma_{LH}^{01} > 0$ then

$$-2 + F(\bar{p}_{11} g_H^A)(g_H^P + g_L^P) = -1 + F(\bar{p}_{10} g_H^A) g_H^P + F(\bar{p}_{01} g_H^A) g_L^P.$$

Rearranging this gives

$$\left(F(g_H^A) - F(\bar{p}_{11} g_H^A)\right) g_H^P + \left(1 - F(\bar{p}_{11} g_H^A) g_L^P\right) = 0. \tag{1.6}$$

To get this expression I have made use of the fact that $\sigma_{LL}^{01} = \sigma_{LL}^{10} = 0$ and $\sigma_{HH}^{11} = 1$ meaning that $\bar{p}_{10} = 1$ and $\bar{p}_{01} = 0$. Since $\bar{p}_{11} < 0$, if g_H^P is sufficiently large and g_L^P is sufficiently small, then the LHS of equation 1.6 is strictly positive—which leads to a contradiction. □

Step 8.

Lemma 14. *In any equilibrium where* $\sigma_{HH}^{11} = \sigma_{LL}^{00} = 1$ *and* $\sigma_{HL}^{10} = \sigma_{LH}^{01} > 0$ *it must be that* $\sigma_{HL}^{00} = \sigma_{LH}^{00} = 0$.

Proof. If $\sigma_{HL}^{00} = \sigma_{LH}^{00} > 0$ and $\sigma_{HL}^{10} = \sigma_{LH}^{01} > 0$ then

$$F(\bar{p}_{00} g_H^A)(g_H^P + g_L^P) = -1 + F(g_H^A) g_H^P.$$

Since $\bar{p}_{00} < 1$, if g_H^P is sufficiently large this equality will no longer hold. □

So it must be that $\sigma_{HL}^{10} = \sigma_{LH}^{01} = 0$. This completes the proof of Proposition 7.

1.2 Proof of Proposition 8 and Corollary 2

The proof just requires the calculation of D_{TN} in the two examples, $F(x) = x$ and $F(x) = x^{1/4}; q = 0.5; p_0 = 0.5, g_H^P = 11; g_L^P = 1.2; g_H^A = 0.4$.

Corollary 2. *Assume that in both subgames the pure strategy separating equilibrium exists and is selected. If F is uniform transparency is optimal.*

Proof. Substitute $F(x) = x$, p_N and p_T into 3.8 to get

$$\frac{1}{2}\left(\mathbb{E}V(T) - \mathbb{E}V(N)\right) = \frac{2g_H^A(g_H^P - g_L^P)(1-p_0)p_0^3(1-q)^2 q}{(1-p_0 q)(1-2p_0 q + p_0^2 q)}.$$

The numerator is clearly positive, and the denominator is also positive since

$$1 - 2p_0 q + p_0^2 q = q(1-p_0)^2 + (1-q) > 0.$$

□

1.3 Proof of Proposition 9

Part (1) is immediate from 3.8 since the first half (which is positive) is increasing in g_H^P and the second half is not dependent on g_H^P.

Part (2) is as follows. First, note that $g_H^P = \bar{g}^P + (1-p_0)\Delta$ and $g_L^P = \bar{g}^P - p_0\Delta$. Applying this to 3.8 gives

$$\begin{aligned}\frac{1}{2}\left.\frac{\partial D_{TN}}{\partial \Delta}\right|_{\bar{g}^P} &= \frac{1}{2}\frac{\partial D_{TN}}{\partial g_H^P}\frac{\partial g_H^P}{\partial \Delta} + \frac{1}{2}\frac{\partial D_{TN}}{\partial g_L^P}\frac{\partial g_L^P}{\partial \Delta} \\ &= q(1-p_0)p_0\left(F(p_N g_H^A) - (1-p_0)F(p_T g_H^A)\right).\end{aligned}$$

So $\left.\frac{\partial D_{TN}}{\partial \Delta}\right|_{\bar{g}^P} > 0$ if and only if

$$\frac{F(p_T g_H^A)}{F(p_N g_H^A)} < \frac{1}{1-p_0}.$$

1.4 Corollary 3 and proof

Corollary 3. *When F is uniform the comparative static in part (2) of Proposition 9 is always strictly positive.*

When F is uniform:

$$\begin{aligned}\left.\frac{\partial D_{TN}}{\partial \Delta}\right|_{\bar{g}^P} &= F(p_N g_H^A) - (1-p_0)F(p_T g_H^A) \\ &= \frac{p_0^2(1-q)^2}{(1-p_0 q)(1-2p_0 q + p_0^2 q)} g_H^A > 0.\end{aligned}$$

It follows that condition to ensure that the comparative static in part (2) of Proposition 9 satisfied, and so the comparative static is strictly positive.

1.5 Proof of Proposition 10

Recall that Δ and $\bar{g}^P$ are defined as $\Delta \equiv g_H^P - g_L^P$ and $\bar{g}^P \equiv p_0 g_H^P + (1-p_0)g_L^P$. Substituting these into 5.1 and taking the derivative gives

$$\frac{1}{2}\left.\frac{\partial D'_{TN}}{\partial \Delta}\right|_{\bar{g}^P} = p_0(1-p_0)\left[(1-q)(1-p_0)(F_1 - F_{\hat{p}_N}) - qp_0(F_2 - F_{\hat{p}_N})\right] + (F_{\hat{p}_T} - F_{\hat{p}_N})p_0(1-p_0)(p_0-q). \quad (1.7)$$

Part (1) just follows from the fact that when $p_0 = 1/2$, $\hat{p}_N = \hat{p}_T$ meaning that the second part of 1.7 is zero.

For part (2) when $F(x) = x$, then

$$F_{\hat{p}_N} = \hat{p}_N F_1 + (1-\hat{p}_N)F_2,$$
$$F_{\hat{p}_T} = \hat{p}_T F_1 + (1-\hat{p}_T)F_2.$$

Substituting this into 1.7 gives

$$\begin{aligned}\frac{1}{2}\left.\frac{\partial D'_{TN}}{\partial \Delta}\right|_{\bar{g}^P} &= (F_1 - F_2)p_0(1-p_0)\times \\ &\quad \left[((1-q)(1-p_0)(1-\hat{p}_N) - qp_0\hat{p}_N) + (\hat{p}_T - \hat{p}_N)(p_0-q)\right], \\ &= (F_1 - F_2)p_0(1-p_0)\frac{(1-q)q(q(p_0-1)^3 - p_0^3(1-q))}{((p_0-q)^2 + q(1-q))(-q(1-p_0) - p_0(1-q))}.\end{aligned}$$

The final expression is clearly positive if and only if $F_1 > F_2$, or equivalently if and only if $g^A(H, \lambda_H) > g^A(L, 1)$.

Part (3) follows from the fact that when $g^A(H, \lambda_H) > g^A(L, 1)$ the first part of 1.7 is positive and when $p_0 > \max\left\{q, \frac{1}{2}\right\}$ the second part is also positive.

Part (4) follows from the fact that when $g^A(L, 1) > g^A(H, \lambda_H)$ the first part of 1.7 is negative and when $q > p_0 > \frac{1}{2}$ or $\frac{1}{2} > p_0 > q$ the second part is also negative.

1.6 Proof of Proposition 11

The unknown state (which was previously $(\lambda, \theta_1, \theta_2)$) can effectively be reduced to the bonus that the principal will pay following a^P which is $(b_1, b_2) \in \{0,1\}^2$.[41] Note that if $\lambda = \lambda_H$ then the state is $(0,0)$ for any θ_1 and θ_2. I begin by explaining why such a simplification is possible. Within the state $(0,0)$ the principal has different productivity types and thus different incentives to deviate, but for any

[41] In this context I refer to this pair as the 'state'.

productivity type pair it is the case that the principal prefers to induce a higher posterior belief for either agent. Note that when the state is $(0,0)$ the posterior of the agents is determined purely through the choice of a^P since there will never be a bonus. If there was an equilibrium in which the principal chose a different a^P within the state $(0,0)$, and these choices induced different posterior beliefs for the agents then the principal would have an incentive to deviate to the a^P that induced the higher posterior belief. This means that the principal must play the same a^P for any realisation within the state $(0,0)$.

Now I show that in any equilibrium the principal must choose transparency ($a^P = T$) in every state. I begin by considering what happens in the state $(1,1)$. Here, since the beliefs of the agents will both be that $\theta_i = H$ with probability 1, the principal is indifferent between $a^P = T$ and $a^P = N$ and so by assumption the principal chooses $a^P = T$. Now consider the other states $(1,0)$, $(0,1)$ and $(0,0)$. Assume that in an equilibrium the principal plays $a^P = N$ in some state. I show by contradiction that such an equilibrium cannot exist.

First, consider a possible equilibrium in which the principal plays $a^P = N$ only in state $(0,0)$. Here the posterior belief of agent i following $a^P = N$ and seeing $b_i = 0$ is $\Pr[\theta_i = H | a^P = N, b_i = 0] = p_T$ and following $a^P = T$ and seeing $b_i = 0$ is $\Pr[\theta_i = H | a^P = T, b_i = 0] = 0$. So the principal can gain from a deviation in which she chooses $a^P = N$ when the state is $(0,1)$. Here the posterior belief of agent i is increased from 0 to p_T and the posterior belief of agent $j \neq i$ remains unchanged.

Second, consider a possible equilibrium in which the principal plays $a^P = T$ when the state is $(0,0)$ and $a^P = N$ otherwise. Here the posterior belief of the agents will remain unchanged if $a^P = T$ is played in place of $a^P = N$ (since when agent i sees $a^P = N$ and $b = 0$ he has a posterior of 0). This means that by assumption the principal will play $a^P = T$ instead of $a^P = N$.

Finally, consider a possible equilibrium in which the principal plays $a^P = N$ in all three states $(1,0)$, $(0,1)$ and $(0,0)$. Here the posterior belief of agent i will be $\Pr[\theta_i = H | a^P = N, b_i = 0] = p_N$ if they get no bonus and 1 if they get a bonus. When the state is $(0,0)$ the principal can deviate to $a^P = T$ and increase the posterior belief of agent i from p_N to p_T and so this cannot be an equilibrium.

This means that in any equilibrium it must be the case that $a^P = T$.[42]

[42]Note that there are different equilibria, that have different off-path beliefs in the no transparency subgame.

1.7 Proof of Proposition 12

The expected payoff (in the equilibrium selected by the intuitive criterion) when $a^P = N$ is

$$\frac{1}{2}\mathbb{E}V(N) = q\left(p_0(-b_N^* + F(g_H^A)g_H^P) + (1-p_0)(0 + F(p_N g_H^A)g_L^P)\right) + (1-q)(\bar{g}_P F(p_N g_H^A)),$$

where $\bar{g}_P \equiv p_0 g_H^P + (1-p_0)g_L^P$. Substituting in $b_N^* = \Delta_N g_L^P$ gives

$$\frac{1}{2}\mathbb{E}V(N) = q\left(p_0(F(g_H^A)(g_H^P - g_L^P) + F(p_N g_H^A)g_L^P)\right) + (1-q)(\bar{g}_P F(p_N g_H^A)).$$

The expected payoff (in the equilibrium selected by the intuitive criterion) when $a^P = T$ is

$$\begin{aligned}\frac{1}{2}\mathbb{E}V(T) = q(p_0^2(-\bar{b}_T^* + F(g_H^A)g_H^P) + p_0(1-p_0)(-\underline{b}_T^* + F(g_H^A)g_H^P) + (1-p_0)^2 F(p_T g_H^A)g_L^P) \\ + (1-q)(\bar{g}_P F(p_T g_H^A)).\end{aligned}$$

Substituting in $\underline{b}_T^* = \Delta_T g_L^P - F(p_T g_H^A)g_L^P$ and $\bar{b}_T^* = \Delta_T g_L^P$ gives

$$\frac{1}{2}\mathbb{E}V(T) = q(p_0(F(g_H^A)(g_H^P - g_L^P) + F(p_T g_H^A)g_L^P + (1-q)(\bar{g}_P F(p_N g_H^A)).$$

Combining these gives

$$\begin{aligned}\frac{1}{2}D''_{TN} &\equiv \frac{1}{2}(\mathbb{E}V(T) - \mathbb{E}V(N)) \\ &= \left(F(p_T g_H^A) - F(p_N g_H^A)\right)\left(q g_L^P + (1-q)\bar{g}_P\right) \\ &> 0.\end{aligned} \tag{1.8}$$

1.8 Proof of Proposition 13

Both parts are immediate from the functional form of D''_{TN} in 1.8.

