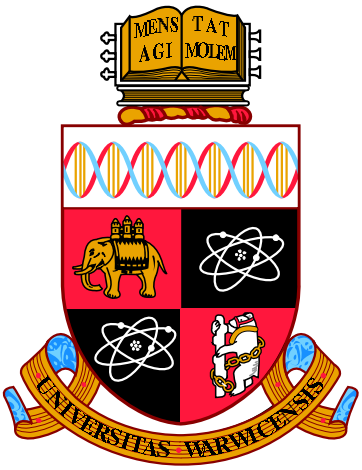


Essays on Economics of Information and Organization

by

Zeinab Aboutalebi

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Economics

September 2019

Contents

List of Figures

Acknowledgments

I want to express my immense gratitude and special appreciation to my supervisors Motty Perry, Ilan Kremer and Jacob (Kobi) Glazer, for their fundamental role in teaching me to seek the economic intuition beyond models and math. Motty has taught me to see economic models not as a set of symbols and equations but as means to explain inefficiencies and phenomena in the real world. With his guidance, I learned to build my ideas and models in their simplest way. He helped me understand if something cannot be shown in a simple context; complexity is not the solution. Ilan and Kobi taught me, what makes an economist a good researcher, is how to interpret and connect intuitions to models and their results. I am also extremely grateful to Robert Akerlof and Debraj Ray for their guidance in my research. My warmest thanks also go to the theory group at the Department of economics for their valuable support and help during my PhD, especially Phil Reny, Herakles Polemarchakis, Dan Bernhardt and Costas Cavounidis. I am thankful to the financial and administrative support of the Department of Economics at University of Warwick and ESRC.

I am also indebted to all those who helped me develop professionally and grow in economics discipline, Bishnupriya Gupta, Gilat Levy, Farshid Vahid, William Coleman, Firouzeh Khalatbari and Fereydoon Tafazoli. Without your support, I would never get here and obtain the ability to do research as a PhD student. Thank you for believing in me more than I believed in myself.

My friends and co-authors at Warwick formed my second family. In them, I found the warmth of friendship and happiness to share, Martina, Federico, Ayush, Gianni and Emama thank you for always being there for me. Daniel, Michella, Akansha thank you for your friendship and kindness. My friends and co-authors are one of the sweetest, most valuable gifts of this PhD.

I now want to thank the most amazing people of my life, my parents, whose love, support and guidance are beyond parenthood. They are and will always be my continued source of inspiration and my driving force to go forward. Maman Parnia,

Baba Hamid, thank you for all you have done for me and most importantly, what you gave me; a strength not to give up ever! I dedicate this thesis to both of you. My sisters, beloved Zahra, you have been a patient listener, closest friend and enthusiastic motivator to me and dearest Reyhanh, you are the joyful voice of wisdom and source of happiness, thank you both for not giving up on me. I also want to thank my grandmother, whose firm believes in women's education, help women in my family to grow. Your house has always been our peaceful second library, thank you Mamani Parvin. I want to pay respect to the memory of my grandfather, who was impatiently waiting to see me graduate but passed away a few months before the end of my PhD. May you rest in peace Babaei. Finally, I want to pay tribute to the memory of two people who were my constant source of strength and courage when I felt weak, and things looked impossible. Amo Majid and Amo Farid, my uncles, thank you!

Declarations

This thesis is submitted to the University of Warwick in accordance with the requirements of the degree of Doctor of Philosophy. I declare that the thesis is my own and original work. Chapter 1 is joint work with Ayush Pant, University of Warwick, who can attest to my significant contribution to the project in terms of the original idea, modeling, model-solving, and writing. Chapters 2 is fully done by me. I also declare that any material contained in this thesis has not been submitted for a degree to any other university.

Zeinab Aboutalebi
September 24, 2019

Abstract

This thesis consists of two essays on economics of information and organization. In general it studies the optimal strategies of acquisition and disclosure of information in different types of relationships within an organization. Information asymmetry shapes the strategic interactions between agents within an organization. Therefore the essays help in obtaining a broader understanding of the role of information in the organizations and the inefficiencies created by its asymmetry in organizations. In chapter One we look at feedback in organizations and study the supervisor's problem. Supervisors face the following tradeoff: while honest feedback encourages employees to discard bad ideas, it can also be demotivating. We obtain three main results. First, the supervisor only gives honest feedback to high self-opinion agents. Second, receiving honest feedback leads high self-opinion agents to exert more effort. Third, overconfidence is potentially welfare improving. In the second chapter, I look at how the incentives to discriminate within an organization induces the managers to manipulate information about subordinates and cause failure in projects. I study a principal manager career concern relationship where manager and principal may not have identical bias toward diversity. In such a setting the misaligned manager faces the following trade off: while hiring from minorities will reduce his utility, not hiring them might cost him his career. I show that when success of employees depends on their ability and manager's effort, positive bias of the principal induces sabotage of minority groups. If the principal has no bias toward diversity, diversity marginally improves. But if the principal has a positive bias toward diversity, the misaligned manager improves reputation by hiring more from minority groups but sabotages them. This forms the diversity paradox, if there is no positive bias toward diversity, diversity does not improve much. But if there is, the diversity improves at the cost of increased sabotage. We show minorities in low productivity jobs are more likely to be sabotaged.

Chapter 1

Feedback on Ideas

1.1 Introduction

Employees are often assigned tasks with two distinct phases: in the first phase, ideas are generated; in the second phase, the best idea is implemented. Furthermore, it is common for supervisors to give feedback to their employees in this process. For instance, a partner in a law firm supervises an associate developing a litigation strategy, a project manager in a technology firm supervises an engineer solving a bug in app development, and a senior designer in an architecture firm supervises a junior designer looking for a design solution. One can trace such examples of feedback and supervision outside of corporate organizations as well; for instance, a professor supervising her grad student in a university.

This paper studies the supervisor's problem. Supervisors face the following trade-off. On the one hand, honest feedback encourages employees to discard bad ideas. On the other hand, such feedback can be demoralizing and discourage both idea generation and effort implementation. We build a model to describe how this trade-off shapes the supervisor's feedback, the employee's effort, and the employee's trust in the supervisor.

We consider a supervisor-agent model with two phases: experimentation and implementation. In the experimentation phase, the agent sequentially generates ideas at a cost, receives feedback from the supervisor on her ideas, and selects an idea to implement. In the implementation phase, the agent decides how much effort to put into completing her chosen idea. The agent's ability is initially unknown, and the agent and supervisor share a common prior. Importantly, we assume the supervisor does not internalize the agent's cost of effort. This misalignment of preferences means that

dishonesty is a possibility.[1]

Ability plays a central role in our model. We assume a high-ability agent both generates and implements ideas better than a low-ability agent. As a result, the agent's self-opinion (prior belief about her ability) affects both the agent's decision regarding how much to experiment and her choice of implementation's effort. Both of these effects, in turn, impact the supervisor's feedback.

There are three key findings of our model. First, the supervisor never gives a low self-opinion agent honest feedback because doing so is demotivating: it discourages effort in both the experimentation and implementation phases. When negative feedback discourages further experimentation, the supervisor prefers to falsely encourage the agent to induce her to put a higher effort in implementation instead. Therefore, negative feedback is only forthcoming for a high self-opinion agent. Moreover, a high self-opinion agent, independent of her actual ability, is repeatedly informed about her bad ideas and can end up being "treated more harshly".

Second, receiving supervisor feedback magnifies performance differences between high and low self-opinion agents. Because high self-opinion agents receive honest feedback, they have confidence both in their ability and in the quality of their ideas, which leads to high effort. Low self-opinion agents, in contrast, lack confidence, which leads to low effort. Receiving more honest feedback with a higher self-opinion allows the agent not only to experiment more but also to exert an optimal effort in implementing her chosen idea. Such an opportunity might not be available to a slightly lower self-opinion agent because she does not receive honest feedback as often. As a result, she has lower confidence in her idea. Therefore she might end up exerting too much effort on a bad idea, and too little effort on a good idea.

Third, overconfidence can be welfare improving. The discontinuous change in the supervisor's feedback strategy as the agent incorrectly goes from a low self-opinion to a high self-opinion gives rise to this possibility. The cost of overconfidence in ability is that it leads to too much effort exertion. However, the benefit of overconfidence is that it can lead to honest feedback. This benefit may outweigh the cost.

Our results find support in The Sensitivity to Criticism Test from PsychTests which collected responses from more than 3,600 participants.[2] The study revealed that

[1] Note that if providing feedback is costly to the supervisor (such as time costs) this could realign the principal's and agent's interests, thereby restoring honesty. We show that the supervisor is more (less) honest when he is more (less) time constrained, and therefore less (more) willing to supervise.

[2] https://eu.usatoday.com/story/money/columnist/kay/2013/02/15/at-work-criticism-sensitivity/1921903/

those who tended to be defensive about negative feedback had lower performance ratings and lower self-esteem. Moreover, managers were skeptical to give feedback to workers who get defensive. "If there was an esteem problem, both men and women seemed to block out the constructive part of the equation and only focus on the criticism", revealed a manager. This further meant that the manager would rather "develop the more (talented and) mature employee," instead of spending time counseling those who easily got defensive. These ideas further find support in the situational leadership theory developed by Paul Hersey and Ken Blanchard in mid-1970s. According to Ken Blanchard, " Feedback is the breakfast of champions."

Related Literature. Our paper relates to two distinct strands of literature: experimentation and dynamic communication games. Within experimentation, our work falls under models of motivating experimentation. Previous research has looked at how information can be optimally delivered to the agents arriving sequentially to experiment (such as Kremer, Mansour and Perry (2014) and Che and Horner (2015)) or at how information should be designed for a single agent to motivate her to experiment (such as Renault, Solan and Vieille (2017) and Ely (2017)).[3] Among the two, our setting falls in the latter category. Ely and Szydlowski (2017), Smolin (2017) and Ali (2017) are the closest in this respect.[4] In each of these papers, a principal must reveal information by balancing the positive effect of good news with the discouraging effect of no or bad news. Nonetheless, these papers do not address situations where ex-ante commitment to a disclosure rule is not possible. How the same tradeoff shapes the honesty in strategic feedback with no commitment is our point of departure from these papers. Thus, our model is one of communication rather than information design. To the best of our knowledge, we are the first to study such settings without commitment.

Another point of departure is how the agent responds to honest feedback. In our setting, the supervisor tries to motivate the agent to exert effort in both the experimentation and implementation phases. As a consequence, honest feedback can discourage the agent at two levels. The first is stopping experimentation too early, and the second is exerting low effort in implementation. Introducing this novel objective makes our setting unique in feedback and experimentation literature.

Some older papers like Lizzeri, Meyer and Persico (2002) and Fuchs (2007) have looked at feedback in dynamic settings without experimentation and show that often it

[3]See Hörner and Skrzypacz (2016) for a survey on the recent advancements in experimentation and information design.

[4]Some other related papers have looked at settings in which a sender commits to dynamic information design to influence a receiver. See, for example, Bizzotto, Rüdiger and Vigier (2018).

is not optimal to provide feedback.[5] Orlov (2013) considers a setting in which providing information to the agent might benefit the principal in the short-run but may lead to long term agency costs. There the principal designs an optimal information sharing rule along with a compensation scheme. Boleslavsky and Lewis (2016) also study dynamic settings with commitment in which the agent has new information every period. The principal makes sequential decisions, after which he observes a private signal of the state. These works consider the effect of feedback in settings with commitment but no experimentation. Our paper connects these two types of literature in a no-commitment setting.[6]

The other strand of literature related to our work is dynamic communication games. A few papers like Aumann and Hart (2003), Krishna and Morgan (2004), Forges and Koessler (2008) and Goltsman, Hörner, Pavlov and Squintani (2009) look at repeated communication with an action at the end. Our setup is different in that the receiver should decide after each round whether she wants to experiment again. Golosov, Skreta, Tsyvinski and Wilson (2014) and Renault, Solan and Vieille (2013) are closer in this sense. They look at situations where the receiver decides after every round of communication. However, neither has the above-stated feature of persuasion in two phases.

In this respect, our work relates to dynamic persuasion games. Morris (2001a), Honryo (2018) and Henry and Ottaviani (2019) are a few papers that do not assume commitment by the sender of information. The seminal paper by Morris (2001a) deals with a potentially biased advisor persuading a decision-maker to choose actions dynamically when reputation matters. Honryo (2018) and Henry and Ottaviani (2019), however, are closer to our setting. In these papers, a sender (entrepreneur or researcher) tries to persuade a receiver (venture capitalist or publisher) to take a favorable action by sequentially disclosing some verifiable or costly information. We instead have a tradeoff with cheap talk communication. In our model, when the supervisor persuades the agent to experiment again, he inadvertently also persuades her to exert lower effort in implementation. It is this feature that creates the main honesty/dishonesty tradeoff in our model.