2 Definition of Intuitive Criterion with multiple receivers

As mentioned in the main text, the intuitive criterion is not defined for the class of games with multiple receivers (agents in my terminology). In general this potentially complicates things. For example, following the action of the sender (and beliefs induced), if the receivers actions influence each other's payoffs (i.e. there is a game), it is not clear what set of potential outcomes should be

considered for each possible belief induced. In the setting with a single receiver, the intuitive criterion compares the payoff for the principal when the receiver best responds to the sender's equilibrium payoff. In my setting, since the receivers (agents) have payoffs that do not depend of the belief of the other agent—and so a game is not induced following the sender (principal's) action—the best responses can still be used as before, thus allowing the intuitive criterion to be extended in a natural way.

The procedure I propose is as follows. As in the standard intuitive criterion, at the stage where some types are excluded from a particular deviation, I consider whether each type can possibly get a higher payoff than their (expected) equilibrium payoff from that particular deviation given any possible beliefs of *all* agents (receivers).[43] Then having excluded these types, as in the single receiver definition, an equilibrium fails the intuitive criterion if there is a type such that a deviation gives a strictly higher payoff compared to the equilibrium payoff for any best response of the agents (receivers) given that the type cannot be any of those excluded.

More formally, I adapt the Definition 11.4 in Fudenberg and Tirole (1991). Before going into the definition, I simplify the latter part of the game tree so that the agents' action is a choice of cutoff for which outside option they will accept, rather than having a realisation of outside option then a binary stay/quit decision—this is strategically equivalent to the game described in the text. Denote the vector of cutoffs that the agents choose by $\bar{u}^A = \left(\bar{u}_1^A, \bar{u}_2^A\right) \in [0,1]^2$, so that $a_i^A = Q$ if and only if $u_i < \bar{u}_i^A$. I also introduce notation for best responses given an action from the principal and belief that it induces. Let $T \subseteq \Theta$ where Θ is the set of all possible $\theta = (\theta_1, \theta_2)$ 'types' of the principal (productivity of the agents she employs). Let T_i denote the i^{th} projection map of T, $T_i = proj_i(T)$. Let $\mu_i \equiv \mu(\theta_i|b)$ be the beliefs of agent i following bonuses $b = (b_1, b_2)$.[44] Denote the (expected) payoff of agent i in terms of the cutoff strategy $\bar{u}_i^A$ by $U_i(b, \bar{u}_i^A, \theta_i)$—note that $\bar{u}_i^A$ and θ_i do not affect the payoff of agent $j \neq i$. Denote the principal's (expected) payoff by $V(b, \bar{u}_1^A, \bar{u}_2^A, \theta)$. The set of best response vectors for agent i when $\theta \in T$ and a bonus b is paid is given by

$$\mathrm{BR_i}(T, b) = \bigcup_{\mu_i : \mu(T_i|b)=1} \mathrm{BR}(\mu_i, b),$$

where

[43] The standard definition considers only the belief and action of the single receiver.

[44] Note that since θ_1 and θ_2 are independent, $\mu(\theta_i|b)$ does not impose any restrictions on the values that $\mu(\theta_j|b)$ can take.

$$\text{BR}(\mu_i, b) = \arg\max_{\bar{u}_i^A \in [0,1]} \sum_{\theta_i \in T_i} \mu(\theta_i|b) U_i(b, \bar{u}_1^A, \theta_i).$$

Definition 3 (Multi-receiver intuitive criterion)**.** *Fix a vector of equilibrium payoffs for the principal: $V^*(\cdot)$. For each strategy vector b, let $J(b)$ be the set of all θ such that*

$$V^*(\theta) > \max_{(\bar{u}_1^A, \bar{u}_2^A) \in (\text{BR}_1(\Theta, b), \text{BR}_2(\Theta, b))} V(b, \bar{u}_1^A, \bar{u}_2^A, \theta).$$

If there exists θ' and b such that

$$V^*(\theta') < \min_{(\bar{u}_1^A, \bar{u}_2^A) \in (\text{BR}_1(\Theta \setminus J(b), b), \text{BR}_2(\Theta \setminus J(b), b))} V(b, \bar{u}_1^A, \bar{u}_2^A, \theta'),$$

then the equilibrium fails the multi-receiver intuitive criterion.

The key features of the setting that allow the intuitive criterion to be extended in this way are:

1. Agent i's payoff does not depend on agent j's action. If this were not the case then $\text{BR}(\mu_i, b)$ would not be defined in the way it has been above. Instead the principal's choice of b and the corresponding belief would induce a game played between the two agents which means that the definition would have to be adapted further;

2. Agents' productivities are independent and so $\mu(\theta_i|b)$ does not impose any restrictions on the values that $\mu(\theta_j|b)$ can take;

3. Agent i's payoff does not depend on θ_j. Combined with independence assumption above, this means that for a given b, $\text{BR}_i(T, b)$ does not restrict $\text{BR}_j(T, b)$ in any way. This means that the 'max' and 'min' over the principal's expected utility within the definition can be done dimension by dimension.

3 Constraints from Section 5.1

In this Appendix I provide sufficient conditions for the equilibrium described in Section 5.1 to exist in both subgames. I also verify that there exist parameter values that satisfy these conditions.

3.1 No transparency

Consider an equilibrium in the no transparency subgame where the principal pays agent i $b_i = 0$ if $\eta = \eta_0$, $b_i = 1$ if $\eta = \eta_1$ or η_2, and $b_i = 2$ if $\eta = \eta_3$. There are 8 incentive constraints ensuring that each type does not deviate to one of the other two possible bonus levels. As with 3.1 and 3.2, in Section 3, these will be satisfied when g_L^P is small and g_H^P is large. It must also be the case that λ_H is not too different from g_H^P/g_L^P. Intuitively, this latter condition means that types η_1 and η_2 have similar incentives to deviate which means that in equilibrium they pool.

The 8 constraints are as follows. Recall the notation from the main text—the posterior probability of agent i being of high productivity given that he receives a bonus $b_i = 1$ is given by

$$\begin{aligned}\hat{p}_N &\equiv \Pr[\theta_i = H | b_i = 1; a^P = N] \\ &= \frac{(1-q)p_0}{(1-q)p_0 + q(1-p_0)},\end{aligned}$$

$F_k \equiv F(g^A(\eta_k))$ and $F_{\hat{p}_N} \equiv F\left(\hat{p}_N g^A(\eta_1) + (1-\hat{p}_N)g^A(\eta_2)\right)$. I also normalise $F_0 = 0$.

The first two constraints ensure that type η_0 does not choose $b_i = 1$ and type η_1 does not choose $b_i = 0$.[45] These are given by

$$\begin{aligned}-0 + F_0 g_L^P &\geq -\lambda_H + F_{\hat{p}_N} g_L^P, \\ -\lambda_H + F_{\hat{p}_N} g_H^P &\geq -0 + F_0 g_H^P.\end{aligned}$$

The next sets of constraints ensure that type η_2 does not choose $b_i = 2$ and type η_3 does not choose $b_i = 1$. These are given by

$$\begin{aligned}-1 + F_{\hat{p}_N} g_L^P &\geq -2 + F_3 g_L^P, \\ -2 + F_3 g_H^P &\geq -1 + F_{\hat{p}_N} g_H^P.\end{aligned}$$

Finally, there are constraints that ensure that type η_2 does not choose $b_i = 0$ and type η_1 does not choose $b_i = 2$. These are given by

$$\begin{aligned}-1 + F_{\hat{p}_N} g_L^P &\geq -0 + F_0 g_L^P, \\ -\lambda_H + F_{\hat{p}_N} g_H^P &\geq -2\lambda_H + F_3 g_H^P.\end{aligned}$$

[45] In this context 'type' refers to the agent i that the principal is facing given the agent's productivity and the principal's costs $\eta = (\theta_i, \lambda)$.

Note that if these constraints are satisfied there is no need for constraints that ensure that type η_0 does not choose a bonus level $b_i = 2$ and type η_3 does not choose a bonus level $b_i = 0$. Combining the 6 constraints above gives

$$\begin{aligned}(F_3 - F_{\hat{p}_N})g_H^P &\leq \lambda_H \\ F_{\hat{p}_N} g_L^P \leq \lambda_H &\leq F_{\hat{p}_N} g_H^P \\ (F_3 - F_{\hat{p}_N})g_L^P \leq 1 &\leq (F_3 - F_{\hat{p}_N})g_H^P \\ 1 &\leq F_{\hat{p}_N} g_L^P.\end{aligned}$$

3.2 Transparency

Consider an equilibrium in the transparency subgame where the principal again the principal pays agent i $b_i = 0$ if $\eta = \eta_0$, $b_i = 1$ if $\eta = \eta_1$ or η_2, and $b_i = 2$ if $\eta = \eta_3$. This equilibrium has an off-path action of paying a bonus pair $(2, 0)$ (or $(0, 2)$). To support the equilibrium the agents' beliefs following this action is that they are both type η_0, the most pessimistic belief.

There are now 8 possible pairs of states that can occur for each agent (note some of the 16 possible pairs such as (η_0, η_2) are not possible since λ must be the same for both agents) and so there are potentially many incentive constraints to consider. It is sufficient to consider only 18 of these constraints. Since there are so many constraints it may seem problematic to find parameters that satisfy all of these constraints, however when $g_H^P/g_L^P = \lambda_H$ and $g^A(\eta_1) = g^A(\eta_2)$ these conditions simplify significantly (this is described in more detail later) and so when this is satisfied it is much more straightforward to find parameters that satisfy all the constraints.

The 18 constraints are as follows. Recall the notation from the main text—the posterior probability of agent i being of high productivity given that he receives a bonus $b_i = 1$ and he sees the other agent receives a bonus $b_j = 1$ is given by

$$\begin{aligned}\hat{p}_T &\equiv \Pr[\theta_i = H | b_i = b_j = 1; a^P = T] \\ &= \frac{(1-q)p_0^2}{(1-q)p_0^2 + q(1-p_0)^2},\end{aligned}$$

and $F_{\hat{p}_T} \equiv F\left(\hat{p}_T g^A(\eta_1) + (1-\hat{p}_T)g^A(\eta_2)\right)$.

The first 6 constraints are for types where $\lambda = \lambda_H$.[46] First, type (η_0, η_0) cannot prefer to pay bonuses $(1, 0)$, and type (η_1, η_0) cannot prefer to pay bonuses

[46]Type in this context is a pair (η_i, η_j) for the two agents.

$(0,0)$.[47] These constraints can be simplified to

$$F_1 g_L^P \leq \lambda_H \leq F_1 g_H^P.$$

Next, there is a pair of constraints that ensure that type (η_1, η_0) does not pay bonuses $(1,1)$ and type (η_1, η_1) does not pay bonuses $(1,0)$. These constraints can be simplified to

$$(F_{\hat{p}_T} - F_1) g_H^P + F_{\hat{p}_T} g_L^P \leq \lambda_H \leq (F_{\hat{p}_T} - F_1) g_H^P + F_{\hat{p}_T} g_H^P.$$

Finally, there is a pair of constraints that ensure that type (η_0, η_0) does not pay bonuses $(1,1)$ and type (η_1, η_1) does not pay bonuses $(0,0)$. These constraints can be simplified to

$$F_{\hat{p}_T} g_L^P \leq \lambda_H \leq F_{\hat{p}_T} g_H^P.$$

Notice that when $g_H^P / g_L^P = \lambda_H$ and $g^A(\eta_1) = g^A(\eta_2)$ these three conditions are all equivalent.

The next 6 constraints for types where $\lambda = 1$ and are similar to those above. First, type (η_2, η_2) cannot prefer to pay bonuses $(2,1)$, and type (η_3, η_2) cannot prefer to pay bonuses $(1,1)$. These constraints can be simplified to

$$(F_3 - F_{\hat{p}_T}) g_L^P + (F_2 - F_{\hat{p}_T}) g_L^P \leq 1 \leq (F_3 - F_{\hat{p}_T}) g_H^P + (F_2 - F_{\hat{p}_T}) g_L^P.$$

Next, there is a pair of constraints that ensure that type (η_3, η_2) does not pay bonuses $(2,2)$ and type (η_3, η_3) does not pay bonuses $(2,1)$. These constraints can be simplified to

$$(F_3 - F_2) g_L^P \leq 1 \leq (F_3 - F_2) g_H^P.$$

Finally, there is a pair of constraints that ensure that type (η_2, η_2) does not pay bonuses $(2,2)$ and type (η_3, η_3) does not pay bonuses $(1,1)$. These constraints can be simplified to

$$(F_3 - F_{\hat{p}_T}) g_L^P \leq 1 \leq (F_3 - F_{\hat{p}_T}) g_H^P.$$

There next 2 constraints that ensure that type (η_1, η_1) does not deviate to

[47] There will obviously be identical constraints for bonuses $(0,1)$ and types (η_0, η_1). I omit these here and also later on when similar symmetric constraints need to be satisfied.

higher bonuses of either $(2,1)$ or $(2,2)$. These constraints can be simplified to

$$(F_3 - F_{\hat{p}_T})g_H^P + (F_2 - F_{\hat{p}_T})g_H^P \leq \lambda_H,$$

and

$$(F_3 - F_{\hat{p}_T})g_H^P \leq \lambda_H.$$

Similar to the 2 constraints above, there are 2 constraints that ensure that type (η_2, η_2) does not deviate to lower bonuses of either $(1,0)$ or $(0,0)$. These constraints can be simplified to

$$1 \leq (2F_{\hat{p}_T} - F_1)g_L^P,$$

and

$$1 \leq F_{\hat{p}_T} g_L^P.$$

The final set of 2 constraints ensure that type (η_1, η_0) does not pay bonuses $(2,1)$ and that type (η_3, η_2) does not pay bonuses $(1,0)$. These constraints can be simplified to

$$\frac{1}{2}(F_3 g_H^P + F_2 g_L^P - F_1 g_H^P) \leq \lambda_H,$$

and

$$1 \leq \frac{1}{2}(F_3 g_H^P + F_2 g_L^P - F_1 g_H^P).$$

The following parameters satisfy all 6 constraints under no transparency and all 18 constraints under transparency: $g_H^P = 10; g_L^P = 2; \lambda_H = 5; p_0 = .5; q = .5; g^A(\eta_1) = .7; g^A(\eta_2) = .8; g^A(\eta_3) = .91; F(x) = x$. For these values $\eta_2 \succ \eta_1$, if instead $g^A(\eta_2) = .7; g^A(\eta_1) = .8$ the conditions are still satisfied and $\eta_1 \succ \eta_2$.

4 Analysis of continuous choice of bonus

Throughout this section, I make a parametric assumption that $F(g_H^A) - 2F(p_T g_H^A) > 0$, which ensures it is possible to a pure strategy separating equilibrium under transparency. Under no transparency a (pure strategy) separating equilibrium must have a bonus, $b_i = b_N > 0$ when $\theta_i = H$ and $\lambda = 1$ and no bonus, $b_i = 0$, otherwise. To support such an equilibrium the off path beliefs must be such that if $b_i \in (0, b_N)$ then agent i assigns probability 1 to $\theta_i = L$ (i.e. pessimistic beliefs).[48] The incentive constraints of the principal to ensure that she pays a bonus when $\theta_i = H$ and $\lambda = 1$ and does not pay a bonus when $\theta_i = L$ and $\lambda = 1$

[48]The beliefs assigned to $b_i > b_N$ do not matter since regardless of what these are neither type would deviate to this choice of b_i.

are given by

$$-b_N + F(g_H^A)g_H^P \geq -0 + F(p_N g_H^A)g_H^P,$$
$$-0 + F(p_N g_H^A)g_L^P \geq -b_N + F(g_H^A)g_L^P,$$

respectively. Combining these gives

$$\Delta_N g_L^P \leq b_N \leq \Delta_N g_H^P, \tag{4.1}$$

where $\Delta_N \equiv F(g_H^A) - F(p_N g_H^A)$. Applying the intuitive criterion refinement results in the unique equilibrium with $b_N^* = \Delta_N g_L^P$.[49]

Under transparency a (pure strategy) separating equilibrium potentially has different bonus levels depending on what both agents' productivities are. When $\lambda = 1$ the bonuses are given as in the table below.[50]

	H	L
H	$\bar{b}_T, \bar{b}_T$	$\underline{b}_T, 0$
L	$0, \underline{b}_T$	0,0

As before, the off-path beliefs that support such an equilibrium are that agent i assigns probability 1 to $\theta_i = L$ when the principal pays a pair of bonuses $(b_i, b_j) \neq (0, \underline{b}_T), (\underline{b}_T, 0), (\bar{b}_T, \bar{b}_T)$.

There are now 6 incentive constraints that ensure: type (L, L) doesn't deviate to (H, L) and vis-versa, type (H, L) doesn't deviate to (H, H) and vis-versa, and type (L, L) doesn't deviate to (H, H) and vis-versa.[51] The constraints can be simplified to

$$\begin{aligned} \Delta_T g_L^P \leq \bar{b}_T &\leq \Delta_T g_H^P, \\ F(g_H^A)g_L^P \leq 2\bar{b}_T - \underline{b}_T &\leq F(g_H^A)g_H^P, \\ \Delta_T g_L^P - F(p_T g_H^A)g_L^P \leq \underline{b}_T &\leq \Delta_T g_H^P - F(p_T g_H^A)g_L^P, \end{aligned}$$

where $\Delta_T \equiv F(g_H^A) - F(p_T g_H^A)$. The unique equilibrium that survives the intuitive criterion is at the intersection at the lowerbound of all of the above

[49] I omit the arguments for the intuitive criterion refinements in this section, since they are very similar to the arguments given for Proposition 7.

[50] Note that in any pure strategy separating equilibrium an agent with productivity $\theta_i = L$ must get a bonus of 0. If this was not the case the principal could deviate and pay a lower bonus.

[51] There are obviously identical constraints for (L, H) but I omit these.

inequalities. This means that

$$\underline{b}_T^* = \Delta_T g_L^P - F(p_T g_H^A) g_L^P,$$
$$\bar{b}_T^* = \Delta_T g_L^P.$$

Note that since it is assumed that $F(g_H^A) - 2F(p_T g_H^A) > 0$ it must be that $\underline{b}_T^* > 0$.

Proposition 12. *Assume an equilibrium exists and is selected in which:*

- *in the no transparency subgame, the principal pays an agent b_N^* if and only if $\lambda = 1$ and the agent is of high productivity, and the principal pays 0 otherwise;*
- *in the transparency sugame, the principal pays both agents $\bar{b}_T^*$ if and only if $\lambda = 1$ and both agents are of high productivity, the principal pays $\underline{b}_T^*$ and 0 to the agent with high and low productivity respectively if and only if $\lambda = 1$ and there is one agent with high productivity and one agent with low productivity, and the principal pays 0 to both agents otherwise.*

In such an equilibrium $a^P = T$ is always strictly preferred by the principal.

When $\lambda = 1$, under transparency the principal does not need to pay bonuses as high as under no transparency. There are two cases where at least one agent has high productivity and so a bonus is paid. The first has one agent with high productivity and the other with low productivity. In this case, under transparency, the cost of revealing to the low productivity agent that he is definitely the low type is internalised by paying a lower bonus to the agent with high productivity in order to signal to him that he is of high productivity. In effect the principal can signal to agents that this is the state by paying lower bonuses (compared with no transparency) because the agents know that this comes at a cost.[52] The second case has both agents with high productivity. Again, under transparency, the principal does not need to pay bonuses as high as she does under no transparency. The reason is that when both agents aren't paid a bonus agents are more likely to stay at the firm under transparency, but when both agents are paid a bonus there is no difference under transparency and no transparency. This results in the bonus having to be higher under no transparency to

[52] In the discrete bonus set-up of the previous section, when there is one high and one low productivity agent, the principal was forced to pay the *same* signalling cost. Since the posterior beliefs in the no transparency subgame were more favourable for this realisation of states, this resulted in a higher expected payoff for no transparency compared to transparency. Due to the lower signalling costs, there is no difference in the expected payoff with the continuous bonuses.

ensure that when the agents have low productivity the principal does not have an incentive to deviate (and realise a greater increase in the posterior belief). In addition to the lower bonuses, when no bonus is paid transparency is better for the principal (since $p_T > p_N$).

As before, I analyse the comparative statics on the difference in the expected value of transparency and no transparency.

Proposition 13. *Assume both transparency and no transparency lead to the pure strategy equilibrium selected by the intuitive criterion, then:*

1. *increasing g_H^P while keeping all other parameters ($g_L^P, g_H^A, p_0, q, F(\cdot)$) constant leads to an increase in D''_{TN};*

2. *increasing $g_H^P - g_L^P$ while keeping all the parameters $g_H^A, p_0, q, F(\cdot)$, and $\bar{g}^P$ constant leads to an increase in D''_{TN}.*

Part (2) of the result means it must be that the difference in payoff between transparency and non-transparency is increasing as there is more heterogeneity in productivity. As in the previous sections, this suggests that a firm facing a greater level of heterogeneity in productivity would want to commit to transparency.

5 Correlation in agents' productivity and outside option and agents receiving informative signals

Formally, I assume that $u_i \sim F(u_i|\theta_i)$, and that this has full support for all θ_i. The constraints for the principal can be derived in the same way as before. Note that the belief of the agent is only updated with the additional information learned from u_i when the belief is interior—i.e. not at 0 or 1. For example, 3.1 becomes

$$[F(g_H^A|H) - F(\hat{p}_N(H)g_H^A|H)]g_H^P \geq 1, \tag{5.1}$$

where $\hat{p}_N(\theta_i)$ is the expected belief of the agent given that his productivity is θ_i and he was not paid a bonus under no transparency. This can be computed by Bayes rule, for example for $\theta_i = H$

$$\hat{p}_N(H) \equiv \int_0^1 \frac{p_N f(u|H)}{p_N f(u|H) + (1-p_N)f(u|L)} du.$$

Because $F(\cdot|\theta_i)$ has full support, it follows that $\hat{p}_N(H) < 1$. So, as before, if g_H^P is sufficiently large, then this constraint will be satisfied. For the other 3

constraints—equivalent to 3.2, 3.3, and 3.4—it can also be shown that these are satisfied if g_H^P and g_L^P are sufficiently large and small.

Next, the difference between transparency and no transparency can be derived as in 3.8,

$$\begin{aligned}\frac{1}{2}\hat{D}_{TN} &\equiv \frac{1}{2}\left(\mathbb{E}\hat{V}(T) - \mathbb{E}\hat{V}(N)\right) \\ &= 2g_L^P\left((1-q)(1-p_0) + q(1-p_0)^2\right)\left(F(\hat{p}_T(L)g_H^A|L) - F(\hat{p}_N(L)g_H^A|L)\right) \\ &\quad + 2g_H^P(1-q)p_0\left(F(\hat{p}_T(H)g_H^A|H) - F(\hat{p}_N(H)g_H^A|H)\right) \\ &\quad - 2qp_0(1-p_0)g_L^P F(\hat{p}_N(L)g_H^A|L),\end{aligned}$$

where $\hat{p}_T(\theta_i)$ is is the expected belief of the agent given that his productivity is θ_i and he was not paid a bonus under transparency. This can be computed by Bayes rule in a similar way to $\hat{p}_N(\theta_i)$. Since $p_T > p_N$, it follows that $\hat{p}_T(\cdot) > \hat{p}_N(\cdot)$. This means that the comparative statics on $\hat{D}_{TN}$ with respect to the parameters g_H^P and g_L^P do not change—i.e. Proposition 9 is still valid.

6 No bonus as perfect bad news

In this appendix I provide the formal results for the alternative model of Section 5.5 where no bonuses act as perfect bad news, rather than having a bonus act as perfect good news (as in Section 3).

The model is as in Section 3, but with λ_H taking a finite value chosen so that there exists an equilibrium in which the principal chooses $b_i = 0$ if and only if $\lambda = \lambda_H$ and $\theta_i = L$. In order to establish the conditions on the parameters λ_H, g_L^P and g_H^P for existence, I consider the IC constraints of the principal in such an equilibrium.

Define:

$$\begin{aligned}\tilde{p}_N &\equiv \Pr[\theta_i = H | b_i = 1] = \frac{p_0}{q + (1-q)p_0}, \\ \tilde{p}_T &\equiv \Pr[\theta_i = H | b_i = b_j = 1] = \frac{qp_0 + (1-q)p_0^2}{q + (1-q)p_0^2},\end{aligned}$$

and note that $\tilde{p}_N > \tilde{p}_T$.

I start with no transparency ($a^P = N$). The principal must always pay a bonus when $\theta_i = H$, clearly it is sufficient to ensure that this is the case when $\lambda = \lambda_H$—the constraint is weaker when $\lambda = 1$. The constraint is

$$-\lambda_H + F(\tilde{p}_N g_H^A)g_H^P \geq 0.$$

When $\theta_i = L$ the principal must only want to pay a bonus when $\lambda = 1$, so there are two constraints

$$\begin{aligned} -1 + F(\tilde{p}_N g_H^A) g_L^P &\geq 0, \\ -\lambda_H + F(\tilde{p}_N g_H^A) g_L^P &\leq 0. \end{aligned}$$

Now I consider the case of transparency ($a^P = T$). When $\theta_i = H$ the principal must always pay a bonus. When $\lambda = 1$, since the other agent will always be paid a bonus, the constraint is

$$-1 + F(\tilde{p}_T g_H^A) g_H^P \geq 0.$$

When $\lambda = \lambda_H$, clearly it is sufficient to ensure that the principal pays the bonus when $\theta_j = H$, the constraint is

$$-\lambda_H + F(\tilde{p}_T g_H^A) g_H^P \geq 0.$$

When $\theta_i = L$ and $\lambda = 1$, for any θ_j the constraint to ensure the principal pays the bonus is

$$-1 + F(\tilde{p}_T g_L^P) g_L^P \geq 0.$$

When $\theta_i = L$ and $\lambda = \lambda_H$, it is sufficient that the constraint

$$-\lambda_H + F(g_H^A) g_L^P \leq 0$$

is satisfied.