Finally, our result on the importance of beliefs in final performance is related

[5] Both these papers are also concerned with the issue of dynamic moral hazard, and feedback plays an assistive role to contracting.

[6] Orlov, Skrzypacz and Zryumov (2018) is an exception. They look at commitment and no commitment case in a setting in which an agent tries to convince the principal to wait before exercising a real option. Again, however, their model does not have experimentation.

to some of the older research starting with Bénabou and Tirole (2002). This vast line of economics research is itself based on the original psychology research of Bandura (1977). However, such research usually looks at the importance of belief absent any external supervision. The presence of a supervisor drives our results on the effect of higher self-opinion and overconfidence.[7]

The rest of the paper is structured as follows. In Section 2.3, we describe the basic model. In Section 1.3, we solve two benchmark cases of the model without supervision, which help us build intuition and solve the complete game. Then, in Section 1.4, we present the main analysis of the game with a supervisor without commitment. We move onto presenting how our results are qualitatively the same in a few extensions and offer new interpretations of our model in Section 1.5. Finally, we conclude in Section 1.6.

1.2 The model

We consider a setting in which an *agent* (she) works on a project and a *supervisor* (he) is responsible for providing feedback. The project involves two distinct stages that proceed sequentially. The first stage is *planning* or *experimenting with ideas*, and the second stage is *execution* or *implementation of a chosen idea*. The agent is responsible for both experimenting with and implementing ideas for the completion of the project. The supervisor has no commitment power or verifiable signals and provides cheap talk feedback based on what he observes.

***Stage 1*: Idea generation.** The process of idea generation involves multiple rounds $t = 1, 2, \ldots$. In each round t, the agent decides whether she wants to draw a new idea. The quality of an idea is determined by its *ex-ante potential to succeed* θ_t which could be either high ($\hbar$) or low (ℓ). The distribution of θ_t is given by

$$\theta_t = \begin{cases} \hbar & \text{with probability } \alpha, \\ \ell & \text{otherwise} \end{cases}$$

where α is the *ability* of the agent. $\alpha \in \{0, q\}$ where zero is "low", and $q \in (0, 1)$ is "high". Therefore, only a high-ability agent can come up with a high potential idea, which happens with probability q. The ability (unlike the idea) remains persistent throughout the play. The agent and the supervisor only know the distribution of the ability; neither observes it. The belief that the agent is high-ability at the beginning of

[7] Koellinger, Minniti and Schade (2007) and Hirshleifer, Low and Teoh (2012) are two papers that empirically show the importance of overconfidence in the context of innovation and creativity.

round t is denoted by β_t, with a common prior $\beta_1 \in (0,1)$ at the beginning of the game in round 1. For much of the text, we use belief and self-opinion interchangeably. We assume that the agent possesses a low potential outside option idea at the beginning in round 1 denoted by $\bar{\theta} = \ell$.

Actions and timing: In each round of experimentation the agent chooses $I_t \in \{0,1\}$. $I_t = 0$ denotes the agent's decision to stay in Stage 1 and experiment with another idea in round t, i.e., not implement. There is a cost c of experimentation. It could arise from searching the Internet, looking up for data, reading material, previous works, and seeking inspiration. The agent produces an idea θ after privately incurring c.

Importantly, we assume that only the supervisor can see the potential of the idea generated. The supervisor privately observes θ_t and chooses an announcement about its observed potential, $m_t \in \{\ell, \hbar\}$.[8] We initially assume limited recall of the agent and the supervisor so that they only talk about the last idea produced (and not the entire history of past ideas). We present the analysis of perfect recall in which the supervisor is allowed to make backdated messages in Section 1.5.2.

Alternately, the agent could decide to implement the last idea after the supervisor's message. This is denoted by $I_t = 1$.

***Stage 2*: Idea implementation.** If the agent decides to move to the idea implementation stage in $t+1$ following the last message of the supervisor m_t, then her idea gets fixed at $\theta \equiv \theta_t$.

Actions and timing: The agent chooses effort $e \in [0,1]$ at cost $\frac{e^2}{2}$ to complete the project. The final outcome of the project, success or failure, is determined by the following distribution function

$$\Pr(\text{success}) = \begin{cases} e & \text{if } \theta = \hbar, \\ ke & \text{if } \theta = \ell \text{ and agent is high-ability, } k \in (0,1), \\ 0 & \text{otherwise.} \end{cases}$$

The probability of success is a function of the potential of the chosen idea θ, effort exerted by the agent e and the ability of the agent α. It must be noted that only

[8] We can also start with an arbitrary message space M but since we consider a game of cheap talk with binary types and we focus on pure strategy equilibria, what matters are the equilibrium mappings from the supervisor type (what potential of the ideas he observes) to the message space, i.e. what is the meaning of the messages. Here, messages ℓ and $\hbar$ have their natural meaning and are understood as the potential of the idea developed.

the high-ability agent is capable of successful completion of the project. Moreover, only she may obtain a success even with a low potential idea. Therefore, when the ability is unknown there is an incentive to implement a low potential idea instead of experimenting again.

We will make the following assumption for mathematical convenience.

$$q \geq (q + (1-q)k)^2 \geq k \tag{A}$$

Intuitively, this assumption implies that in case the agent has a low potential idea, the supervisor finds it beneficial for the agent to experiment than to implement that idea (with the maximum possible effort of 1). Further, an additional round of experimentation with feedback is preferred to an additional round of experimentation without feedback. We explain these ideas further when presenting the main analysis in Section 1.4.[9]

Payoffs: Completion of the project yields V. If the completed project is successful, it yields a normalized value of 1, and zero otherwise. The payoff of the agent is given by

$$u_A = V - Tc - \frac{e^2}{2}$$

where T is the number of rounds for which the agent has experimented. The payoff of the supervisor is given by

$$u_S = V.$$

The payoffs highlight the incentive misalignment between the agent and the supervisor. While both players prefer success over failure, the agent alone bears the cost of experimentation and implementation.

Once the payoffs are realized, the game ends. A summary of the timing of the game is provided in Figure 1.1. We provide an alternate interpretation of the model and additional examples in Section 1.5.3.

[9]This assumption helps simplify the proofs by providing sufficient conditions. In the absence of this assumption, all our proofs go through but will be belief dependent, which makes them less obvious and more cumbersome.

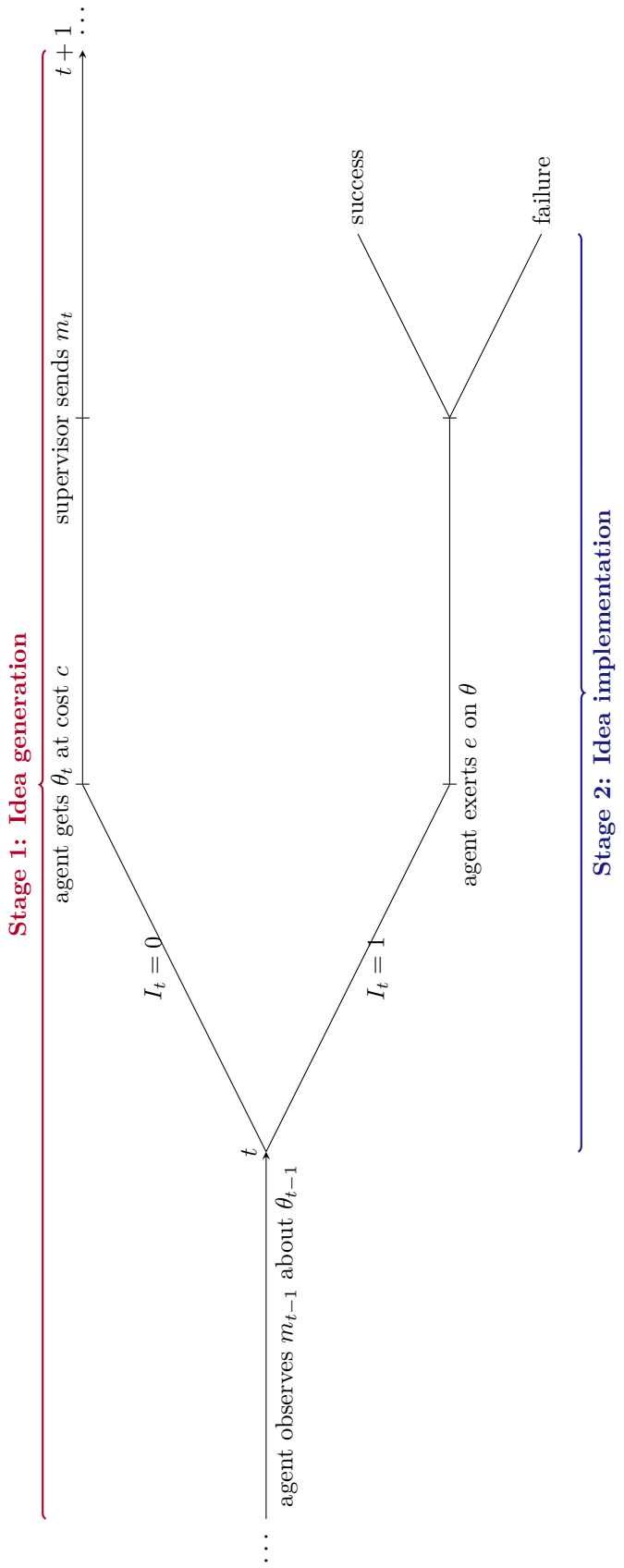


Figure 1.1: Summary of the timing of the game

We now turn to the analysis of the game. Before we describe the behaviour of a strategic supervisor, we describe the benchmark case in the following section without

the supervisor. We then introduce the supervisor in Section 1.4 and search for honest equilibrium feedback strategies.

1.3 Benchmark: Single agent problem

In this section we look at a setting in which an agent works on the project without any supervision. This preliminary analysis helps us put bounds on the behavior of the agent and supervisor when they interact with each other. Two cases are possible – the agent does not observe the potential of her idea, or she does so perfectly.

1.3.1 No information (NI) about θ

If the agent does not observe the potential of her idea θ from attempting experimentation at belief β and there is no outside support, then the two alternatives available to her are as follows:

1. The agent can choose to not experiment and directly implement the project using the outside option idea. In this case, the agent $\max_e \beta k e - \frac{e^2}{2}$, which yields a maximized payoff of $\frac{(\beta k)^2}{2}$.

2. The agent can choose to experiment once and then execute the resulting idea. In this case, the agent $\max_e \beta(q+(1-q)k)e - \frac{e^2}{2} - c$, which gives a maximized payoff of $\frac{\beta^2(q+(1-q)k)^2}{2} - c$.

Observe that the agent does not want to try experimenting more than once in this setting because experimenting is an additional cost without any added benefit. She will not learn the quality of the new idea and the odds of coming up with a high potential idea remain unchanged. The only reason she might want to experiment once is to take the gamble of coming up with a high potential idea. She will do so if her belief is high enough. This is illustrated in the following condition:

$$\overbrace{\frac{\beta^2(q+(1-q)k)^2}{2}}^{\text{expected benefit of experimentation}} \geq \overbrace{\underbrace{\frac{(\beta k)^2}{2}}_{\text{opportunity cost}} + \underbrace{c}_{\text{actual cost}}}^{\text{cost of experimentation}}, \tag{C1}$$

which leads to the following lemma:[10]

[10] A similar lemma with a belief threshold condition can also be obtained if the agent has no outside option idea. Denote such a cutoff by β_ϕ^{NI}. Then it can be shown that such a cutoff exists and is given by $\beta_\phi^{NI} = \frac{(2c)^{1/2}}{q+(1-q)k}$. Obviously, $\beta_\phi^{NI} < \beta_0^{NI}$. However, we make use of β_0^{NI} in the main analysis – we assume away the possibility of quitting when there is no support from a supervisor.

Lemma 1.1. *Let* $c < \frac{(q+(1-q)k)^2-k^2}{2}$. *If there is no information about* θ, *there exists a unique threshold* $\beta_0^{NI} := \left(\frac{2c}{(q+(1-q)k)^2-k^2}\right)^{\frac{1}{2}}$ *such that*

1. *if the prior belief* $\beta_1 \geq \beta_0^{NI}$ *then the agent experiments once before finishing the project by exerting effort* $\beta_1(q+(1-q)k)$, *and*
2. *if the prior belief* $\beta_1 < \beta_0^{NI}$, *the agent uses the outside option idea* $\bar{\theta} = \ell$ *to finish the project by exerting effort* $\beta_1 k$.