In order to satisfy these constraints, fixing λ_H, p_0, q, the parameters g_H^A, g_H^P and g_L^P can be chosen to satisfy

$$\begin{aligned} &g_H^P \geq \max\left\{1/F(\tilde{p}_T g_H^A), \lambda_H / F(\tilde{p}_N g_H^A)\right\}; \\ &1/F(\tilde{p}_T g_H^A) \leq g_L^P \leq \lambda_H / F(g_H^A). \end{aligned}$$

Now as in 3.8, I provide an expression for the difference between the expected value of transparency and no transparency for the principal and then show that the comparative statics are analogous to those in Proposition 9.

The expression for the difference is given by

$$\begin{aligned}\frac{1}{2}\tilde{D}_{TN} &\equiv \frac{1}{2}\left(\mathbb{E}V(T) - \mathbb{E}V(N)\right) \\ &= -q\left(F(\tilde{p}_N g_H^A) - F(\tilde{p}_T g_H^A)\right)\bar{g}^P \\ &\quad + (1-q)p_0 g_H^P\left(-F(\tilde{p}_N g_H^A) + (1-p_0)F(g_H^A) + p_0 F(\tilde{p}_T g_H^A)\right).\end{aligned}$$

If g_L^P is decreased (and all other parameters are held fixed), then the first line is increased and the second line stays constant. If $\bar{g}^P$ is held fixed and the difference between g_H^P and g_L^P is increased, then the first line is constant and the second line increases if and only if the expression in the brackets is positive (which it is if there is not too much density in F between $\tilde{p}_T$ and $\tilde{p}_N$). This means the comparative statics are analogous to those in Proposition 9.

Chapter 3

Pay Transparency in Organisations—A Dynamic Model

1 Introduction

In the static signalling game analysed in Chapter 2, the continuation payoffs of the players are *exogenous*. Natural assumptions are made so that when the productivity of an agent is higher, the expected future surplus (or continuation payoff) going to both the agent and the principal is higher. I provide micro-foundations for these assumptions by analysing a dynamic model in which the agents' productivities evolve over time and the continuation values of the players arise *endogenously*.[1] The motive for the principal to pay a bonus is as in the static game—paying a bonus to an agent signals to him that he has high productivity today and so he is more likely to have high productivity tomorrow (and receive further bonuses). I show that an equilibrium exists, in which, in every stage game, the principal's strategy is the same as the equilibrium of interest in the static model. In this equilibrium, the principal's and the agents' continuation values are increasing in the agents' productivities—this matches the 'reduced-form' assumptions in the static model. To demonstrate the robustness of the comparative statics in the static model, I analyse setups that are analogous to no transparency and transparency in the static model. I show numerically that increasing the value of employing a high productivity agent makes transparency

[1] There is one slight discrepancy: As will become clear, in the dynamic model, given the productivity of the agent, the principal prefers to retain an agent with a higher belief. This is not built into the reduced form payoffs in the static model of Chapter 2. However, if they were, the results would not qualitatively change.

more favourable.

I start with a single agent which is analogous to no transparency in the static model.[2] Then, I analyse what I call 'partial transparency', which is analogous to the transparency in the static model. Under partial transparency, I assume for simplicity that only one of the two agents is 'active'—and so chooses whether or not to quit based on his (changing) belief. The other agent is assumed to always get a bonus if the principal has funds available. The difference between a single agent/no transparency and this setting is that the active agent is still able to learn from the other agent. For example, if the active agent did not receive a bonus and the passive agent did, the active agent infers that he must be of low productivity. First, I show that in both settings an equilibrium exists for some parameter values. Then, I numerically approximate the expected value of the principal, and show that the comparative static result in Chapter 2 continues to hold.

My model is related to reputation models with changing types—for example, Phelan (2006). The agent is unsure of the 'type' of the principal (his own productivity), and this evolves over time. His beliefs are updated by observing the actions of the principal. A key difference between my model and the existing literature is that both the agents and the principal are long-run players—in contrast in Phelan (2006), only the principal is a long-run player.

2 Model with a single agent

2.1 Set-up

I analyse a model with a single agent and principal. Time is discrete and infinite and is denoted by $t = 1, 2, ...$

In period t, the agent has productivity given by $\theta_t \in \{H, L\}$. This changes stochastically over time through a Markov process, with persistence $\rho \in (\frac{1}{2}, 1)$, so $\Pr[\theta_t = \theta_{t-1}] = \rho$. The assumption that $\rho > \frac{1}{2}$ means that the productivity is positively correlated across consecutive periods. At $t = 1$ productivity is drawn such that $\Pr[\theta_1 = H] = p_0$. In every period, there is a cost shock for the principal $\lambda_t \in \{1, \infty\}$. This is drawn independently over time from the

[2]Note that in reality it might be that even if the principal committed to no transparency about bonuses, an agent could observe (and make inferences) from the exit of other agents. Adding this into the analysis makes the problem intractable as the agent would have to keep track of the (distribution over) higher order beliefs of the other agent. For this reason, I assume that when the principal commits to no transparency, the agent observes neither the bonus nor the exits of the other agent.

distribution $\Pr[\lambda_t = 1] = q$.[3] In every period, the agent gets an outside offer u_t that is modelled as a lump sum payment. This is also drawn independently over time from the distribution $u_t \sim F[0,1]$ where F has full support and no mass points.

Stage game.

1. The principal privately learns the productivity of the agent (θ_t) and the cost shock (λ_t).

2. The principal chooses whether or not to pay the agent a bonus, $b_t \in \{0, b\}$, $b > 0$.[4]

3. The agent learns his outside option u_t.

4. The agent chooses whether to stay at the firm or to quit, denote this decision by $a_t^A \in \{S, Q\}$.

5. The players receive their payoffs for period t as described below. If the agent chose to stay at the firm ($a_t^A = S$) then the players repeat the stage game in period $t+1$, if the agent chose to quit the firm ($a_t^A = Q$) the game ends.

Payoffs. Both players discount future periods with a common discount factor $\delta \in (0,1)$. The principal's period t payoff is

$$V_t = -\lambda_t b_t + \mathbb{1}[\theta_t = H]v,$$

where $v > b$ is the payoff from employing a high productivity worker. The payoff for employing a low productivity worker is normalised to 0. The agent's period t payoff is

$$U_t = b_t + \mathbb{1}[a_t^A = Q]u_t.$$[5]

[3] The independence across time corresponds to the set up of Section 3 in Chapter 2 in which the agent does not have preferences over the principal's costs today when deciding whether or not to stay. It is also a simplify assumption made to ensure the problem is tractable. In reality the principal's costs may be positively correlated across consecutive periods as discussed in Section 5.1 of Chapter 2. If there was correlation across periods, the principal would retain private information making the problem much more challenging.

[4] Note that in the static model it was without loss to assume that $b = 1$ since it was in effect a normalisation with the exogenously given continuation payoffs. Here due to the repeated nature of the game, continuation payoffs become endogenous, this is no longer the case.

[5] Note that the lump sum payoff for quitting is equivalent to the discounted stream of future payoffs of $\bar{u}_t$ in every future period where $\bar{u}_t = \frac{\delta}{1-\delta}u_t$.

Equilibrium. I focus on Markov perfect equilibria (for the rest of the section this is just 'equilibrium') where the publicly known 'state' variable is the agent's belief at the start of period t, $p_t \equiv \Pr[\theta_t = H|b^{t-1}]$, where $b^{t-1} = (b_1, ..., b_{t-1})$ is the publicly observed history of bonuses.[6]

2.2 Analysis of the single agent model

I construct an equilibrium in which, in the period t stage game, when the agent has a belief p_t:

- For any p_t, the principal pays a bonus $b_t = b$ if and only if $\theta_t = H$ and $\lambda_t = 1$;
- The agent chooses to stay at the firm ($a_t^A = S$) if and only if his expected value of staying at the firm is greater than quitting the firm, which is the case when
$$\delta U(p_{t+1}(p_t, b_t)) \geq u_t,$$
where $U(p)$ is the (expected) value for the agent of having a belief p (defined formally below), and $p_{t+1}(p_t, b_t) = \Pr[\theta_{t+1}|p_t, b_t]$ is the belief in the next period given the current belief and bonus in the current period (p_t, b_t).

The belief in period t is updated twice to get the belief in period $t+1$. First, the agent observes b_t and updates using Bayes rule. Defining $\hat{p}_t(b_t, p_t) \equiv \Pr[\theta_t = H|b_t, p_t]$ as the update to the belief following b_t. Given the strategy described above, this is

$$\hat{p}_t(1, p_t) = 1 \text{ for all } p_t,$$

$$\hat{p}_t(0, p_t) = \frac{p_t(1-q)}{1-p_t q}.$$

Second, this posterior belief undergoes the Markov transition that dictates how the productivity evolves between periods, so

$$p_{t+1} = \rho\hat{p}_t + (1-\rho)(1-\hat{p}_t).$$

[6] Note that this is technically slightly different from the standard concept of Markov equilibrium since the principal can condition her action on her private information and so her action is not measurable with respect to the state. The reason that there is this difference is that the stage game is in effect an extensive form game, whereas usually the stage game is a normal form game. However, note that the principal's strategy will be required to be measurable with respect to the private information she learns in the current period, and in an equilibrium she cannot, for example, condition her strategy on private information from previous periods.

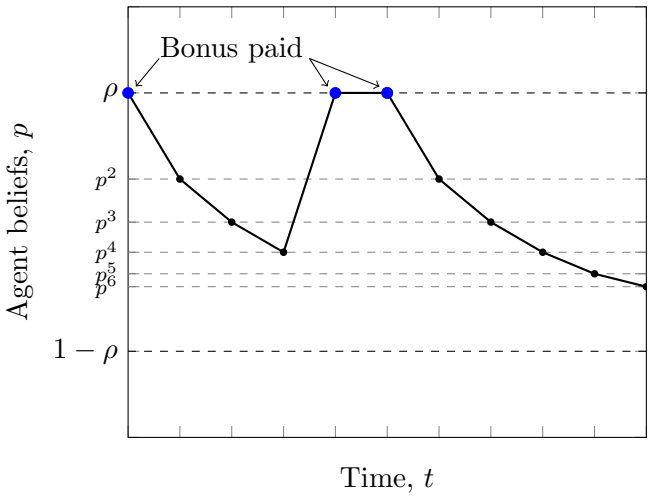

Figure 3.1: *An example of agent's belief path with a single agent.*

It will be helpful to define a sequence of beliefs $\{p^n\}_{n=1}^{n=\infty}$ where

$$p^1 = \rho,$$

$$p^{n+1} = \frac{p^n(1-q)}{1-p^n q}\rho + \left(1 - \frac{p^n(1-q)}{1-p^n q}\right)(1-\rho) \text{ for } n \geq 1.$$

The sequence starts with the most optimistic belief the agent can have, $p_t = p^1 = \rho$, which occurs following a bonus in the previous period. Each step in the sequence represents the lower belief when there has been an additional period since the last bonus was paid—so a belief p^n means that it has been n periods since the last bonus was paid. The sequence is bounded below by $1-\rho$, since even if the agent had a posterior after a history of $\hat{p}_t = 0$, the transition probability brings the belief back up to $1-\rho$. Since the sequence is monotonic, this means that the sequence converges to p^∞, which solves the equation

$$p^\infty = \frac{p^\infty(1-q)}{1-p^\infty q}\rho + \left(1 - \frac{p^\infty(1-q)}{1-p^\infty q}\right)(1-\rho). \tag{2.1}$$

I illustrate an example of a path of the agent's beliefs in Figure 3.1.

Denote the expected value of an agent with belief p^n by $U^n \equiv U(p^n)$.[7] These

[7]Note that this assumes that $p_t \in \{p^1, p^2, ..., p^n, ...\}$ for all t which will be the case if $p_0 \in \{p^1, p^2, ..., p^n, ...\}$. The results in the section continue to hold if this is not the case, but the sequence of beliefs before the first bonus is paid cannot be written in this way, and so the difference equations that are derived would also be different.

can be written as difference equations:

$$\begin{aligned} U^n &= bqp^n + qp^n \left(\Pr[u \leq \delta U^1]\delta U^1 + (1 - \Pr[u \leq \delta U^1])\mathbb{E}[u|u > \delta U^1]\right) \\ &+(1-qp^n) \left(\Pr[u \leq \delta U^{n+1}]\delta U^{n+1} + (1 - \Pr[u \leq \delta U^{n+1}])\mathbb{E}[u|u > \delta U^{n+1}]\right), \text{ for all } n. \end{aligned}$$

The first term is the probability that the agent will receive a bonus multiplied the value of the bonus. The second term on the first line is the continuation value of the agent when he gets a bonus. If he continues to the next period his belief will be p^1. There are two possibilities, his outside option is below the (discounted) expected value and so he will stay at the firm; and that his outside option is better than staying and so he quits. The final line is similar to the second term on the first line, but the case where the agent does not get paid a bonus, meaning his belief falls from p^n to p^{n+1}. This simplifies to

$$U^n = bqp^n + qp^n \left(\delta F^1 U^1 + (1 - F^1)I^1\right) + (1-qp^n) \left(\delta F^{n+1} U^{n+1} + (1 - F^{n+1})I^{n+1}\right), \tag{2.2}$$

where $F^n \equiv F(\delta U^n)$ and $I^n \equiv \mathbb{E}[u|u > \delta U^n]$.[8]

In order to ensure that the strategy profile described above is an equilibrium, it must be that the principal is best responding by choosing $b_t = b$ if and only if $\lambda_t = 1$ and $\theta_t = H$ for any belief $p_t \in \left\{p^1, p^2, ..., p^\infty\right\}$. Before writing the principal's incentive constraints, the principal's value for employing an agent with given productivity and belief in the candidate equilibrium must be defined. Define V_θ^n as the principal's expected value from period t onwards, given the agent has productivity $\theta_t = \theta$ and belief $p_t = p^n$, but before the principal learns λ_t. These values satisfy the difference equations:

$$\begin{aligned} V_H^n &= v - bq + qF^1\delta(\rho V_H^1 + (1-\rho)V_L^1) + (1-q)F^{n+1}\delta(\rho V_H^{n+1} + (1-\rho)V_L^{n+1}), \\ V_L^n &= F^{n+1}\delta(\rho V_L^{n+1} + (1-\rho)V_H^{n+1}), \text{ for all } n. \end{aligned} \tag{2.3}$$

Notice that in the first equation, since the agent has high productivity, the principal enjoys a payoff of v, and with probability q pays the bonus of value b. In contrast, in the second equation, these benefits and costs do not occur. The continuation probability and payoffs depend on the agent's beliefs, which is determined by whether or not the agent received a bonus. In the first equation, where it is possible that the agent received the bonus, the belief in the following

[8] The parameters must be such that for some n, $\delta U^n < 1$ so that the agent eventually quits the firm with some probability. For n where $\delta U^n > 1$, I^n is undefined, in such a case define $I^n \equiv 1$ (since $1 - F^n = 0$ this is without loss).

period is either p^1 or p^{n+1} depending on whether or not the agent received the bonus. The two terms are the probability of the agent staying at the firm and the (discounted) expected continuation value of the principal in each of the two cases. The second equation only has the one case where the agent does not receive a bonus.

The incentive constraints of the principal ensure that when she sees the agent has high productivity and she can pay a bonus, she does, and that when she sees that the agent has low productivity she never pays a bonus. These need to be satisfied for all possible beliefs that the agent can have. The first set of constraints are

$$-b + F^1\delta(\rho V_H^1 + (1-\rho)V_L^1) \geq F^{n+1}\delta(\rho V_H^{n+1} + (1-\rho)V_L^{n+1}), \text{ for all } n.$$

The right hand side is the value when the principal deviates and doesn't pay a bonus when the agent has high productivity. Note that because the transitions of θ_t are Markovian and the cost shocks λ_t are i.i.d., the principal does not retain any instrumental private information. This is the reason that the continuation value functions (V_H^{n+1} and V_L^{n+1}) remain the same off the equilibrium path—as on the right hand side of the constraint. The constraints can be simplified to

$$F^1(\rho V_H^1 + (1-\rho)V_L^1) - F^{n+1}(\rho V_H^{n+1} + (1-\rho)V_L^{n+1}) \geq b/\delta, \text{ for all } n. \quad (2.4)$$

The second set of constraints are

$$F^{n+1}\delta(\rho V_L^{n+1} + (1-\rho)V_H^{n+1}) \geq -b + F^1\delta(\rho V_L^1 + (1-\rho)V_H^1), \text{ for all } n.$$

These can be simplified to

$$F^1(\rho V_L^1 + (1-\rho)V_H^1) - F^{n+1}(\rho V_L^{n+1} + (1-\rho)V_H^{n+1}) \leq b/\delta, \text{ for all } n. \quad (2.5)$$

In order to focus on the relevant constraint, some properties of the value functions need to be understood. It is not possible to solve the difference equations 2.2 and 2.3, that determine U^n, V_H^n and V_L^n, since there is an infinite set of equations. However, showing that the value functions are monotonic, means that two constraints will be sufficient to satisfy all constraints.

First, it can be shown that the sequence $\{U^n\}_{n=0}^{n=\infty}$ is decreasing. Intuitively this is because lower beliefs for the agent, correspond to lower expected values for the agent from future bonus payments.

Lemma 15. *$U^n > U^{n+1}$ for all n.*

Although this result is intuitive, it does not follow directly from the difference equation 2.2. The proof writes the problem with the belief of the agent as a continuous state variable. The agent's maximisation problem (taking the principal's strategy as given) can then be written as an operator that is shown to be a contraction. It follows from the contraction mapping theorem that the agent's value function has a unique fixed point. It can then be shown that this is a strictly increasing function in the belief of the agent.

Define $\tilde{V}_H^n \equiv \rho V_H^n + (1-\rho) V_L^n$. This is the expected value of the principal from period t onwards, given that in the previous period the agent had productivity $\theta_{t-1} = H$ and in the current period has belief $p_t = p^n$, and also that neither the current productivity (θ_t) nor the current cost shock (λ_t) have been learned by the principal yet. Intuitively, it must be that $\tilde{V}_H^n$ is decreasing in n. This is because conditional on the productivity of the agent, the principal always wants the agent to have a higher belief since this will make them more likely to stay at the firm.

Lemma 16. *$\tilde{V}_H^n > \tilde{V}_H^{n+1}$ for all n.*

The proof again writes the problem with a continuous state variable and by applying the contraction mapping theorem, it shows that the unique pair of value functions for the principal must be strictly increasing in the belief of the agent.

Now attention can be restricted to just two of the principal's incentive constraints. In particular, I show that if the value functions are monotonic, constraint 2.4 with $n = 1$ and constraint 2.5 with $n = \infty$ are sufficient for all incentive constraints to be satisfied—i.e. they ensure that the respective constraints for $n > 1$ and $n < \infty$ are satisfied. This is shown formally in the proof of Proposition 14. For constraint 2.4, the intuition is as follows. When a bonus was paid in the previous period ($n = 1$), the agent's beliefs are at their maximum and so the principal has the least to gain from not paying a bonus in the next period—the beliefs will only go down to p^2 rather than staying at p^1 if a bonus is not paid. So if she is incentivised to pay a bonus when $n = 1$, she will also be incentivised when beliefs are lower, i.e. when $n > 1$. The intuition is reversed for constraint 2.5.

Proposition 14. *There exist parameter values $(F(\cdot), v, \rho, q, \delta, b)$ for which there is an equilibrium with the strategies described above. This equilibrium has the following properties: the probability that the agent will stay ($a_t^A = S$) is always increasing in his belief that $\theta_t = H$; and given the agent's beliefs, the principal always prefers to retain a high rather than low productivity agent.*

This result establishes the existence of an equilibrium where the continuation value for the agent is increasing in his belief. These two properties match the reduced-form assumptions made in the static model of Chapter 2. Note that this equilibrium is not necessarily unique (as was established under sufficient conditions in Chapter 2). The proof involves bounding the value functions to find sufficient conditions for the two incentive constraints that need to be satisfied. Parameter values can then be found that satisfy these conditions.

I am unable to characterise the set of parameter values for which this equilibrium exists. However, it is clear that both v and ρ must have intermediate values. If v is too high then the value of retaining even the low productivity agent would be sufficiently high meaning the principal would deviate and pay a bonus to a low productivity agent, similarly if v is too low then the principal would have no incentive to pay a bonus to retain even a high productivity agent. If ρ is very high (close to 1), then once the agent knows he has high productivity it is very likely that he will continue to have high productivity and so if the agent has a high belief the principal will have an incentive to deviate. If ρ is very low (close to 1/2), then there is not much more incentive for the principal to keep a currently high productivity agent compared to a low productivity agent since it is not that much more likely he will be of high productivity in the future.

3 Dynamic model with (partial) transparency

In this subsection I continue to analyse the interaction of a principal and a single agent in a dynamic relationship, and in order to capture the notion of transparency, I assume that there is a 'passive' second agent whose bonus the active agent can observe. I do not model the interaction between the principal and the passive agent explicitly. Instead, I assume that the principal pays a bonus to the passive agent if and only if $\lambda_t = 1$—this is as if the passive agent had a fixed belief of $p = 1$ (or productivity $\theta_t = H$ for all t) and the principal plays the same strategy as with a single agent.[9] All other features of the model remain unchanged.

As before, I construct an equilibrium in which, in the period t stage game, when the agent has a belief p_t:

- For any p_t, the principal pays a bonus $b_t = b$ if and only if $\theta_t = H$ and $\lambda_t = 1$;

[9]Fixing $p = 1$ demonstrates the effect of transparency in the simplest way possible. The analysis remains similar if the agent had a fixed belief $p \in (0, 1)$. I discuss this in more detail at the end of the subsection. If $p = 0$ the setting is identical to the single agent setting.

- The agent chooses to stay at the firm $a_t^A = S$ if and only if his expected value of staying at the firm is greater than quitting the firm, which is the case when

$$\delta \hat{U}(p_{t+1}(p_t, b_t, b_t^2)) \geq u_t,$$

where $\hat{U}(p)$ is the (expected) value for the agent of having a belief p, b_t^2 denotes the bonus paid to the passive agent and $p_{t+1}(p_t, b_t, b_t^2) = \Pr[\theta_{t+1}|p_t, b_t, b_t^2]$ is the belief in the next period given the current belief and bonuses in the current period (p_t, b_t, b_t^2).

The belief in period t is updated twice to get the belief in period $t+1$. First, the agent observes b_t and b_t^2 and updates using Bayes rule. Define $\hat{p}_t(b_t, b_t^2) \equiv \Pr[\theta_t = H|b_t, b_t^2, p_t]$. Given the strategy described above, unlike under no transparency, there are now three cases:

$$\begin{aligned}
\hat{p}_t(b, b_t^2) &= 1 \text{ for any } b_t^2, \\
\hat{p}_t(0, 0) &= p_t, \\
\hat{p}_t(0, b) &= 0.
\end{aligned}$$

Note in the second case, since it must be that $\lambda = \infty$, the agent does not update his belief when he receives no bonus.

Second, as before, this posterior belief undergoes the Markov transition that dictates how the productivity evolves between periods. So

$$p_{t+1} = \rho \hat{p}_t + (1 - \rho)(1 - \hat{p}_t).$$

It will be helpful to define two sequences of beliefs $\{p^n\}_{n=1}^{n=\infty}$ and $\{p^n\}_{n=-1}^{n=-\infty}$ where

$$\begin{aligned}
p^1 &= \rho, \\
p^{n+1} &= p^n \rho + (1 - p^n)(1 - \rho) \text{ for } n \geq 1;
\end{aligned}$$

and

$$\begin{aligned}
p^{-1} &= 1 - \rho, \\
p^{-n-1} &= p^{-n} \rho + (1 - p^{-n})(1 - \rho) \text{ for } n \geq 1.
\end{aligned}$$

The first sequence starts with the most optimistic belief, $p_t = p^1 = \rho$, that occurs following a bonus in the previous period. Each step represents the decline in the belief following no bonus for either agent. Note that this is less steep than

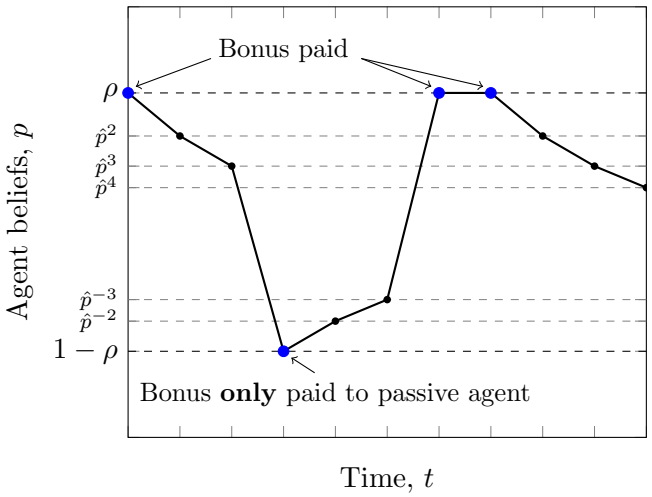

Figure 3.2: *An example of agent's belief path with a second (passive) agent.*

with a single agent/no transparency, since here it is definitely the case that $\lambda_t = \infty$. The sequence converges to $p^{\infty} = \frac{1}{2}$. The second sequence starts with the most pessimistic belief, $p_t = p^{-1} = 1 - \rho$, that occurs following no bonus for the agent and a bonus for the passive agent in the previous period. Each step represents the increase in beliefs following no bonus for either agent, and again this converges to $p^{-\infty} = \frac{1}{2}$. I illustrate an example of a path of the agent's beliefs in Figure 3.2—note how this contrasts with Figure 3.1.