In the text we will also be interested in how β_0^{NI} responds to changes in the cost of experimentation c. It is easy to see that a higher cost of experimentation raises this threshold as it reduces the incentives to experiment *ceteris paribus* (see Appendix B for other comparative statics result).

1.3.2 Full information (FI) about θ

When the agent can perfectly observe the outcome of each round of experimentation, then she potentially wants to experiment at least once. This, as before, depends on her belief about her ability. But now she uses Bayes' rule sequentially to update her belief after observing the potential of the idea from the latest round of experimentation in a way that

$$\beta_t = \begin{cases} \frac{(1-q)\beta_{t-1}}{1-\beta_{t-1}q} & \text{if } \theta_{t-1} = \ell, \\ 1 & \text{otherwise.} \end{cases}$$

As is standard in good-news models, the agent revises her belief downwards each time she generates a low potential idea, but her belief jumps to 1 if she generates a high potential one. The agent enters the implementation phase and finishes the project upon observing $\theta_{t-1} = \hbar$. At this point, she does not have an incentive to experiment further as she only bears an additional cost without any extra benefit. She finalizes the project with the maximum effort of 1 which leads to the project being successful with certainty, and yields a maximized payoff of $\frac{1}{2}$ (the previous cost of experimentation is sunk). Thus, independent of which round of experimentation she is at if $\theta_{t-1} = \hbar$ then $I_t^{FI}(\beta_t = 1) = 1$ is optimal with $e^{FI}(\beta_t = 1) = 1$.

On the other hand, after observing $\theta_{t-1} = \ell$ (with the agent observing low potential ideas $\theta_{t'} = \ell$ for all the previous rounds $t' < t-1$ as well) the agent holds a belief $\beta_t < 1$ about her ability. The agent again faces two choices – to implement the low potential idea or to continue experimenting. If she chooses to implement her low potential idea then she chooses the optimal effort to $\max_e \beta_t k e - \frac{e^2}{2}$. This yields a maximized payoff of $\frac{(\beta_t k)^2}{2}$ where she exerts effort $\beta_t k$ according to her belief β_t.

Depending on her belief β_t she might be a high-ability agent with a positive probability of success. If she chooses to experiment once more, then with probability $\beta_t q$ she comes up with a high potential idea and exerts maximal effort of 1 thereafter to finish the project (from above). With probability $1-\beta_t q$ she comes up with a low potential idea and she faces the same decision problem but with a lower belief $\beta_{t+1} < \beta_t < 1$. Denote the value function of the agent at the beginning of round t with belief β_t when her last observed outcome is $\theta_{t-1} = \ell$ by $\mathcal{V}^\ell(\beta_t)$, such that

$$\mathcal{V}^\ell(\beta_t) = \max\left\{\frac{(\beta_t k)^2}{2}\ ,\ -c + \frac{\beta_t q}{2} + (1-\beta_t q)\mathcal{V}^\ell(\beta_{t+1})\right\}.$$

Assuming that the agent wants to start experimenting (the condition for which we will outline below), we are interested in if and when the agent stops experimenting with repeated low potential ideas. To do so, let the maximum number of rounds the agent experiments be T. The agent at belief $\beta_T \equiv \beta$ after $T-1$ rounds will attempt another *final* round of experimentation knowing that irrespective of the outcome she will move to implementing her idea in the following round. So

$$\begin{aligned}\mathcal{V}^\ell(\beta) &= \max\left\{\frac{(\beta k)^2}{2}\ ,\ -c + \frac{\beta q}{2} + (1-\beta q)\mathcal{V}^\ell(\beta')\right\} \\ &= -c + \frac{\beta q}{2} + (1-\beta q)\mathcal{V}^\ell(\beta') \quad \geq \quad \frac{(\beta k)^2}{2}\end{aligned}$$

where

$$\beta' = \frac{(1-q)\beta}{1-\beta q} \text{ and } \mathcal{V}^\ell(\beta') = \frac{(\beta' k)^2}{2}\ ,$$

which can be rearranged to

$$\overbrace{\frac{\beta q}{2} + (1-\beta q)\frac{(\beta' k)^2}{2}}^{\text{expected benefit of experimentation}} \quad \geq \quad \overbrace{\underbrace{\frac{(\beta k)^2}{2}}_{\text{opportunity cost}} + \underbrace{c}_{\text{actual cost}}}^{\text{cost of experimentation}}. \tag{C2}$$

Lemma 1.2 follows from condition (C2) and captures the optimal behaviour of the agent under full information about θ. (All proofs are presented in Appendix A.)

Lemma 1.2. *If there is full information about θ, the optimal decision rule of the agent*

I_t^{FI} is a unique belief threshold rule such that

$$I_t^{FI} = \begin{cases} 0 & \text{if } \theta_{t-1} = \ell \text{ and } \beta_t \geq \beta_0^{FI} \ , \\ 1 & \text{otherwise.} \end{cases}$$

for $c < \frac{q(1-k^2)}{2}$. Further, the optimal effort that the agent exerts to implement her idea is given by

$$e^{FI} = \begin{cases} \beta_{T+1}k & \text{if } \theta_T = \ell \ , \\ 1 & \text{otherwise.} \end{cases}$$

When $c \geq \frac{q(1-k^2)}{2}$ the agent does not experiment for any belief, and implements her outside option idea with effort $\beta_1 k$.

Figure 1.2 plots the expected benefit from experimentation (LHS plotted in green) and the cost of experimentation (RHS plotted in red) from condition (C2) for different levels of beliefs β. It illustrates the uniqueness result of Lemma 1.2 under the cost condition $c < \frac{q(1-k^2)}{2}$. Note that both the benefit and the costs are declining in belief about ability. A lower belief in ability means that the agent is less likely to get a high potential idea, which reduces the expected benefit of experimentation. At the same time, for the same reason, it induces the agent to exert lower effort when implementing the outside option idea, thereby reducing the opportunity cost of experimentation. However, the fixed component c of the total costs of experimentation ensures that the costs never go down to zero, which in turn guarantees the existence of the unique threshold.

Observe that the optimal decision rule does not depend on t but only on the belief β, which is a function of the potential of the last observed idea. For a given set of parameters, the maximum number of rounds the agent experiments T is only defined by the prior belief β_1. The agent wants to start experimenting with ideas if $\beta_1 \geq \beta_0^{FI}$, and goes on doing so with repeated low potential ideas as long as the belief hits β_0^{FI}. T is therefore determined by how far β_1 is from β_0^{FI}.

It only remains to show how β_0^{FI} varies with a change in parameters. Again, we'll be interested in how β_0^{FI} responds to a change in the cost of experimentation. As expected, an increase in the cost of experimentation raises the threshold belief β_0^{FI} as the agent wants to experiment fewer rounds now (for any prior).

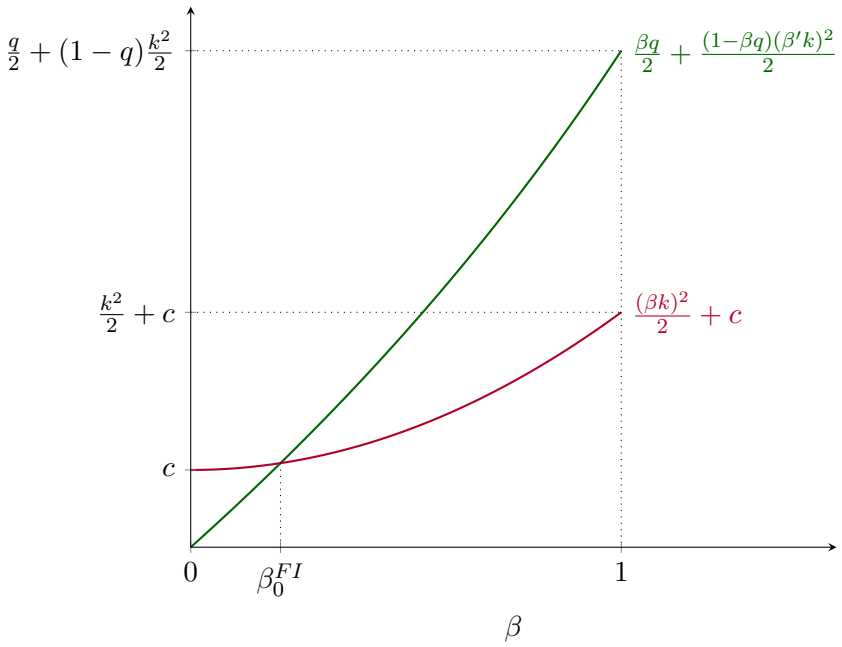


Figure 1.2: The optimal belief threshold β_0^{FI} for the complete information about θ case

1.3.3 Comparing β_0^{NI} and β_0^{FI}

Lemma 1.3. *If* $c < \frac{(q+(1-q)k)^2-k^2}{2}$*, then both* β_0^{NI} *and* β_0^{FI} *exist and are unique with* $\beta_0^{NI} > \beta_0^{FI}$.

Figure 1.3 illustrates why $\beta_0^{NI} > \beta_0^{FI}$. It shows that for any belief β the value of experimenting is always lower in the case when the agent has no information about her output of experimentation. Experimentation is merely a gamble to try luck without any learning. This makes the threshold for experimentation higher under the no information case.

1.3.4 An important definition

Before moving to the main analysis, we introduce some additional terminology that we will use extensively in the following sections.

Given the no information and the full information belief thresholds β_0^x for $x \in \{NI, FI\}$, define recursively a sequence of belief thresholds $\{\beta_i^x\}_{i=0}^{\infty}$ such that $0 < \beta_i^x < 1$ and $\beta_{i+1}^x = \frac{\beta_i^x}{1-q(1-\beta_i^x)}$. Starting with the threshold β_0^x the sequence identifies β_1^x, the belief that leads to β_0^x when the agent correctly finds out that her idea has a low potential to succeed, and so on. Therefore, β_{i+1}^x is the belief which when updated with the correct information about a low potential outcome leads to the belief β_i^x, and this is recursively defined all the way down to the belief β_0^x.

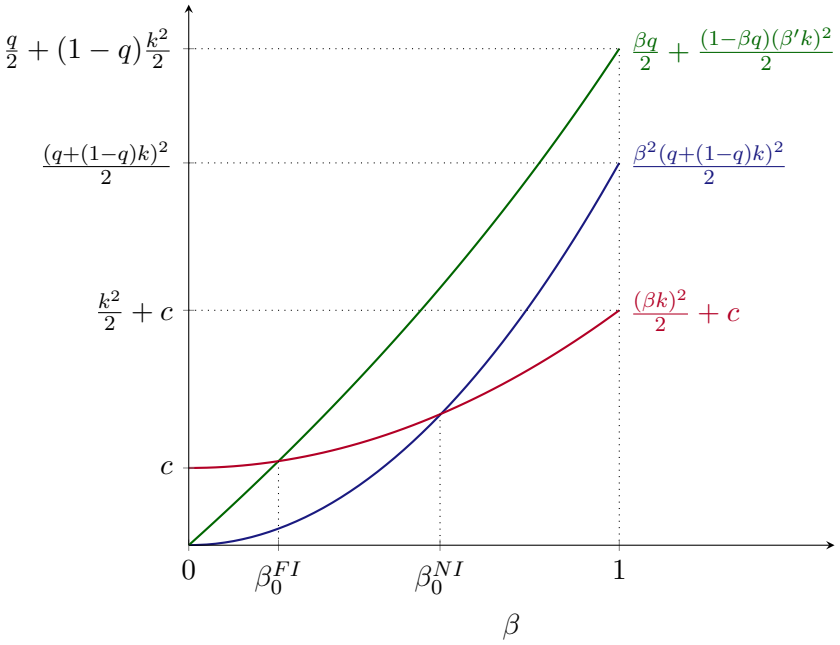


Figure 1.3: Comparing β_0^{NI} and β_0^{FI}

1.4 Strategic supervisor

1.4.1 Preliminaries

The game between a strategic supervisor and an agent in Stage 1 is one of dynamic cheap talk. The supervisor can costlessly send either of the two messages independent of the true potential of the idea. Our solution concept is (perfect) Bayesian Equilibrium.

To define the strategies of the agent and the supervisor at any time, we would need to define the history for each player when they are called upon to make a decision. Round t begins for the agent after having observed the last message sent by the supervisor m_{t-1}. Accordingly, a realized history for the agent includes the set of all previous messages sent by the supervisor until and including the last message m_{t-1} and the sequence of past decisions made. Round t begins for the supervisor after observing the last idea of the agent θ_t. Accordingly, a realized history for the supervisor includes, in addition to the history viewed by the agent, the sequence of all the realized idea potential from the past experimentation.[11]

[11] Let $I^t := (I_1, \ldots, I_t)$ and $m^t := (m_1, \ldots, m_t)$ be the sequence of decisions made by the agent and the public messages given by the supervisor until round t. Define the set of histories for the agent and the supervisor at the beginning of round t by H_t^A and H_t^S respectively. The history for the agent at the beginning of round t is

$$h_t^A = (I^{t-1}, m^{t-1}) \in H_t^A \subset (\{0\}^{t-1} \times \{\ell, \hbar\}^{t-1}).$$

For most of the paper, we focus on pure strategy equilibria and limited recall, i.e. we are interested in whether the supervisor is honest with the agent when he can only send a message about the last idea generated. A pure strategy for the supervisor in round t is a mapping from the realized history to the message space $\{\ell, \hbar\}$. The supervisor is honest with the agent if for any realization of the history the supervisor sends a message that matches the observed potential of the idea. If the supervisor reveals to the agent the outcome of her last experimentation in round t starting from a prior β_t the agent's updated posterior in round $t+1$ is as in the full information case:

$$\beta^{\ell}_{t+1} = \frac{(1-q)\beta_t}{1-q\beta_t} \text{ if } m_t = \ell \text{ , and} \tag{1.1}$$

$$\beta^{\hbar}_{t+\tau} = 1 \text{ otherwise.} \tag{1.2}$$

If the supervisor uses the same message independent of the realized history the supervisor is said to lie or babble (see footnote 12). In this case the agent's posterior belief is the same as her prior belief. We will assume that when the supervisor is expected to lie the agent does not consult the supervisor. This rules out the possibility of the supervisor privately learning and not revealing to the agent the outcome, and the arising deviations.

Given our focus on pure strategies and that the two players share a common prior , the agent and the supervisor symmetrically update their belief on the agent's ability. If the agent stops experimenting (and implements her last idea) because the supervisor is babbling, neither the agent nor the supervisor have any new information. There is learning only insofar as the supervisor is honest.

1.4.2 Analysis

What feedback strategy the supervisor employs will depend on how he expects the agent will respond to it, both in the experimentation phase and the implementation phase. We begin by discussing the obvious babbling equilibria. Babbling is always an equilibrium for any prior β_1 in the first stage of the game. The agent does not learn about the true potential of the last idea as the supervisor is always expected to send the same message. This is equivalent to the single agent decision-making problem without

This is also the public history of the play of the game up to round t. In addition to the public history, the supervisor observes $\theta^t := (\theta_t, \ldots, \theta_t)$ and an extra decision of the agent to experiment $I_t = 0$. The history for the supervisor at the beginning of round t is

$$h^S_t = (\theta^t, I_t, h^A_t) \in H^S_n \subset (\{\ell, \hbar\}^t \times \{0\}^t \times \{\ell, \hbar\}^{t-1}).$$

advice and Lemma 1.1 applies. Thus, the agent experiments once before finishing the project if $\beta_1 \geq \beta_0^{NI}$, otherwise she uses the outside option idea to finish the project. Neither supervisor type can profitably deviate from such an equilibrium given the beliefs. The supervisor sends meaningless messages, the agent correctly believes that there is no information content in the recommendations and she makes her decision only on the basis of her prior belief.[12]

In what follows we determine if there exist any pure strategy equilibria in which the supervisor is honest, and under what conditions. The approach will be to determine the existence for different ranges of beliefs starting with low ones.[13]

Proposition 1.1. *For any belief $\beta < \beta_0^{FI}$, any communication strategy is an equilibrium and none induces the agent to experiment.*

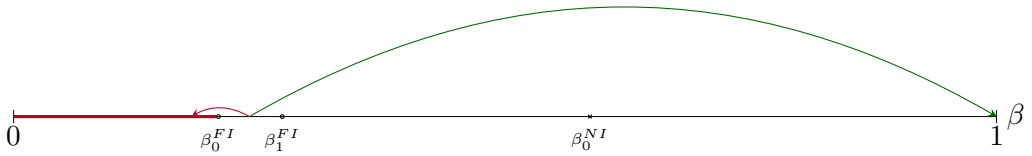
From Lemma 1.3, we know that $\beta_0^{FI} < \beta_0^{NI}$. The region of beliefs $\beta < \beta_0^{FI} < \beta_0^{NI}$ is the one in which the agent does not want to experiment with ideas independent of how much information is provided to her. So all communication strategies are equally informative to the agent and are an equilibrium. The agent does not consult the supervisor in any equilibria as she is very pessimistic about her ability to come up with a high potential idea. She does not want to bear the cost of experimentation at such low beliefs. She simply implements her low potential outside option idea $\bar{\theta} = \ell$ with an effort βk.

A concern when evaluating whether the supervisor can be honest for higher beliefs will be what he thinks is the possibility of the agent experimenting again after a negative message. As in the the full information case outlined in Section 1.3.2, the agent experiences a decline in both the benefit and cost of coming up with a new idea after receiving a truthful negative messages. With continued discouragement the agent must stop experimenting at belief β_0^{FI}. However, the supervisor's payoff is contingent

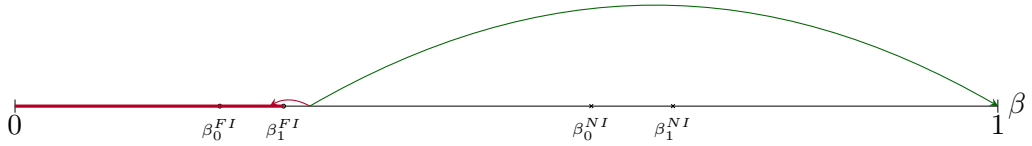
[12] When the supervisor babbles, it might be useful to think of babbling in mixed strategies rather than in pure strategies (see description of mixed strategies in Appendix A). A supervisor babbling in mixed strategies makes use of both the messages in equilibrium, and the posterior β_t after either message remains unchanged. There are also babbling equilibria in pure strategies. Say the agent conjectures that the supervisor only says $m = \hbar$ on-the-equilibrium path. We have that $\Pr(m = \hbar | \theta = \hbar) = 1 - \Pr(m = \ell | \theta = \ell)$ and a potential babbling equilibrium. While there is no update of beliefs on path, the message $m = \ell$ is off path and we would need to specify beliefs in the information set following this message. Such an equilibrium is supported by any belief $\beta^{\text{offpath}} \in [0, \beta_1)$.

[13] The proofs will be presented in terms of a generic belief β wherever possible. The intuition is the same – whether the agent starts out in the given range with a low potential outside option idea or whether she lands there after continued experimentation (and ending up with a low potential idea that she is aware of), if she finds herself there her behavior is the same. If she finds herself in any of the ranges with the knowledge that her idea was definitely a high potential idea, then she will always immediately implement her idea by exerting effort 1.

Step 1: Babbling is unique for $\beta_0^{FI} \leq \beta_1 < \beta_1^{FI}$



Step 2: Babbling is unique for $\beta_1^{FI} \leq \beta_1 < \beta_0^{NI}$



Step 3: Babbling is unique for $\beta_0^{NI} \leq \beta_1 < \beta_1^{NI}$

Figure 1.4: Uniqueness of babbling equilibria for priors $\beta_1 < \beta_1^{NI}$

on the agent's success. This implies that the he faces a discontinuous drop in the benefit of being honest at β_0^{FI}, whiile the cost is that the agent exerts a lower effort in implementation. Our first main result and proposition builds on this intuition. It defines the range of beliefs for which the cost of being honest are higher than the benefits.

Proposition 1.2. *For any belief $\beta_0^{FI} \leq \beta < \beta_1^{NI}$, babbling is the unique equilibrium strategy.*

The intuition for this proposition is illustrated in steps using Figure 1.4.[14]

We begin by showing that babbling must be a unique equilibrium strategy of the supervisor in the range of priors $\beta_0^{FI} \leq \beta_1 < \beta_1^{FI}$ (see Step 1 of Figure 1.4). In this range of priors, a message about the idea being low potential if expected in equilibrium must lead to a posterior about ability $\beta_2^{\ell} < \beta_0^{FI}$. At this point the agent does not want to experiment any more (from Proposition 1.1). Moreover, after experimenting and learning that her idea had a low potential to succeed she reduces her effort when implementing the idea. As a result, the expected probability of success further reduces with the low potential idea. This leads the supervisor observing a low potential idea to deviate from honesty and always send a positive message instead.

A positive message is believed by the agent pushing up the posterior of the agent

[14]Here we discuss the intuition of why honesty cannot be an equilibrium strategy but the proposition is stronger. The argument will also hold to prove that no informative equilibria will survive in this range of beliefs. Our proof in Appendix A presents a general proof that allows for mixed strategies as well.

to 1. The agent best responds by implementing the chosen idea with the maximal effort of 1, which increases the expected probability of success with a low potential idea. The supervisor is at the very least able to extract a higher effort on a low potential idea by deviating. Thus, no equilibria in which the supervisor is honest will survive – babbling is unique in this range of priors. In such a babbling equilibrium, the agent best responds by not experimenting because this is identical to a situation with no supervisor and $\beta_1 < \beta_0^{NI}$ (from Lemma 1.1).

Now, in Step 2 consider the range of priors which when updated with negative messages lead to posteriors below β_1^{FI}. The same argument as the one highlighted above holds because such low posteriors lead the agent to implementing the low potential idea with a lower effort. This time because the supervisor is expected to babble if updated with an honest discouraging message. Therefore, an agent expecting information can be taken advantage of by supervisor type who has only observed low potential ideas. This kills honesty and only the babbling equilibria survive. The same logic can now be extended all the way up to all the prior beliefs which when updated with a discouraging message about the idea lead to posteriors below β_0^{NI}. Below β_0^{NI} the agent does not want to experiment when no information is provided by the supervisor. Such is the case for all prior beliefs $\beta_1 < \beta_1^{NI}$ (illustrated in Step 3).

The total communication breakdown between the supervisor and the agent in this range of beliefs is driven by the fear of the supervisor to discourage the agent to the point of no further experimentation. This is why we call this region of beliefs as those in which the agent has a *low self-opinion*. When he sees that the agent has produced a low potential idea the supervisor finds it beneficial to cajole the agent by calling it a high one, so that at the very least the agent exerts a high effort to implement a low potential idea. But lying is counter-productive as the agent expects the supervisor to only provide fake encouragement; neither does she consult the supervisor nor does she experiment.

This region of beliefs $\beta_0^{FI} \leq \beta < \beta_1^{NI}$ where the agent has a low self-opinion reflect pure inefficiencies in the supervisor-agent relationship. From Lemma 1.2 we know that the agent would continue experimenting with ideas until she produces a high potential idea for beliefs $\beta \geq \beta_0^{FI}$ if she receives honest feedback. At the same time, the supervisor is also (always) better off with repeated experimentation until a high potential idea is produced. But neither can achieve this better outcome because the supervisor is unable to commit to honestly revealing the result of the agent's

experimentation. Even though the agent is willing to listen to honest feedback, her reaction to negative feedback is too extreme from the supervisor's point of view. If the agent must give up, he prefers she exert the maximum effort instead. Such inefficiency will be a feature of any communication equilibrium we can construct as babbling is unique. The supervisor cannot offer any information in equilibrium.

The extent of babbling and that of the resulting inefficiency is determined by the gap between β_0^{FI} and β_1^{NI}, which is a function of the parameters. An increase in the cost of experimentation (c) increases both these thresholds and causes babbling for even higher beliefs (and also no experimentation for higher beliefs). An increase in the probability of generating a high potential idea (q) reduces the region of babbling. An increase in the success rate from implementing a bad idea (k) can *decrease* the inefficiency by reducing the babbling region as it makes the agent want to experiment more without supervision by reducing β_0^{NI}.

Note, however, the difference in the agent's best response to such an uninformative strategy of the supervisor. Since the supervisor babbles in the entire region of beliefs below β_1^{NI}, from Lemma 1.1 the agent best responds by not experimenting in the region below β_0^{NI} and by experimenting once in the region between β_0^{NI} and β_1^{NI}. This produces an added source of inefficiency when she experiments in this region i.e. when the belief is above β_0^{NI} but below β_1^{NI}. In this case, the agent exerts an inefficient level of effort to implement the idea as she is unable to observe the potential of her idea without honest supervision. She exerts more effort on a low potential idea and a lower effort on a high potential idea.