Proposition 15. *In the partial transparency game, there exist parameter values $(F(\cdot), v, \rho, q, \delta, b)$ for which there is an equilibrium with the strategies described above. This equilibrium has the following properties: the probability that the agent will stay ($a_t^A = S$) is always increasing in his belief that $\theta_t = H$; and given the agent's beliefs, the principal always prefers to retain a high rather than low productivity agent.*

The result is exactly the same as with a single agent, and there are parameter values for which the equilibrium exists in both settings. In the proof, I construct the equilibrium in the same way as before: I show that the value functions of the principal and agent are monotonic (i.e. higher beliefs lead to higher values for the players), and then I bound the incentive constraints of the principal to ensure that the strategy above is indeed an equilibrium.

4 Discussion

It is not possible to derive closed form expressions for the principal's value functions. In order to understand if the comparative static results in Chapter 2 (Section 3) continue to hold, I approximate the value functions by value function iteration. Then, I compute the difference between partial transparency and no transparency as v—the value to the principal of having a high productivity agent—varies. Figure 3.4 (in Appendix 2) shows the results for a set of parameters for which the equilibrium exists in both settings. The results are in line with the comparative statics in the static game—as v increases, transparency becomes relatively more favourable for the principal.[10,11] As in the static model, the intuition for this is because under transparency following no bonus, there are two possible beliefs that can be induced—a more optimistic belief when the other agent is also not paid a bonus and a more pessimistic belief when the other agent is paid a bonus. The posterior beliefs under transparency and under no transparency are illustrated in Figure 3.3.[12] The high productivity agent can never get the more pessimistic belief $(1-\rho)$. This means when no bonus is paid, he will always be less pessimistic under transparency compared to no transparency. This makes him more likely to stay at the firm which becomes more beneficial for the firm as v increases.[13,14]

I have analysed a game where the passive agent is always paid a bonus if $\lambda = 1$—in effect has a belief $p = 1$. Now, I discuss briefly how the analysis changes when $p \in (0,1)$. The main difference is that if both agents receive no bonus $b = (0,0)$, it is now possible that $\lambda = 1$, meaning that the agent's belief changes following the bonuses with

$$\hat{p}_t(0,0) = \frac{(1-q)p_t}{(1-q)+q(1-p_t)(1-p)}.$$

The sequence of beliefs, including their limit point p^∞, are now different. In particular, the beliefs fall more sharply for lower p when both agents are not

[10] Results continue to hold for other parameters for which the equilibria exist in both setups, at beliefs other than $p^1 = \rho$ and also if the value functions $\tilde{V}_H^n$ and $\tilde{\tilde{V}}_H^n$ are used.

[11] The comparative static here is on v rather than on the continuation values g_H^P and g_L^P used in the static model since these are now endogenous.

[12] Note that the way the beliefs update is similar to Figures 2.1 and 2.2 in Chapter 2 that illustrate the case of the static model.

[13] This is more complicated than the static game because the continuation values are endogenised and the agent's productivity can change over time.

[14] Another comparative static of interest is how the difference between transparency and no transparency changes as the persistence parameter, ρ, varies. Numerical results suggest that this is non-monotonic—meaning there is no clear prediction from the model.

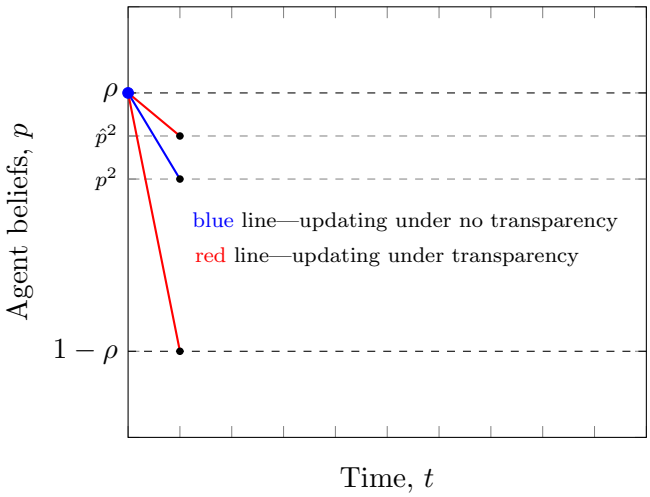

Figure 3.3: *Comparison following no bonus with a single agent (no transparency) and with a second passive agent (transparency).*

paid a bonus. Existence of an equilibrium can be proved in a similar way to Proposition 15.

Analysing the game with two active agents is more complex—the main reason being that there is now a two dimensional state variable (the belief of each of the two agents). From a modelling perspective, what happens after one agent quits would need to be modelled since this affects the continuation value (and consequently the quitting decisions) of the agents—this is not a problem in the setting with a single active agent. A possible solution is to have the agent replaced with an agent holding an identical belief to the agent who quit would have had if he had stayed. In addition, to ensure that the principal has an incentive to retain agents, she would have to incur a cost equal to the continuation value she would have had from that agent if he had stayed. Despite these difficulties, some new insights may arise from the analysis of this setting and so it warrants investigation in future work. I also conjecture that it would be possible to rule out an equilibrium in which, when there is one agent with high and one of low productivity, the principal only pays a bonus to the agent with low productivity.[15]

[15]Recall, in the static game, it was necessary to explicitly rule out such an equilibrium.

Appendix to Chapter 3

1 Proofs

1.1 Proof of Lemma 15

Let $I \equiv [p^{\infty}, \rho]$ be the domain of beliefs—recall that p^{∞} is defined in 2.1. Let $\beta(p)$ be the Bayesian updating formula for the agent's beliefs between period t and $t+1$ when $b_t = 0$, formally defined as

$$\beta(p) \equiv \frac{p(1-q)}{1-pq}\rho + \left(1 - \frac{p(1-q)}{1-pq}\right)(1-\rho).$$

Note that $\beta : I \to I$, that $\beta(\cdot)$ is a strictly increasing function.

Let $\mathcal{I}(I)$ be the set of all functions $h : I \to [0, \bar{u}]$ where $\bar{u} \equiv \max\left\{q\rho b + 1, \frac{q\rho b}{1-\delta}\right\}$. Define the operator $T : \mathcal{I} \to \mathcal{I}$ as

$$\begin{aligned} T \circ h(p) \equiv \max_{x,y \in [0,1]} & qp\left[b + F(x)\delta h(\rho) + (1 - F(x))J(x)\right] \\ & + (1-qp)\left[F(y)\delta h(\beta(p)) + (1-F(y))J(y)\right], \end{aligned}$$

where $J(x) \equiv \mathbb{E}[u|u > x]$ is the expected value of the outside option if the agent decides to quit for all $u > x$. Here the agent's decision over whether to stay or quit is given by the choice variables x and y. These define the cutoffs where if the outside option is above this level the agent quits in the case that a bonus was paid and not paid respectively. $h(\cdot)$ is the continuation value of the agent given his belief at the start of the period.

The operator $T : \mathcal{I} \to \mathcal{I}$ is a contraction on the 'sup norm' vector space. This can be verified with Blackwell's sufficient conditions for a contraction (see Stokey et al. (1989)). The two conditions (with notation appropriately adapted) are:

- Monotonicity: for $h, h' \in \mathcal{I}$, if $h(p) \leq h'(p)$ for all $p \in I$ then $T \circ h(p) \leq T \circ h'(p)$ for all $p \in I$;

- Discounting: there exists some $\delta' \in (0,1)$ such that

$$[T \circ (h+a)](p) \leq (T \circ h)(p) + \delta' a$$

for all $h \in \mathcal{I}$, $p \in I$ and $a \geq 0$.

Monotonicity follows immediately from the definition of the operator T. To verify discounting it is sufficient to show

$$\begin{aligned}
[T \circ (h+a)](p) = \max_{x,y \in [0,1]} & qp\left[b + F(x)\delta(h+a)(\rho) + (1-F(x))J(x)\right] \\
& + (1-qp)\left[F(y)\delta(h+a)(\beta(p)) + (1-F(y))J(y)\right] \\
\leq \max_{x,y \in [0,1]} & qp\left[b + F(x)\delta h(\rho) + (1-F(x))J(x)\right] \\
& + (1-qp)\left[F(y)\delta h(\beta(p)) + (1-F(y))J(y)\right] + qp\delta a + (1-qp)\delta a \\
\leq & (T \circ h)(p) + \delta a,
\end{aligned}$$

where in the second line x and y have been taken to be 1 in the part that multiplies a inside the brackets on the first line.

It follows from the contraction mapping theorem (again, see Stokey et al. (1989)), that there exists a unique function $v \in \mathcal{I}$ such that

$$\begin{aligned}
v(p) = \max_{x,y \in [0,1]} & qp\left[b + F(x)\delta v(\rho) + (1-F(x))J(x)\right] \\
& + (1-qp)\left[F(y)\delta v(\beta(p)) + (1-F(y))J(y)\right].
\end{aligned}$$

Next, I show that $v(\cdot)$ must be a strictly increasing function. First, observe that the operator T maps non-decreasing functions to non-decreasing functions. To see this consider for $p, p' \in I$ with $p > p'$

$$\begin{aligned}
T \circ v(p) - T \circ v(p') &= qpA + (1-qp)B(p) - qp'A - (1-qp')B(p') \\
&\geq q(p-p')(A - B(p)) \\
&\geq q(p-p')b \\
&> 0,
\end{aligned}$$

where

$$A \equiv \max_{x\in[0,1]} b + F(x)\delta v(\rho) + (1-F(x))J(x),$$
$$B(p) \equiv \max_{y\in[0,1]} F(y)\delta v(\beta(p)) + (1-F(y))J(y).$$

The first inequality follows since $B(\cdot)$ is non-decreasing if $h(p)$ is non-decreasing; and the second inequality follows since $A \geq b + B(p)$ if $h(p)$ is non-decreasing.

Next, note that the set of non-decreasing functions is a closed subset of $\mathcal{I}$. By Corollary 1 of Stokey et al. (1989) (p.52), the fixed point $v(\cdot)$ must be an non-decreasing function.[16] Finally, note that T maps non-decreasing functions to strictly increasing functions (a subset of the set of non-decreasing functions), and thus the fixed point $v(\cdot)$ must be a strictly increasing function.

It follows that since $p^n > p^{n+1}$ for all n that $U^n > U^{n+1}$ for all n.

1.2 Proof of Lemma 16

As in the proof of Lemma 15, let $I \equiv [p^\infty, \rho]$ be the domain of beliefs and let $\beta(p)$ be the Bayesian updating formula for the agent's beliefs between period t and $t+1$ when $b_t = 0$.

Let $\hat{\mathcal{I}}(I)$ be the set of all (bounded) functions $h : I \to [0, \frac{v}{1-\delta}]^2$. Define the operator $\hat{T}$:

$$\hat{T}\circ\begin{bmatrix} h_1(p) \\ h_2(p) \end{bmatrix} = \begin{bmatrix} \max_{x\in\{0,1\}} \Big\{ v + q\Big(\mathbb{1}[x=1](-b + (F(\delta U(\rho))\delta\big(\rho h_1(\rho) + (1-\rho)h_2(\rho)\big) \\ +(1-\mathbb{1}[x=1])(F(\delta U(\beta(p)))\delta\big(\rho h_1(\beta(p)) + (1-\rho)h_2(\beta(p))\big)\Big) \\ +(1-q)\Big((F(\delta U(\beta(p)))\delta\big(\rho h_1(\beta(p)) + (1-\rho)h_2(\beta(p))\big)\Big)\Big\}, \\ \max_{y\in\{0,1\}} \Big\{ 0 + q\Big(\mathbb{1}[y=1](-b + (F(\delta U(\rho))\delta\big(\rho h_2(\rho) + (1-\rho)h_1(\rho)\big) \\ +(1-\mathbb{1}[y=1])(F(\delta U(\beta(p)))\delta\big(\rho h_2(\beta(p)) + (1-\rho)h_1(\beta(p))\big)\Big) \\ +(1-q)\Big((F(\delta U(\beta(p)))\delta\big(\rho h_2(\beta(p)) + (1-\rho)h_1(\beta(p))\big)\Big)\Big\} \end{bmatrix} \tag{1.1}$$

Here the principal's decision over whether to pay a bonus or not after observing that the agent has high (low) productivity is given by x (y), where 1 represents a bonus, and 0 no bonus. $h_1(\cdot)$ and $h_2(\cdot)$ are the value functions of

[16] The proof is in their text. Intuitively start from some $f_0 \in \mathcal{I}'$, where $\mathcal{I}' \subset \mathcal{I}$ is the set of non-decreasing functions. Then repeatedly applying the operator T will get to the fixed point $v(\cdot)$, and this must also be non-decreasing since the set of non-decreasing functions is closed.

the principal in the case that the agent is of high and low productivity.[17] $U(\cdot)$ denotes the agent's value function, and this affects the probability that he will quit or not for a given belief.