We are now in a position to determine if there are any honest equilibria. The possibility of honesty opens up for beliefs $\beta > \beta_1^{NI}$ because the agent is now willing to experiment at least once without the supervisor's support. This happens in the region of beliefs between β_0^{NI} and β_1^{NI}. The previous threat point for the supervisor now potentially disappears as the supervisor can guarantee that the agent will experiment even when she is discouraged. In this sense, we call this the region of *high self-opinion*. We are now in a position to analyse whether this one extra round of experimentation (without the consultation of the supervisor) and a high self-opinion is sufficient for the supervisor to be honest.

Proposition 1.3. *For* $c \geq \frac{\kappa k - (\kappa k)^2}{2}$ *where* $\kappa \equiv \frac{k}{(q+(1-q)k)^2}$ *and for all* $t \geq 1$,

1. *truth-telling is an equilibrium strategy for the supervisor for* $\beta_t \geq \beta_1^{NI}$, *and*

2. babbling is the unique equilibrium strategy for the supervisor for $\beta_t < \beta_1^{NI}$.

The agent's equilibrium strategy is given by

$$I_t^* = \begin{cases} 0 & \text{if } m_{t-1} = \ell \text{ and } \beta_t \geq \beta_1^{NI} \text{ , or } \beta_0^{NI} \leq \beta_t < \beta_1^{NI} \text{ ,} \\ 1 & \text{otherwise.} \end{cases}$$

The agent's optimal effort is given by

$$e^* = \begin{cases} 1 & \text{if } m_{t-1} = \hbar, \\ \beta_t(q + (1-q)k) & \text{otherwise.} \end{cases}$$

Proposition 1.3 identifies the necessary and sufficient condition for an honest equilibrium to arise in the *entire* region above babbling equilibria, i.e. one of high self-opinion. This is shown to be when the agent's cost of experimentation is sufficiently *high*. To see this, let us first look at the supervisor's incentives to be honest in the region of priors $\beta_1^{NI} \leq \beta_1 < \beta_2^{NI}$. Here the agent experiments once even when discouraged. At most the agent's belief can fall down to β_0^{NI} after a negative message. The supervisor is then willing to discourage the agent with a negative message only if he can ensure that even after discouragement the agent does not reduce her effort significantly. In the absence of further supervision, he can only expect a higher expected probability of success if she exerts a high enough effort in implementation.

A supervisor who has observed a low potential idea expects the project to be successful with probability $(\beta_2^\ell(q + (1-q)k))^2$ from being honest. After receiving a message $m_1 = \ell$, the agent correctly believes her current idea has a low potential to succeed and experiments once again but does not seek supervision because the supervisor is expected to babble. In this case, the agent then implements the next idea with effort $e = \beta_2^\ell(q + (1-q)k)$. On the other hand, if such a supervisor deviates from honesty and announces $m_1 = \hbar$, then he expects the probability of success to be $\beta_2^\ell k$. The agent incorrectly believes that her idea had a high potential to succeed and exerts effort of 1 in implementing a low potential idea. For such a conjectured strategy to be an equilibrium, we must have that

$$\begin{aligned} &(\beta_2^\ell)^2(q + (1-q)k)^2 \geq \beta_2^\ell k \\ \implies &\beta_1 \geq \frac{k}{qk + (1-q)(q + (1-q)k)^2} := \beta^{\text{truth}} \end{aligned}$$

Thus, the supervisor requires agent's belief to be sufficiently high even after discouragement, which in turn requires the prior to be large enough. This ensures that the agent exerts a higher effort in implementing her idea of unknown potential. We call this truth-telling threshold on prior β^{truth}.

The truth-telling threshold β^{truth} is a conditional threshold. It identifies how high the prior should be such that the supervisor has an incentive to reveal the truth about the agent's negative outcome *if the agent experiments again without supervision following the negative message*. The supervisor does not directly care about the agent's cost of experimentation in so far as she attempts to experiment again with an idea. So β^{truth} does not depend on c.

Now all we need to do is identify whether the range of priors we are considering delivers honesty by the supervisor, that is we are interested in if $\beta^{\text{truth}} < \beta_2^{NI}$. Specifically, if $\beta^{\text{truth}} \leq \beta_1^{NI}$ then truth-telling is an equilibrium for the full range of beliefs above β_1^{NI} and up to β_2^{NI}. If this condition is satisfied, the supervisor has an incentive to be honest because the prior is sufficiently high given the parameters. As outlined above, β^{truth} does not depend on the cost of experimentation c while β_1^{NI} does. The one free parameter can be used to determine if truth-telling is an equilibrium. The condition $\beta^{\text{truth}} \leq \beta_1^{NI}$ can then be rearranged to

$$c \geq \frac{\kappa k - (\kappa k)^2}{2} \text{ where } \kappa \equiv \frac{k}{(q + (1-q)k)^2} < 1.$$

Intuitively, a lower bound on the cost of experimentation ensures that the agent's no information thresholds β_0^{NI} and β_1^{NI} are high enough. Thus, when the agent decides to experiment and consult the supervisor her belief in her ability is already high. The supervisor can then be content with revealing the truth about low potential ideas to the agent. Discouragement does not lead to quitting with low effort; the agent still experiments once more and does so by exerting a sufficiently high effort. While the conditional truth-telling threshold β^{truth} is not a function of the cost of experimentation c, whether truth-telling is an equilibrium depends on it. An increase in the cost of experimentation raises the threshold β_0^{NI} (increasing the region of babbling) but has no effect on β^{truth}, making it easier to satisfy the condition $\beta^{\text{truth}} \leq \beta_1^{NI}$ and ensuring truth-telling above β_1^{NI}.

We are now only left with determining why if the supervisor is honest in the range of beliefs $\beta_1^{NI} \leq \beta_1 < \beta_2^{NI}$, then he should be honest in the range of beliefs above β_2^{NI}. For expositional convenience start now with the range of beliefs $\beta_2^{NI} \leq \beta_1 < \beta_3^{NI}$

when it is an equilibrium for the supervisor to be honest in the next lower range of beliefs. Consider whether a conjectured strategy of honesty is an equilibrium for the supervisor. A supervisor who observes a low potential idea can induce another two rounds of experimentation by being honest at this stage, one with supervision and one without. If, however, he deviates he induces the agent to exert maximal effort in a low potential task. Under assumption (A), the payoff from being honest are strictly higher than that from deviating as it is evaluated relative to his private updated belief β_2^ℓ. The same line of reasoning can then be extended to any belief above β_3^{NI} as well so that the supervisor always prefers honestly discouraging the agent and getting her to experiment more often than making her implement a low potential idea.

What happens when $c < \frac{\kappa k - (\kappa k)^2}{2}$? The following corollary identifies the honest equilibrium.

Corollary 1.1. *When $c < \frac{\kappa k - (\kappa k)^2}{2}$, $\beta_j^{NI} \leq \beta^{truth} < \beta_{j+1}^{NI}$ exists such that for all $t > 1$ for $j \geq 1$*

1. *truth-telling is an equilibrium strategy for the supervisor for $\beta_t \geq \beta^{truth}$, and*
2. *babbling is an equilibrium strategy for the supervisor for $\beta_t < \beta^{truth}$.*

The agent's equilibrium strategy is given by

$$I_t^* = \begin{cases} 0 & \text{if } m_{t-1} = \ell \text{ and } \beta_t \geq \beta^{truth} \text{ , or } \beta_{j-1}^{NI} \leq \beta_t < \beta_j^{NI} \text{ ,} \\ 1 & \text{otherwise.} \end{cases}$$

The agent's optimal effort is given by

$$e^* = \begin{cases} 1 & \text{if } m_{t-1} = \hbar, \\ \beta_t(q + (1-q)k) & \text{otherwise.} \end{cases}$$

In this case, $\beta^{\text{truth}} > \beta_1^{NI}$ and can lie between any β_j^{NI} and β_{j+1}^{NI}. We can then again construct an honest equilibrium above β^{truth} and a babbling one below. That all of these beliefs are above β_0^{NI} ensures that the agent experiments once more when a low potential idea is revealed to her in the presence of future babbling and makes such a strategy an equilibrium.The two cases discussed here are depicted in Figure 1.5.

It is worth emphasizing at this stage the key intuition driving the results in Propositions 1.2 and 1.3. What action the agent chooses depends on whether she thinks she is capable of drawing a better idea, and the expected strategy of the supervisor. If

1. Honest equilibria when $c \geq \frac{\kappa k-(\kappa k)^2}{2}$

2. Honest equilibria when $c < \frac{\kappa k-(\kappa k)^2}{2}$

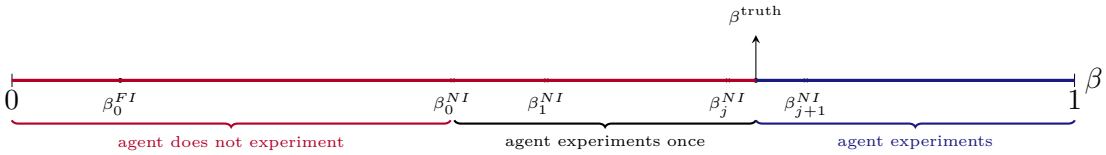


Figure 1.5: Honest equilibria for different c ranges

the agent has produced a low potential idea , the supervisor needs to incorporate the downwards effect that his negative message has on the belief about her ability. A lower belief discourages the agent at two levels. First is the discouragement to experiment, i.e., stopping experimentation too early. Second is the discouragement to implement, i.e., exerting low effort in implementing the idea. The second effect always exists. However, the low self-opinion arises when it is also matched by the first effect. On the contrary, the possibility of a high self-opinion phase arises when the first effect is not present.

We conclude this section by presenting an important corollary and our second main result.

Corollary 1.2. *The expected performance of the agent is better under a higher self-opinion.*

To see this, first note that the supervisor induces a weakly higher number of rounds of experimentation under a prior $\beta' > \beta$. If $\beta_j^{NI} \leq \beta < \beta_{j+1}^{NI}$, then either $\beta_j^{NI} \leq \beta < \beta' < \beta_{j+1}^{NI}$ or $\beta' > \beta_{j+1}^{NI}$. In the former case, the agent experiments an equal number of rounds under the two beliefs. However, in the latter case, the agent experiments more often under belief β' than under β. The reason is that it is easier to support the mutual expectation of honesty and repeated experimentation under a higher belief so that the agent experiments weakly more often under β'.

However, this has consequences on the agent's overall performance. Honest feedback by the supervisor allows the agent to match her effort more closely to the actual potential maximizing the probability of success. If the agent abandons seeking supervision (and experiments one final round) in the kth round under belief β, then she should still be seeking honest supervision in the round k under belief β'. While the

agent with belief β exerts an inefficiently low amount of effort in a high potential idea in round k, an agent with β' will exert the efficient level of effort of 1. An inefficient level of effort reduces the probability of success in a high potential idea.

Finally, if the idea in round k is a low potential one, then an agent with lower belief exerts an inefficiently high level of effort in its implementation while the agent with a higher belief experiments again. Therefore, there is a magnifying effect of a higher belief that results from the combined effect of better experimentation and better implementation. Conditional on being high-ability, an a priori better agent who has a higher belief in her ability does better in expectation.

It is also worth noting that in this context an a priori better agent (who has a higher self-opinion) will face more "criticism" from the supervisor for the same reason. An agent with a higher belief in her ability receives discouraging messages more often conditional on producing the same number of low potential ideas. However, the agent's incentive to experiment more often arises precisely out of the supervisor offering honest criticism. In equilibrium, an agent with a higher belief expects to receive honest feedback more often and is therefore willing to experiment more often. In return, the supervisor expecting more experimentation offers more honest feedback to the agent. When the agent's belief is lower, he fears to discourage the agent with negative messages. In this sense, an agent with a higher belief is more receptive to criticism, and that increases her chances of being successful.

1.4.3 Welfare analysis

The previous result (Corollary 1.2) only talks about the benefit of a higher self-opinion. However, the agent also pays a higher cost under a higher self-opinion owing to the aforementioned magnifying effect. This particularly hurts a low-ability agent who only pays a higher costs of experimentation and/or implementation under a higher belief.

The first part of this section shows that the above is not a concern even when evaluating the agent's welfare under a higher self-opinion. We show, through a series of lemmas below that the ex-ante expected utility of the agent is always higher under a higher belief.[15] The reason is that under a higher belief the agent places a greater ex-ante weight on being high-ability and believes that she is less likely to find herself in the worst situation.

[15]The supervisor is always better off with a higher self-opinion agent because in expectation such an agent performs better. At the same time, the supervisor doesn't have to bear any costs.

The second part of the section then analyzes if holding an incorrect higher belief could also be welfare improving. Surprisingly, we show that this is possible. The reason is the discontinuous change in the supervisor's feedback strategy as he goes from babbling to honesty.[16]

Welfare effect of a correct increase in self-opinion

Lemma 1.4. *Any increase in the prior from β to β' within the region of beliefs $\beta_0^{FI} \leq \beta < \beta' < \beta_0^{NI}$, $\beta_0^{NI} \leq \beta < \beta' < \beta_1^{NI}$, and $\beta_j^{NI} \leq \beta < \beta' < \beta_{j+1}^{NI}$ for $j > 1$ is welfare improving for the agent.*

This lemma relates to increasing the beliefs of the agent in such a way that only the cost of exerting effort increases in the eventuality that the project is implemented with a low potential idea or after not seeking supervision. In such a situation, welfare may increase on account of better implementation (because of higher effort) but may reduce on account higher costs of implementing.

Lemma 1.5. *An increase in the prior from the region $\beta_0^{FI} \leq \beta < \beta_0^{NI}$ to the region $\beta_0^{NI} \leq \beta' < \beta_1^{NI}$ is welfare improving for the agent.*

When the belief increases in such a manner, the agent is expected to conduct a costly round of experimentation which she did not earlier. Moreover, she is not expected to receive any feedback in this round. At the same time, her optimal effort choice increases unambiguously which is both more costly and more beneficial in expectation. From Lemma 1.4, we know that increasing the effort is always welfare improving when the belief increases. In addition, the increase in belief also makes it worthwhile to conduct experimentation without supervision from Lemma 1.1. This leads to an overall increase in welfare.

Lemma 1.6. *Let $2c < q(1-(q+(1-q)k)^2)$. An increase in the prior from $\beta = \beta_{j+1}^{NI} - \epsilon$ to $\beta' = \beta_{j+1}^{NI}$ is welfare improving for the agent.*

Finally, this lemma establishes that just pushing up the belief from an arbitrary region $\beta_j^{NI} \leq \beta < \beta_{j+1}^{NI}$ to the next region $\beta_{j+1}^{NI} \leq \beta' < \beta_{j+2}^{NI}$ is welfare improving. In doing so, the agent is expected to pay not only an additional cost of experimentation c but also that of some minimal increase in effort cost in the event of implementing without supervision.

[16] We prove all the statements here assuming that $c \geq \frac{\kappa k - (\kappa k)^2}{2}$ or that the truth-telling threshold $\beta^{\text{truth}} \leq \beta_1^{NI}$. However, this is not required as the proofs go through with a higher β^{truth} as well.

Proposition 1.4. *Let $2c < q(1-(q+(1-q)k)^2)$. An increase in the prior from β to β' is welfare improving for the agent.*

The above proposition combines the information from the three lemmas and concludes that any increase in prior is welfare improving. This highlights the importance of agent's self-opinion – the agent's confidence in her ability is critical for the overall success of the project.

Welfare effect of overconfidence

Still more interesting is to explain the effect of overconfidence in our environment. To introduce the notion of overconfidence, consider the following. Let the agent and the supervisor hold a common prior belief β about the agent's ability when the true belief is b.

Definition 1.1. *The agent and the supervisor are overconfident about the agent's ability if $\beta > b$.*

Under the above definition of overconfidence, we prove the following proposition:

Proposition 1.5. *Overconfidence is sometimes, but not always, welfare improving.*

To understand the intuition, consider the welfare of the agent when the correct belief is $b = \beta_1^{NI} - \epsilon$ but the common prior is β_1^{NI}. In such a situation, her overconfidence will drive her to experiment once with a round of honest feedback by the supervisor (and then potentially once more without any feedback). This would not have been possible under the true belief wherein she would have simply experimented without any feedback. However, the discontinuous benefit that arises from the change in supervisor's feedback strategy at a higher belief (i.e. receiving honest feedback) outweighs the additional cost that the agent pays for an additional round of experimentation.

In fact, she is able to reduce her inefficient cost of implementation when the supervisor honestly reveals that her idea was a low potential one under the overconfident belief. To see this note that under the true belief she would exert $(\beta_1^{NI} - \epsilon)(q+(1-q)k)$. Whereas under the overconfident belief she would exert $\beta_0^{NI}(q+(1-q)k)$. Thus, overconfidence (and holding an incorrect self-opinion) can be welfare improving.

However, the above argument relies on the discontinuous change in behavior of the supervisor at the threshold. It then follows that when the supervisor's behavior does not change, there might not be a benefit of being overconfident. To illustrate this, we show that overconfidence is welfare reducing when the common prior is β_0^{NI} but the

true belief is any $b < \beta_0^{NI}$. In such a situation, holding the incorrect belief only adds to an added cost of experimentation and implementation without any corresponding benefit. Contrasting this with Lemma 1.6, it is immediate to see that overconfidence is different from a correct increase in belief.

1.5 Extensions

1.5.1 Benevolent supervisor and time-constrained players

We start out by discussing what happens when the supervisor also bears the cost of experimentation and implementation. In some situations, it is possible that a benevolent supervisor partially internalizes the costs borne by the agent. Such internalization may arise from the expert's (i.e. the supervisor's) prior experience from when he as an apprentice (agent), or simply because he works on the project with the agent.

For the two players $i \in \{A, S\}$, agent (A) and supervisor (S), let the cost of experimentation be c_i and the cost of implementation be $\frac{\phi_i e^2}{2}$. The difference between these costs for the two players captures any preference conflict between them. In so far as $c_S < c_A$ and $\phi_S < \phi_A = 1$, the preference conflict persists. For a given $(c_S, \phi_S) > 0$, there will be a "full information" threshold for the supervisor as well. Call this threshold β_{S0}^{FI}. This reflects the preferences of the supervisor and determines what are the maximum number of rounds the supervisor desires the agent to experiment (or the belief threshold equivalently) with full information about the potential of the ideas.

In the limiting case of $c_S = \phi_S = 0$ studied in the main text, this threshold did not exist – the supervisor wanted the agent to continue experimenting with complete information until she ended up with a high potential idea. However, when $c_S < c_A$ and $\phi_S < \phi_A$, we have $\beta_{S0}^{FI} < \beta_{A0}^{FI}$ so that the supervisor would still like the agent to experiment more than she would like. In this case, all our results from the main text go through as the fear of discouragement and the agent abandoning experimentation still persists.

One possible interpretation of such a situation are time-constrained players. To keep things simple, let $\phi_S = \phi_A = 1$ so that the supervisor fully internalizes the time cost of implementing to the agent. Now let c_S denote the time cost that the supervisor pays for providing feedback to the agent. This could happen when the supervisor has some alternate tasks to perform or requires time to understand the true potential of the agent's ideas. The following proposition follows from our discussion.

Proposition 1.6. *Let $\phi_S = \phi_A = 1$.*

1. *If $c_S < c_A$ then Propositions 1.1, 1.2 and 1.3 capture the optimal strategies of the agent and the supervisor.*

2. *If $c_S \geq c_A$ then the supervisor offers honest feedback until he reaches the belief β_{S0}^{FI} and the agent experiments with ideas till that point absent a high potential idea.*

The intuition is as follows. When the supervisor is time-constrained, he cares both about the success *and* about costly supervision from the agent experimenting in pursuit of success. In turn, this eliminates the fear of discouragement. Notably, now it is more costly for the supervisor to keep offering feedback beyond a point over letting the agent implement a low potential idea. We can then get honest equilibria for some additional ranges of beliefs. Thus, a more time-constrained supervisor can potentially offer more honest feedback. The next corollary identifies the condition that makes this possible.

Corollary 1.3. *Let $\phi_S = \phi_A = 1$. If $c_S \geq c_A$ such that $\beta_{S0}^{FI} < \beta_1^{NI}$ then the region of beliefs where honest equilibria exist is larger in the case of $c_S \geq c_A$ than $c_S < c_A$.*

Observe that in the case of $c_S < c_A$ honest equilibria exist in the region of beliefs above β_1^{NI} (depending on c_A). But from the above proposition, honest equilibria in the case of $c_S \geq c_A$ exist starting from β_{S0}^{FI}. Thus, the latter case provides the possibility of more honesty if $\beta_{S0}^{FI} < \beta_1^{NI}$. However, since there is no closed form solution of β_{S0}^{FI}, it is not straightforward to translate this into a condition with only the costs.

Finally, note that if the supervisor does not internalize the cost of exerting effort, there is no benefit (in terms of more honest equilibria) of even partially internalizing the costs of experimentation.

Proposition 1.7. *If $\phi_S = 0$, then the equilibrium strategies are given by Propositions 1.1, 1.2 and 1.3.*

To understand the intuition, let $c_S = c_A$ and consider whether honesty is an equilibrium strategy for $\beta_{A0}^{FI} \leq \beta < \beta_{A1}^{FI}$(after all, if the supervisor internalizes the full cost of experimentation then the belief thresholds should match). At this belief, if the supervisor is expected to be honest, then following a negative message the agent abandons experimentation and exerts a low effort level on the idea. If instead, she receives a positive message, she exerts 1 on her idea. Now, for a supervisor who has

seen a low potential idea and does not internalize the cost of implementation, there is a strictly positive deviation to giving a positive message. This breaks down the honest equilibrium (and the existence of β_{S0}^{FI}).[17]

The issue arises here because the supervisor wants the agent to exert the maximal effort independent of the potential of the idea produced. The supervisor fears discouragement leading to lower effort in implementation which precludes honesty.

1.5.2 Perfect recall of previous ideas

Here we describe what happens if the agent and the supervisor have perfect recall of all the previous ideas. In such a situation, the supervisor can potentially make announcements about each of the previous ideas after each round of experimentation. Given our attention to pure strategies, there are two kinds of honest and informative strategies that a supervisor may employ: immediate honesty and delayed honesty.

In the immediately honest strategy, the supervisor reveals to the agent the outcome of her experimentation immediately after she experiments. This is implicitly what we assumed all throughout Section 1.4. In a strategy of delayed honesty, the supervisor provides uninformative messages for certain rounds and then reveals honestly some or all the previous outcomes. Observe that a variety of delayed honesty strategies are possible – the supervisor may babble for any arbitrary number of rounds and then provide information for any arbitrary number of those rounds, and this may change over time. If the supervisor reveals to the agent the $\tau' \leq \tau$ outcomes of her experimentation after τ rounds starting from a prior β_t the agent's updated posterior in round $t+\tau$ is

$$\beta_{t+\tau}^{\ell} = \frac{(1-q)^{\tau'}\beta_t}{1 - q\beta_t \sum_{s=0}^{\tau'-1}(1-q)^s} \text{ if } m_t = \ell \text{ for all } \tau' \text{ ideas, and} \tag{1.3}$$

$$\beta_{t+\tau}^{\hbar} = 1 \text{ otherwise.} \tag{1.4}$$

The case of $\tau = \tau' = 1$ corresponds to immediate honesty where the agent expects the supervisor to reveal the outcome of the experimentation immediately after each round of experimentation. All other cases fall under delayed honesty.

In case the supervisor is expected to babble, the agent's posterior belief is the same as her prior belief. We will assume that when the supervisor is expected to lie about an idea the agent does not consult the supervisor regarding that idea. This rules out the possibility of the supervisor privately learning and not revealing to the agent

[17] It is possible to derive a belief threshold above which the supervisor is expected to be honest in equilibrium for a generic ϕ_S and given c_S and c_A. This is necessarily different from β_{S0}^{FI} because that is contingent on the equilibrium best response of the agent to the supervisor's strategy.

the outcome, and the arising deviations.[18]

Note first that the result of Proposition 1.1 remains unaltered. If the agent does not want to experiment with an immediately honest strategy, she does not want to experiment with a delayed honesty strategy. By experimenting when the supervisor is expected to reveal the outcomes after a delay, the agent only bears a higher cost of experimentation to receive feedback when she is almost convinced that she cannot produce a high potential idea. Thus, implementing the outside option is the best response of the agent, and all strategies of information revelation are an equilibrium.

Corollary 1.4. *Under perfect recall of ideas, for any belief $\beta_0^{FI} \leq \beta < \beta_1^{NI}$, babbling is the unique equilibrium strategy.*

The result of babbling being a unique equilibrium in the region of beliefs $\beta_0^{FI} \leq \beta < \beta_1^{NI}$ *even* under perfect recall follows almost directly from Proposition 1.2. To illustrate this point, start again with a prior belief $\beta_0^{FI} \leq \beta_1 < \beta_1^{FI}$. In the absence of commitment, a supervisor who observes only low potential ideas from all the experimentation rounds (after delaying) is tempted to deviate and call any arbitrary idea a high potential one. This is for the same reason as before – when such a message is believed, the agent exerts maximal effort on such an idea assuming it is a high potential one. The supervisor gains from such a deviation because he increases the effort of the agent on a low potential idea in the absence of more experimentation. As a result, babbling is the unique equilibrium and the agent best responds by implementing the low potential outside option idea. The same reasoning can then be extended to all the beliefs which when updated with a negative message lead to the agent abandoning experimentation (as the supervisor is going to babble in the following round). This happens all the way up to the belief β_1^{NI} as before.