I proceed as in the proof of Lemma 15. First, I show that $\hat{T}$ is a contraction mapping. Then, I show that the pair of value functions must be strictly increasing.

In order to apply the contraction mapping theorem, a metric space in which a function that maps beliefs into two dimensions needs to be defined. A suitable metric is the 'sup-sup' norm formally given by:

$$\|h\| = \sup_{p \in I} \left(\sup_{i \in \{1,2\}} |h_i(p)| \right).$$

It is straightforward to verify that this norm satisfies the required properties and that ensure that it is a normed vector space (p.45 Stokey et al. (1989)). It is also possible to extend Theorem 3.1 in Stokey et al. (1989) to show that set of continuous functions $h : I \to [0, \frac{v}{1-\delta}]^2$ with the sup-sup norm forms a complete normed vector space.[18]

To show that the operator $\hat{T}$ is a contraction mapping, Blackwell's sufficient conditions can be extended to two dimensions as follows:

- Monotonicity: for $h, h' \in \hat{\mathcal{I}}$, $h_i(p) \leq h'_i(p)$ for all $p \in I$, $i \in \{1, 2\}$ then $\hat{T} \circ h_i(p) \leq \hat{T} \circ h_i(p)$ for all $p \in I$;
- Discounting: there exists some $\delta' \in (0, 1)$ such that

$$\left[\hat{T} \circ (h + a \begin{bmatrix} 1 \\ 1 \end{bmatrix}) \right] (p) \leq (\hat{T} \circ h)(p) + \delta' a \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

  for all $h \in \hat{\mathcal{I}}$, $p \in I$ and $a \geq 0$.

It is straightforward from 1.1 that these conditions are satisfied. So, from the contraction mapping theorem, it follows that there is a unique value function for the principal. Next, note that the subset of $\hat{\mathcal{I}}(I)$ of functions that are non-decreasing in both dimensions is closed.[19] Using the same Corollary as in the proof of Lemma 15, since in both dimensions the operator maps non-decreasing to non-decreasing functions the value function is non-decreasing in both dimensions. To verify this recall from Lemma 15 that $U(\cdot)$ is strictly increasing. Since

[17] Note that these represent V_H^n and V_L^n rather than $\tilde{V}_H^n$ and $\tilde{V}_L^n$ in the main text.

[18] In the proof: $|h_n(x) - h_m(x)|$ can be replaced with $\sup_{i \in \{1,2\}} |h_{ni}(x) - h_{mi}(x)|$; and continuity can be shown in each dimension of the limiting function in turn.

[19] The product of two closed sets is also a closed set.

$U(\cdot)$ and $\beta(\cdot)$ are both strictly increasing it follows that if $h_1(\cdot)$ and $h_2(\cdot)$ are both non-decreasing, then

$$\left(\hat{T} \circ \begin{bmatrix} h_1(\cdot) \\ h_2(\cdot) \end{bmatrix}\right)_1 \text{ and } \left(\hat{T} \circ \begin{bmatrix} h_1(\cdot) \\ h_2(\cdot) \end{bmatrix}\right)_2$$

are also both strictly increasing.[20] Finally, since the operator maps a pair of non-decreasing functions to a pair of strictly increasing functions, the fixed point must be strictly increasing.

It follows that since $p^n > p^{n+1}$ for all n that $V_H^n > V_H^{n+1}$ and $V_L^n > V_L^{n+1}$ for all n and so $\tilde{V}_H^n > \tilde{V}_H^{n+1}$ and $\tilde{V}_L^n > \tilde{V}_L^{n+1}$ for all n.

1.3 Lemma 17 and proof

Lemma 17. *$\tilde{V}_H^n > \tilde{V}_L^n$ for all n.*

From the proof of Lemma 16, there is a unique pair of value functions for the principal that satisfy the operator in 1.1. It is straightforward to show that the operator $\hat{T}$ maps functions where $h_1(p) \geq h_2(p)$ for all p, to functions where

$$\left(\hat{T} \circ \begin{bmatrix} h_1(\cdot) \\ h_2(\cdot) \end{bmatrix}\right)_1 > \left(\hat{T} \circ \begin{bmatrix} h_1(\cdot) \\ h_2(\cdot) \end{bmatrix}\right)_2 \text{ for all } p.$$

The subset of $\hat{\mathcal{I}}(I)$ where the first output is weakly greater than the second output is a closed subset of the set of $\hat{\mathcal{I}}(I)$. Thus by the Corollary of the contraction mapping theorem used above, the fixed point that is the principal's pair of value functions must have this property. Finally, as before, since the output from the operator $\hat{T}$ has the first output being *strictly* greater than the second output, the fixed point must also have this property.

It follows that since $\tilde{V}_H^n > \tilde{V}_L^n$ for all n.

1.4 Proof of Proposition 14

To prove this result, I provide sufficient conditions for the two incentive constraints 2.4 and 2.5 to hold, under the assumption that $F \sim U[0,1]$. I start with 2.4, this incentive constraint that ensures that when the agent has high productivity the principal will want to always pay a bonus (assuming she can, i.e. that $\lambda_t = 1$). Define $\tilde{p}^n \equiv qp^n$.

Lemma 18. *If $F \sim U[0,1]$ and $\tilde{p}^1 - \tilde{p}^2 \geq \frac{1}{\rho v \delta^2}$ then 2.4 is satisfied for all n.*

[20] This can easily been seen by computing the difference between the output of the operator for $p, p' \in I$ with $p > p'$ as before.

Proof. With $F \sim U[0,1]$, 2.4 becomes

$$U^1\tilde{V}_H^1 - U^n\tilde{V}_H^n \geq b/\delta^2, \text{ for all } n \geq 2. \tag{1.2}$$

Since $U^n > U^{n+1}$ and $\tilde{V}^n > \tilde{V}^{n+1}$ (and so $U^2 \geq U^n$ and $\tilde{V}^2 \geq \tilde{V}^n$ for all $n \geq 2$), if

$$U^1\tilde{V}_H^1 - U^2\tilde{V}_H^2 \geq b/\delta^2,$$

then 2.4 is satisfied.

Next, from 2.2

$$\begin{aligned} U^1 - U^2 &= b(\tilde{p}^1 - \tilde{p}^2) + \frac{1}{2}\tilde{p}^1(\delta U^1)^2 + \frac{1}{2}(1-\tilde{p}^1)(\delta U^2)^2 - \frac{1}{2}\tilde{p}^2(\delta U^1)^2 - \frac{1}{2}(1-\tilde{p}^2)(\delta U^3)^2, \\ &= b(\tilde{p}^1 - \tilde{p}^2) + \frac{1}{2}(1-\tilde{p}^2)\left((\delta U^1)^2 - (\delta U^3)^2\right) - \frac{1}{2}(1-\tilde{p}^1)\left((\delta U^1)^2 - (\delta U^2)^2\right), \\ &> b(\tilde{p}^1 - \tilde{p}^2), \end{aligned}$$

where the final line follows from $\tilde{p}^1 > \tilde{p}^2$ and $U^2 > U^3$.

Finally, it must be that for all n

$$\begin{aligned} \tilde{V}_H^n &\geq \rho V_H^n, \\ &\geq \rho v, \end{aligned}$$

where the final line is because even if the bonus could be paid, the principal's expected value from paying the bonus today, must be less than the future (discounted) benefit. This means that

$$\begin{aligned} U^1\tilde{V}_H^1 - U^2\tilde{V}_H^2 &\geq (U^1 - U^2)\tilde{V}_H^2, \\ &\geq (U^1 - U^2)\rho v. \end{aligned}$$

So if

$$b(\tilde{p}^1 - \tilde{p}^2)\rho v \geq b/\delta^2,$$

then 2.4 is satisfied. The result follows. □

Now I provide sufficient conditions for 2.5: this incentive constraint that ensures that when the agent has low productivity the principal will never want to pay a bonus. I start by defining some combinations of the parameters which will be used in the result.

Define

$$\bar{U}^n \equiv \frac{(1 - \sqrt{1 - (2b\tilde{p}^n + 1)\delta^2}}{\delta^2}.$$

Define

$$\bar{V}_H^1 \equiv \frac{(1-\delta\rho)(v-bq)}{(1-\delta)(1+\delta-2\delta\rho)},$$
$$\bar{V}_L^1 \equiv \frac{\delta(1-\rho)(v-bq)}{(1-\delta)(1+\delta-2\delta\rho)}.$$

Lemma 19. *If $F \sim U[0,1]$ and*

$$\bar{U}^1(\rho\bar{V}_L^1 + (1-\rho)\bar{V}_H^1) - \bar{U}^\infty(\rho v\delta^2(1-\rho)\bar{U}^\infty + (1-\rho)v) \leq b/\delta^2$$

then 2.5 is satisfied for all n.

Proof. With $F \sim U[0,1]$, 2.5 becomes

$$U^1(\rho V_L^1 + (1-\rho)V_H^1) - U^{n+1}(\rho V_L^{n+1} + (1-\rho)V_H^{n+1}) \leq b/\delta^2, \text{ for all } n. \quad (1.3)$$

As in the previous result, I will provide an upper bound on the positive (first part) of the LHS and a lower bound on the negative part (second part).

First note that

$$\begin{aligned} U^1 &= b\tilde{p}^1 + \frac{1}{2} + \frac{1}{2}\tilde{p}^1(\delta U^1)^2 + \frac{1}{2}(1-\tilde{p}^1)(\delta U^2)^2, \\ &\leq b\tilde{p}^1 + \frac{1}{2} + \frac{1}{2}(\delta U^1)^2, \end{aligned}$$

where the inequality is due to $U^1 > U^2$. Solving this quadratic inequality gives

$$U^1 \leq \bar{U}^1. \quad (1.4)$$

Next note that

$$U^n > \bar{U}^\infty \quad (1.5)$$

for all $n \geq 2$. This is because

$$\begin{aligned} U^\infty &= b\tilde{p}^\infty + \frac{1}{2} + \frac{1}{2}\tilde{p}^\infty(\delta U^1)^2 + \frac{1}{2}(1-\tilde{p}^\infty)(\delta U^\infty)^2, \\ &\geq b\tilde{p}^\infty + \frac{1}{2} + \frac{1}{2}(\delta U^\infty)^2, \end{aligned}$$

and so $U^\infty > \bar{U}^\infty$ and $U^n > U^\infty$ for all n. Next, it must be that

$$\begin{aligned} V_H^1 &\leq v - bq + \delta\left(\rho V_H^1 + (1-\rho)V_L^1\right), \\ V_L^1 &\leq \delta\left(\rho V_L^1 + (1-\rho)V_H^1\right). \end{aligned}$$

These both represent the best case for the principal, when the agent always stays to the next period and remains at the most optimistic level (p^1) (since $V_\theta^1 > V_\theta^2$). Solving this system gives

$$V_H^1 \leq \bar{V}_H^1, \tag{1.6}$$

$$V_L^1 \leq \bar{V}_L^1. \tag{1.7}$$

Finally, it must be that for all $n \geq 2$

$$V_H^n \geq v, \tag{1.8}$$

$$V_L^n \geq (\delta U^\infty)(\delta(1-\rho)v) \geq \delta^2(1-\rho)v\bar{U}^\infty. \tag{1.9}$$

The first inequality is from the fact that the principal could just receive today's flow payoff of v, not pay a bonus, and not receive any possible continuation payoff from the relationship. The second inequality follows since δU^∞ is the smallest possible probability that the agent stays and $\delta(1-\rho)v$ is the expected payoff that the principal can guarantee in the next period (if she never pays a bonus and the relationship never continues).

The result follows from applying 1.4, 1.5, 1.6, 1.7, 1.8 and 1.9 to 2.5.

□

For the parameter values $v = 7, \rho = .9, q = .99, \delta = .5, b = .5$ there is an equilibrium with the principal paying a bonus in period t only if $\theta_t = H$ and $\lambda_t = 1$.

Finally, by Lemma 16 the probability that the agent will stay ($a_t^A = S$) is increasing in his belief that $\theta_t = H$ and given the agent's beliefs; and by Lemma 17 the principal prefers to retain a high rather than low productivity agent.

1.5 Proof of Proposition 15

As with a single agent, the principal and the(active) agent have value functions depending on the state (agent's belief) and the realisation of the agent's productivity. Use 'hatted' variables to denote these value functions, i.e. the value functions of the principal are given by $\hat{V}_H^n$ and $\hat{V}_L^n$, and the value function of the agent is given by $\hat{U}^n$.