For beliefs above β_1^{NI}, we have already identified the condition for *immediate honesty* to arise in Proposition 1.3. It is, however, possible to have other equilibria with some delayed honesty. We identify here a critical feature of such equilibria (if they exist) that allows us to compare it with the immediately honest equilibrium.

Observation 1.1. *In a delayed equilibrium, the supervisor can only induce as many rounds of experimentation as the ones for which he provides honest feedback eventually.*

[18] A formal definition of strategies in this case is complicated. But it is easy to describe what a strategy for the two players are in words. A strategy for the supervisor when the agent consults him in round t is a mapping from all the ideas she observes to the set of messages, one for each round of experimentation. A strategy for the agent in round t is a mapping from the observed messages to a decision to experiment again or implement. If she decides to implement, she must also decide which idea to implement given the message history.

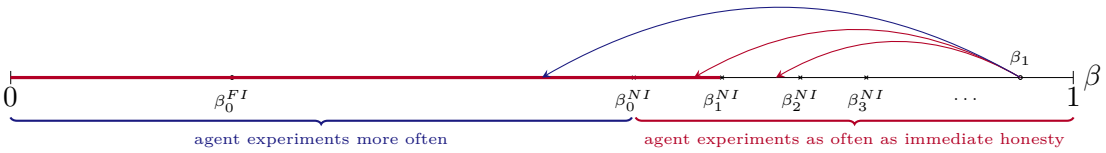


Figure 1.6: Terminal belief possibilities in potential delayed equilibria

The above observation merely states that if the supervisor never provides feedback on some rounds of experimentation that the agent performs, then the agent has no incentive to experiment. Since the agent never consults the supervisor for rounds in which he is expected to babble, there is no benefit to the agent from experimenting these extra rounds. This allows us to focus attention on those strategies in which the outcome of all the rounds of experimentation is eventually revealed.

Proposition 1.8. *The number of rounds of experimentation that an equilibrium strategy of delayed honesty induces can be no more than that induced by the equilibrium immediate honesty strategy.*

What matters when evaluating the supervisor's incentive to be honest at the time of final revelation is the belief from truthfully announcing that all the ideas produced are low potential. Say that the belief after such a revelation at round τ is β_τ^ℓ. This belief can be in one of the following three ranges: $\beta_\tau^\ell \geq \beta_1^{NI}$, $\beta_0^{NI} \leq \beta_\tau^\ell < \beta_1^{NI}$ or $\beta_\tau^\ell < \beta_0^{NI}$ (See Figure 1.6).

Observe that a terminal belief in the first and second range can also be attained by an immediately honest strategy, which is also an equilibrium. For any prior β_1, for the agent to experiment more rounds than what she does under immediately honest strategy her terminal belief after all the revelations should fall in the third case, i.e. $\beta_\tau^\ell < \beta_0^{NI}$. However, we argue that such a strategy cannot be an equilibrium. This is for the same reason as before – a supervisor who has only observed low potential ideas will prefer to deviate and claim any one of the ideas to be of high potential than inducing the agent to stop experimenting with a lower belief where the supervisor only babbles. Thus, equilibrium experimentation possibilities under perfect recall can be no more than those under limited recall.

1.5.3 Alternate interpretations

Our model more generally speaks to the following type of settings. An informed sender of information (supervisor) communicates with a less informed receiver of

information (agent) who needs to take a costly action dynamically. Consider, for instance, an entrepreneur who works on a project experimenting with ideas, *privately observing their potential*, and implementing one of them. However, she relies on the finances of a venture capitalist (VC) who pays for such experimentation and implementation. While the entrepreneur would prefer to continue experimenting until she receives a high potential idea, the VC would like to cut funding for experimentation when he is sufficiently pessimistic.

In such a setting, the entrepreneur is the supervisor, while the VC is the agent.[19] Costs c and $e^2/2$ are the money promised by the agent to the supervisor for experimenting with and implementing ideas. Let $\alpha \in \{0, q\}$ be the state of the project which is determined ex-ante and remains persistent but potentially unknown to both the parties. $\theta \in \{\ell, \hbar\}$ denotes the potential of the idea produced by the entrepreneur. The VC decides in each period, whether to fund experimentation for one extra round or force the entrepreneur to implement the last idea.

We then provide answers to the following questions: When can the entrepreneur credibly release information? How many chances of experimentation can the entrepreneur extract from the VC with her revelation strategy? Notably, our inefficiency result shows that even though the VC would like to continue financing the entrepreneur's experimentation and the entrepreneur would like to continue experimenting, she calls off the project too early. However, there are benefits to be had from the VC both correctly and incorrectly believing that the project is good.

1.6 Conclusion

In this paper, we showed how an employee responds to criticism influences whether she receives feedback or not. Supervisors may not provide honest feedback to employees who do not believe in their ability. In turn, this hurts their performance and potentially their future careers. Moreover, it also hurts organizations as the supervisors provide inefficiently low levels of honest feedback. In this sense, organizations should seek to hire employees that *believe* in their ability to succeed. In fact, our model shows that overconfidence can sometimes be welfare-improving.

Our results are based on a model of feedback provision in an agent-supervisor environment. The agent experiments with ideas to try to solve a problem at hand and a supervisor offers feedback on whether her ideas have the potential to be successful. We

[19] Which player is the agent and which one is the supervisor is *not* determined by who is experimenting and implementing, but by who holds the information and who pays for the action.

showed the results for when the supervisor has no commitment power and uses cheap talk messages to communicate with the agent. We identified the region of beliefs for which the supervisor could only uniquely babble in equilibrium leading to inefficiency in the relationship. Driven by the fear of discouraging the agent to the point of abandonment of experimentation, the supervisor is not able to offer any credible information to the agent. We then showed if there are possible equilibria in which the supervisor can honestly communicate his information to the agent. A necessary and sufficient condition for honesty above the babbling threshold was found to be the costs of experimentation being sufficiently high.

However, our analysis focused only on pure strategy equilibria. The problem involving mixed strategies is a complicated one that requires determining how the agent responds to the current message when, in the future, there can be more mixing. Our work shows the further scope of looking at mixed communication strategies in such dynamic environments in the absence of commitment. One may also think of introducing new complications in the model such as those involving different priors of the agent and the supervisor.

1.7 Appendix

A Proofs from the main text

We present general proofs in mixed strategies, wherever we can. The first section provides some new mathematical notation for this purpose.

Mathematical notation for mixed strategies

We focus attention on limited recall of previous ideas so that when the agent experiments one more round, she does not recall the previous ideas she has worked on. As a result, the supervisor does not need to make back dated messages about all the previous ideas. A strategy for the agent ρ_t in round t is a mapping from the last observed message to a possible mixed decision to continue experimenting with ideas or implementing the last one . We let

$$\rho_t^{m_{t-1}} = \Pr(I_t = 1 \mid m_{t-1})$$

be the probability that the agent decides to implement the project following the last message.

Similarly, when the supervisor is called upon, a strategy for the supervisor σ_t in round t is a mapping from the last idea to a possible mixed message about its potential. We let

$$\sigma_t^{\theta_t} = \Pr(m_t = \theta_t \mid \theta_t)$$

be the probability of the supervisor being honest about the potential of the observed idea. Depending on the expected strategy of the supervisor, the agent conditions her action only on the last message received.

Let the sequence $\hat{\boldsymbol{\sigma}} = \{\hat{\sigma}_t^{\hbar}, \hat{\sigma}_t^{\ell}\}_{t=1}^{T}$ denote the conjectured strategy of the supervisor, and let $\hat{\boldsymbol{\rho}} = \{\hat{\rho}_t^{\hbar}, \hat{\rho}_t^{\ell}\}_{t=1}^{T}$ denote the conjectured strategy of the agent. Given the conjectured strategy of the supervisor, the agent updates beliefs about the two unknowns – her ability and the potential of her previous ideas. The belief about her ability is β_t. Let the belief about whether her idea was as announced by the supervisor be denoted by λ_t. Observe that:

1. the public history h_t^A at the beginning of round t can be summarized by the current public belief β_t about the ability of the agent and by the belief about the true potential of the last idea produced λ_t, while

2. the private history of the supervisor h_t^S at the beginning of round t can be summarized by the current private belief β_t about the ability of the agent.[20]

We can now informally describe the notion of equilibrium. We say that a pair of sequences of conjectured strategies $\boldsymbol{\sigma}$ and $\boldsymbol{\rho}$ constitute an equilibrium if (1) they are both the best responses to each other given the beliefs β_t and λ_t for each t, and (2) the beliefs β_t and λ_t are consistent with what the players are conjectured to do, i.e. $\boldsymbol{\sigma}$ and $\boldsymbol{\rho}$. Strategies expressed in the text without a hat constitute an equilibrium.

When both the messages are expected in equilibrium, either one of the messages will lead to a higher and the other to a lower β_t, or β_t remains the same with both the messages. We will call the former *informative* strategy and the latter *babbling* (or lying) strategy. The supervisor is expected to babble in equilibrium in round $t-1$ if $\hat{\sigma}_{t-1}^{\hbar} = 1 - \hat{\sigma}_{t-1}^{\ell}$, i.e. when the probability with which the supervisor is expected to reveal a true high potential idea is the same as the probability with which the supervisor incorrectly calls a low potential idea a high one. Thus, the agent is equally likely to get a positive or a negative message, and in turn does not learn from the messages. When the supervisor is expected to be informative, we will assume without loss of generality that he does so by increasing the posterior after a positive message of $m_{t-1} = \hbar$ (and the posterior beliefs fall after a negative message $m_{t-1} = \ell$). So, we assume that $\hat{\sigma}_{t-1}^{\hbar} > 1 - \hat{\sigma}_{t-1}^{\ell}$ for informativeness.

We will restrict attention here to informative strategies in which $\sigma^{\hbar} = 1$, i.e. the supervisor always truthfully announces that the project has a high potential to succeed when he sees so. The supervisor cannot credibly commit to lying when $\theta_t = \hbar$. In any informative strategy, a positive message $m_t = \hbar$ should increase the posterior belief β_{t+1} of the agent. When the supervisor sees $\theta_t = \hbar$, he has no incentive to discourage the agent. If discouragement leads to another round of experimentation, then the supervisor faces the risk of abandoning the current high potential idea and never getting a new one. Alternately, if discouragement leads to implementation then she will do so with a lower effort. In neither case a supervisor who has observed a high potential idea is better off discouraging the agent. Going forward, we assume $\sigma_t^{\hbar} = 1$, and with some

[20]Note that we are currently not making any notational distinction between the private and the public beliefs about ability. This is to keep things simple. The two will coincide as long as the supervisor is honest. When the supervisor is not honest, the beliefs diverge only when the agent best responds to a dishonest message by experimenting again. This plays a role only in checking for deviations when constructing other informative equilibria.

replace σ_t^ℓ with σ_t. Then the posterior beliefs about ability is

$$\beta_t^\ell = \frac{(1-q)\beta_{t-1}}{1-q\beta_{t-1}} \tag{1.7.A.1}$$

$$\beta_t^\hbar = \frac{(1-\hat{\sigma}_{t-1}(1-q))\beta_{t-1}}{1-\hat{\sigma}_{t-1}(1-q\beta_{t-1})} \tag{1.7.A.2}$$

where $\beta_t^{m_{t-1}} = \Pr(\alpha = q|m_{t-1})$ is the posterior belief of the agent about her ability after receiving message m_{t-1} given the conjecture $\hat{\sigma}_{t-1}$. And

$$\lambda_t^\ell = 1 \tag{1.7.A.3}$$

$$\lambda_t^\hbar = \frac{q}{q+(1-\hat{\sigma}_{t-1})(1-q)} \tag{1.7.A.4}$$

where $\lambda_t^{m_{t-1}} = \Pr(\theta_{t-1} = m_{t-1}|m_{t-1})$ is the belief about whether the supervisor's message m_{t-1} matches the true potential of the idea given the conjectured $\hat{\sigma}_{t-1}$.

Thus, the value of a negative message under any informative strategy is the same as in a truth-telling strategy. When an agent receives $m_t = \ell$ then she can be sure that $\theta_t = \ell$ and she revises her belief about her ability downwards to the maximum extent. Under this condition, the agent must decide what to do following a message of $m_t = \hbar$ since a positive message cannot be trusted.