Analogous results to Lemma 15, 16—that show monotonicity of the value functions—can be derived. In Lemma 15, the main difference is that the operator T is defined differently since the agent needs to choose a cutoff strategy in each of three outcomes—when he receives a bonus, when neither agent receives a bonus

and when he doesn't receive a bonus and the other agent does. Formally the operator becomes:

$$T' \circ h(p) \equiv \max_{x,y,z \in [0,1]} qp\left[b + F(x)\delta h(\rho) + (1 - F(x))J(x)\right]$$
$$+(1-q)\left[F(y)\delta h(\hat{\beta}(p)) + (1 - F(y))J(y)\right] + q(1-p)\left[F(z)\delta h(1-\rho) + (1 - F(z))J(y)\right],$$

where $\hat{\beta}(p) \equiv \rho p + (1-\rho)(1-p)$ is the updating rule in this setting. It follows, using a very similar argument as before, that there is a unique fixed point for the operator and that the fixed point must be a strictly increasing function.

Similarly, for the principal, the operator $\hat{T}$ becomes:

$$\hat{T}' \circ \begin{bmatrix} h_1(p) \\ h_2(p) \end{bmatrix} = \begin{bmatrix} \max_{x \in \{0,1\}} \Big\{ v + q\Big(\mathbb{1}[x=1](-b + (F(\delta U(\rho))\delta\big(\rho h_1(\rho) + (1-\rho)h_2(\rho)\big) \\ +(1 - \mathbb{1}[x=1])(F(\delta U(1-\rho))\delta\big(\rho h_1(1-\rho) + (1-\rho)h_2(\hat{\beta}(p))\big)\Big) \\ +(1-q)\Big((F(\delta U(\hat{\beta}(p)))\delta\big(\rho h_1(\hat{\beta}(p)) + (1-\rho)h_2(\hat{\beta}(p))\big)\Big)\Big\}, \\ \max_{y \in \{0,1\}} \Big\{ 0 + q\Big(\mathbb{1}[y=1](-b + (F(\delta U(\rho))\delta\big(\rho h_2(\rho) + (1-\rho)h_1(\rho)\big) \\ +(1 - \mathbb{1}[y=1])(F(\delta U(1-\rho))\delta\big(\rho h_2(1-\rho) + (1-\rho)h_1(\hat{\beta}(p))\big)\Big) \\ +(1-q)\Big((F(\delta U(\hat{\beta}(p)))\delta\big(\rho h_2(\hat{\beta}(p)) + (1-\rho)h_1(\hat{\beta}(p))\big)\Big)\Big\} \end{bmatrix}$$

where the only difference is that after not paying a bonus when $\lambda = 1$, the agent's belief immediately goes to $1 - \rho$. As before, it follows that there is a unique fixed point that must be strictly increasing in each dimension, and that the first dimension must be strictly greater than the second.

Turning to the incentive constraints of the principal, these are actually much simpler than with a single agent. The reason is that if the principal can pay the agent a bonus (i.e. $\lambda_t = 1$) then the agent's posterior belief does not depend on his belief—the belief must either be ρ or $1 - \rho$ in the next period. In the case of $F \sim U[0,1]$, the two incentive constraints are given by

$$\hat{U}^1 \hat{\tilde{V}}_H^1 - \hat{U}^{-1} \hat{\tilde{V}}_H^{-1} \geq b/\delta^2,$$
$$\hat{U}^1 \hat{\tilde{V}}_L^1 - \hat{U}^{-1} \hat{\tilde{V}}_L^{-1} \leq b/\delta^2.$$

Following a very similar argument to before, a sufficient condition for the first

constraint is

$$(2\rho - 1)\rho q v \geq 1/\delta^2.$$

For the second constraint the sufficient condition is again

$$\bar{U}^1(\rho \bar{V}_L^1 + (1-\rho)\bar{V}_H^1) - \bar{U}^\infty(\rho v \delta^2 (1-\rho)\bar{U}^\infty + (1-\rho)v) \leq b/\delta^2,$$

where everything is defined exactly as before.

These conditions are satisfied for the same parameter values as before ($v = 7, \rho = .9, q = .99, \delta = .5, b = .5$) and so there is an equilibrium with the principal paying a bonus in period t only if $\theta_t = H$ and $\lambda_t = 1$.

2 Numerical results

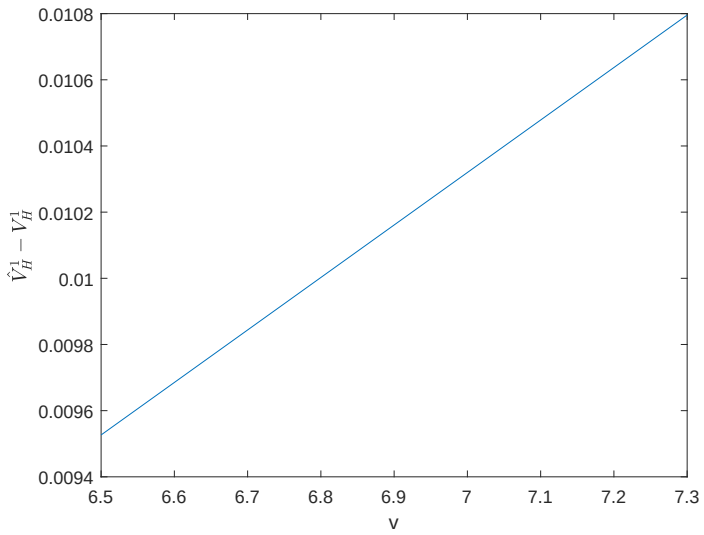

Figure 3.4: *The difference between the value function of the principal (at belief $p = \rho$) under partial transparency ($\hat{V}_H^1$) and no transparency (V_H^1) for different values of v. The other parameter values are $\rho = .9; b = .5; q = .99; \delta = .5$.*

Bibliography

Ali, S. N. and Bénabou, R. (2019). Image Versus Information: Changing Societal Norms and Optimal Privacy. *American Economic Journal: Microeconomics*, (forthcoming).

Amador, M., Werning, I., and Angeletos, G.-M. (2006). Commitment vs. Flexibility. *Econometrica*, 74(2):365–396.

Austen-Smith, D. and Fryer, R. G. (2005). An Economic Analysis of "Acting White". *The Quarterly Journal of Economics*, 120(2):551–583.

Bagwell, K. (2007). Signalling and Entry Deterrence: A Multidimensional Analysis. *The RAND Journal of Economics*, 38(3):670–697.

Baker, M., Halberstam, Y., Kroft, K., Mas, A., and Messacar, D. (2019). Pay Transparency and the Gender Gap. *NBER Working Paper Series*.

Battaglini, M., Bénabou, R., and Tirole, J. (2005). Self-control in peer groups. *Journal of Economic Theory*, 123(2):105–134.

Bewley, T. (1999). *Why Wages Don't Fall during a Recession*. Harvard University Press.

Boleslavsky, R. and Kim, K. (2018). Bayesian Persuasion and Moral Hazard. *working paper*.

Bénabou, R. and Tirole, J. (2002). Self-Confidence and Personal Motivation. *The Quarterly Journal of Economics*, 117(3):871–915.

Bénabou, R. and Tirole, J. (2003). Intrinsic and Extrinsic Motivation. *The Review of Economic Studies*, 70(3):489–520.

Bénabou, R. and Tirole, J. (2004). Willpower and Personal Rules. *Journal of Political Economy*, 112(4):848–886.

Bénabou, R. and Tirole, J. (2006). Incentives and Prosocial Behavior. *American Economic Review*, 96(5):1652–1678.

Card, D., Mas, A., Moretti, E., and Saez, E. (2010). Inequality at Work: The Effect of Peer Salaries on Job Satisfaction. *NBER Working Paper Series*.

Card, D., Mas, A., Moretti, E., and Saez, E. (2012). Inequality at Work: The Effect of Peer Salaries on Job Satisfaction. *American Economic Review*, 102(6):2981–3003.

Carrillo, J. D. and Mariotti, T. (2000). Strategic Ignorance as a Self-Disciplining Device. *The Review of Economic Studies*, 67(3):529–544.

Cho, I.-K. and Kreps, D. M. (1987). Signaling Games and Stable Equilibria. *The Quarterly Journal of Economics*, 102(2):179–221.

CIPD (2017). Reward Management: Focus on Pay. *Chartered Institute of Personnel and Development*.

Cullen, Z. and Pakzad-Hurson, B. (2018). Equilibrium Effects of Pay Transparency. *working paper*.

Cullen, Z. and Perez-Truglia, R. (2018). How Much Does Your Boss Make? The Effects of Salary Comparisons. *working paper*.

Ederer, F. (2010). Feedback and Motivation in Dynamic Tournaments. *Journal of Economics & Management Strategy*, 19(3):733–769.

Engers, M. (1987). Signalling with Many Signals. *Econometrica*, 55(3):663–674.

Esteban, J. and Ray, D. (2006). Inequality, Lobbying, and Resource Allocation. *American Economic Review*, 96(1):257–279.

Farragut, J. and Rodina, D. (2017). Inducing Effort through Grades. *working paper*.

Fehr, E. and Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Frankel, A. and Kartik, N. (2019). Muddled Information. *Journal of Political Economy*, 127(4):1739–1776.

Fuchs, W. (2015). Subjective Evaluations: Discretionary Bonuses and Feedback Credibility. *American Economic Journal: Microeconomics*, 7(1):99–108.

Fudenberg, D. and Tirole, J. (1991). *Game Theory*. MIT Press.

Galperti, S. (2015). Commitment, Flexibility, and Optimal Screening of Time Inconsistency. *Econometrica*, 83(4):1425–1465.

Halac, M., Kartik, N., and Liu, Q. (2017). Contests for Experimentation. *Journal of Political Economy*, 125(5):1523–1569.

Hansen, S. E. (2013). Performance Feedback with Career Concerns. *The Journal of Law, Economics, and Organization*, 29(6):1279–1316.

Hastings, D. F. (1999). Lincoln Electric's Harsh Lessons from International Expansion. *Harvard Business Review*, (May–June 1999).

Ivanov, M. (2015). Optimal Signals in Bayesian Persuasion Mechanisms. *working paper*.

Jehiel, P. (2015). On Transparency in Organizations. *The Review of Economic Studies*, 82(2):736–761.

Kamenica, E. (2008). Contextual Inference in Markets: On the Informational Content of Product Lines. *The American Economic Review*, 98(5):2127–2149.

Kamenica, E. and Gentzkow, M. (2011). Bayesian Persuasion. *The American Economic Review*, 101(6):2590–2615.

Kolotilin, A., Mylovanov, T., Zapechelnyuk, A., and Li, M. (2017). Persuasion of a Privately Informed Receiver. *Econometrica*, 85(6):1949–1964.

Kremer, I., Mansour, Y., and Perry, M. (2014). Implementing the "Wisdom of the Crowd". *Journal of Political Economy*, 122(5):988–1012.

Laibson, D. (1997). Golden Eggs and Hyperbolic Discounting. *The Quarterly Journal of Economics*, 112(2):443–477.

Lazear, E. P. (2000). Performance Pay and Productivity. *The American Economic Review*, 90(5):1346–1361.

Li, J. and Matouschek, N. (2013). Managing Conflicts in Relational Contracts. *The American Economic Review*, 103(6):2328–2351.

Lizzeri, A., Meyer, M., and Persico, N. (2002). The incentive effects of interim performance evaluations. *Penn CARESS Working Papers*.

MacLeod, W. B. (2003). Optimal Contracting with Subjective Evaluation. *The American Economic Review*, 93(1):216–240.

Mariotti, T., Schweizer, N., and Szech, N. (2018). Information Nudges and Self Control. *working paper*.

Mas, A. (2017). Does Transparency Lead to Pay Compression? *Journal of Political Economy*, 125(5):1683–1721.

Milgrom, P. R. (1981). Good News and Bad News: Representation Theorems and Applications. *The Bell Journal of Economics*, 12(2):380–391.

Myerson, R. B. (1986). Multistage Games with Communication. *Econometrica*, 54(2):323–358.

O'Donoghue, T. and Rabin, M. (1999). Doing It Now or Later. *The American Economic Review*, 89(1):103–124.

Phelan, C. (2006). Public trust and government betrayal. *Journal of Economic Theory*, 130(1):27–43.

Prendergast, C. (1999). The Provision of Incentives in Firms. *Journal of Economic Literature*, 37(1):7–63.

Quinzii, M. and Rochet, J.-C. (1985). Multidimensional signalling. *Journal of Mathematical Economics*, 14(3):261–284.

Rege, M. and Solli, I. (2014). Lagging Behind the Joneses: The Impact of Relative Earnings on Job Separation. *working paper*.

Rodina, D. (2017). Information Design and Career Concerns. *working paper*.

Stokey, N. L., Lucas, Jnr, R. E., and Prescott, E. E. (1989). *Recursive Methods in Economic Dynamics*. Harvard University Press.