Proof of Lemma 1.2

Proof. Part 1: Existence of β_0^{FI}

For a given set of parameters, there is no straightforward closed form solution to the equation in condition C2. We therefore need to establish the existence of belief threshold(s). First, it can be verified that both the LHS and RHS of condition C2 are monotonically increasing and convex in β. We have

$$\frac{\partial \text{LHS}}{\partial \beta} = \frac{q}{2} + \frac{(k\beta')^2}{2}\left(\frac{2}{\beta} - q\right) > 0$$
$$\frac{\partial^2 \text{LHS}}{\partial \beta^2} = \frac{k^2(1-q)^2}{(1-\beta q)^3} > 0$$

and

$$\frac{\partial \text{RHS}}{\partial \beta} = k^2\beta \geq 0$$
$$\frac{\partial^2 \text{RHS}}{\partial \beta^2} = k^2 > 0.$$

Second, we show that if $2c < q(1-k^2)$ then the threshold belief β_0^{FI} is unique.

Consider the range of beliefs $0 \leq \beta \leq 1$. Since $c > 0$ and LHS at $\beta = 0$ is zero, RHS cuts the LHS from above at least once. Now, under the assumption $2c < q(1-k^2)$, it can be verified that RHS at $\beta = 1$ is lower than LHS at $\beta = 1$. Since both LHS and RHS are monotonically increasing, they must intersect at exactly one point. Call that belief β_0^{FI}. Thus, β_0^{FI} exists and is unique.

Third, we need to show that if there exists a unique threshold belief β_0^{FI}, then $2c < q(1-k^2)$. If there is a unique belief threshold then it must be the case that there is a unique point of intersection of LHS and RHS in condition C2. Again, RHS cuts the LHS from above because at $\beta = 0$ $c > 0$. Therefore, given the monotonicity of the two functions, a sufficient condition for uniqueness is $\text{LHS}|_{\beta=1} > \text{RHS}|_{\beta=1}$. This gives $\frac{q}{2} + (1-q)\frac{k^2}{2} > \frac{k^2}{2} + c$, which can be rearranged to $2c < q(1-k^2)$.

Lastly, we need to show that the agent does not experiment when $2c \geq q(1-k^2)$. This is so because then the RHS is always above the LHS, so that even experimentation once is not beneficial. When $2c \geq q(1-k^2)$ we have that $\text{LHS}|_{\beta=1} \leq \text{RHS}|_{\beta=1}$. Given that both LHS and RHS of condition (C2) are increasing convex functions, a concern is that there might be two points of intersection. However, it is easy to verify that the slope of the RHS is lower than the slope of the LHS at both $\beta = 0$ and $\beta = 1$. This precludes such a possibility. Therefore, the agent does not want to experiment when $2c \geq q(1-k^2)$ as the RHS is always above the LHS.

Part 2: Optimal decision rule I_t^{FI}

Condition C2 is the condition for experimenting in the worst case scenario, that is when the agent knows she is going to stop after another ℓ idea. Therefore, it follows that $I_t^{FI} = 0$ in $\beta \geq \beta_0^{FI}$ if $\theta_{t-1} = \ell$, i.e the agent continues experimenting.

Next, note that the agent cannot continue experimenting forever after ℓ ideas because at the limit the value of experimentation goes to $-c$. This is so because at the limit the belief about ability goes to zero while the cost of experimentation is a positive constant. Thus, what we need to show is that the agent does not want to experiment even once when condition C2 does not hold, i.e. $I_t^{FI} = 1$ for beliefs $\beta_t < \beta_0^{FI}$ if $\theta_{t-1} = \ell$ is the optimal decision rule.

Suppose not. Say that for some belief $\tilde{\beta} < \beta_0^{FI}$, it does not pay to experiment just once but it pays to experiment at least $\tilde{T}$ times and then stop (Note from above, she does not want to experiment forever). Now at round $\tilde{T} - 1$ when belief is $\tilde{\beta}_{\tilde{T}-1}$ it

must be that condition C2 holds i.e.

$$\frac{\tilde{\beta}_{\tilde{T}-1}q}{2} + (1-\tilde{\beta}_{\tilde{T}-1}q)\frac{(\tilde{\beta}_{\tilde{T}}k)^2}{2} \geq \frac{(\tilde{\beta}_{\tilde{T}-1}k)^2}{2} + c$$

But now since $\tilde{\beta}_{\tilde{T}-1} \leq \tilde{\beta} < \beta_0^{FI}$ and we know that for any belief $\beta < \beta_0^{FI}$ condition C2 does not hold, this is a contradiction.

Finally, we have already shown the proof of the choice of e^{FI} in the main text. □

Proof of Lemma 1.3

Proof. Fix the parameters such that $2c < (q+(1-q)k)^2 - k^2$. Since, $q(1-k^2) > (q+(1-q)k)^2 - k^2$, both β_0^{NI} and β_0^{FI} exist and are unique. To compare β_0^{NI} and β_0^{FI}, we only need to compare the LHS of the equation that defines condition (C1) with the LHS of the equation that defines condition (C2). We can then compare them with a common RHS.

Observe that the LHS of both the conditions are increasing and convex in β. Further, as $\beta \to 0$ the LHS in both the conditions also tend to zero. Thus, to establish a relationship between them it is sufficient to look at the behaviour of the LHS as $\beta \to 1$. This is equal to $\frac{(q+(1-q)k)^2}{2}$ for condition C1 and $\frac{q+(1-q)k^2}{2}$ for condition C2. Again, it can be shown that $\frac{(q+(1-q)k)^2}{2} < \frac{q+(1-q)k^2}{2}$ which is equivalent to $q(1-k^2) > (q+(1-q)k)^2 - k^2$. This implies that the LHS of condition C1 lies below the LHS of condition C2 for all $\beta > 0$. Thus, $\beta_0^{NI} > \beta_0^{FI}$. □

Proof of Proposition 1.2

Proof. We prove this statement in steps by considering different regions of starting prior β_1. There exists a $j \geq 0 \in \{0, 1, 2, \dots\}$ where belief β_j^{FI} is such that $\beta_j^{FI} < \beta_0^{NI} \leq \beta_{j+1}^{FI}$. The value that j takes depends on the parameters.

Step 1: Proving babbling is a unique equilibrium for $\beta_0^{FI} \leq \beta_1 < \beta_1^{FI}$

Consider any informative strategy $\hat{\sigma}_1 \in (0, 1]$ including the truth-telling strategy. In any such strategy a message $m_1 = \ell$ is only used when $\theta_1 = \ell$. So the agent believes such a message ($\lambda_2^\ell = 1$) with the posterior about ability $\beta_2^\ell < \beta_0^{FI}$ which makes the agent experiment only once at $t = 1$ and then exert $e = \beta_2^\ell k$ (see Proposition 1.1). A message $m_1 = \hbar$ instead leads to a higher belief $\beta_2^\hbar \in (\beta_1, 1]$, which can either push the agent to implement her idea with a higher effort or to experiment again (depending on $\hat{\sigma}_1$ and $\hat{\sigma}_2$).

If the agent best responds to $m_1 = h$ implementing her idea, she exerts effort $e = \beta_2^{\hbar}(\lambda_2^{\hbar} + (1 - \lambda_2^{\hbar})k) > \beta_2^{\ell}k$. In this case, the supervisor type $\theta_1 = \ell$ is better off deviating and sending a message $m_1 = \hbar$ and getting a higher expected probability of success of $\beta_2^{\hbar}\beta_2^{\ell}(\lambda_2^{\hbar} + (1 - \lambda_2^{\hbar})k)k$ instead of $(\beta_2^{\ell}k)^2$. If the agent best responds to $m_1 = h$ by experimenting again, then also the supervisor type $\theta_1 = \ell$ is better off always sending the message $m_1 = h$. This is because the supervisor always prefers experimentation when the current idea is low potential. Thus, the supervisor has an incentive to deviate in either case.

Thus, only the babbling strategy remains which is always an equilibrium. The agent's equilibrium strategy is to implement her outside information idea, i.e. $I_1 = 1$ with $e = \beta_1 k$ since $\beta_1 < \beta_0^{NI}$(see Lemma 1.1).

Step 2: Proving babbling is a unique equilibrium for $\beta_1^{FI} \leq \beta_1 < \beta_0^{NI}$

If $j = 0$, then either $\beta_0^{FI} \leq \beta_1 < \beta_0^{NI} < \beta_1^{FI}$ or $\beta_0^{FI} < \beta_0^{NI} \leq \beta_1 < \beta_1^{FI}$. In either case, the scenario highlighted in Step 2 does not exist. Step 1 is sufficient in this case.

If $j = 1$ then it is enough to show that babbling is the unique equilibrium in the range $\beta_1^{FI} \leq \beta_1 < \beta_0^{NI}$ with the knowledge that if the posterior $\beta_2 < \beta_1^{FI}$ then the supervisor babbles (from Step 1 above). Note that any informative messaging strategy conjecture for $t = 1$ with $\hat{\sigma}_1 \in (0, 1]$ must lead to a posterior $\beta_2^{\ell} < \beta_1 < \beta_2^{\hbar}$. Now, as before the value of message $m_1 = \ell$ is the same as in truth-telling so that $\beta_2^{\ell} \in [\beta_0^{FI}, \beta_1^{FI})$. From Step 1 above, the supervisor is then expected to babble in $t = 2$ and the agent best responds by choosing to implement her low potential idea ($I_2 = 1$) from $t = 1$ with effort $e = \beta_2^{\ell}k$. A message $m_1 = \hbar$ again leads to a higher belief $\beta_2^{\hbar} \in (\beta_1, 1]$, which can either push the agent to implement her idea with a higher effort or to experiment again (depending on $\hat{\sigma}_1$ and $\hat{\sigma}_2$). As before now, the supervisor type $\theta_1 = \ell$ is better off deviating and sending a message $m_1 = \hbar$. Thus, babbling is the unique equilibrium strategy of the supervisor.

If $j \in \{2, 3, \dots\}$, then it needs to be shown that babbling is a unique equilibrium strategy in the ranges $\beta_1^{FI} \leq \beta_1 < \beta_2^{FI}$, $\dots$, $\beta_{j-1}^{FI} \leq \beta_1 < \beta_j^{FI}$ and $\beta_j^{FI} \leq \beta_1 < \beta_0^{NI}$. Consider first the range $\beta_1^{FI} \leq \beta_1 < \beta_2^{FI}$. Any posterior β_2^{ℓ} for priors $\beta_1^{FI} \leq \beta_1 < \beta_2^{FI}$ must map in to the range of beliefs highlighted in Step 1. This implies that supervisor type $\theta_1 = \ell$ cannot credibly commit to sending a message $m_1 = \ell$. Such a message leads to the agent implementing with effort $e = \beta_2^{\ell}k$. This makes babbling a unique

equilibrium strategy for $\beta_1^{FI} \leq \beta_1 < \beta_2^{FI}$. The same logic applies to all the ranges of prior belief up to β_j^{FI}. Then, in the range $\beta_j^{FI} \leq \beta_1 < \beta_0^{NI}$ the proof is identical to the above described $j = 1$ case.

Therefore, babbling is the unique equilibrium strategy of the supervisor and the agent does not experiment, i.e. $I_1 = 1$ and $a = \beta_1 k$.

Step 3: Proving babbling is a unique equilibrium for $\beta_0^{NI} \leq \beta_1 < \beta_1^{NI}$

For $j = 0$, we have already shown that babbling is a unique equilibrium strategy for $\beta_0^{FI} \leq \beta_1 < \beta_0^{NI} < \beta_1^{FI}$ or $\beta_0^{FI} < \beta_0^{NI} \leq \beta_1 < \beta_1^{FI}$. Note that since $\beta_0^{FI} < \beta_0^{NI}$, it must be the case that $\beta_1^{FI} < \beta_1^{NI} < \beta_2^{FI}$. So, it remains to show that babbling is unique for $\beta_1^{FI} \leq \beta_1 < \beta_1^{NI}$. This argument is the same as the one presented below.

Any informative mixing for $j \geq 1$ leads to $\beta_2^{\ell} < \beta_0^{NI}$. The supervisor babbles in the range of posteriors $\beta_0^{FI} \leq \beta_2^{\ell} < \beta_0^{NI}$ from Step 1 and 2 above (and for $j = 0$ case the supervisor babbles in the range $\beta_0^{FI} \leq \beta_2^{\ell} < \beta_1^{FI}$), and the agent chooses to implement thereafter (from Lemma 1.1). A message $m_1 = \hbar$, on the other hand, is believed and the agent best responds by either implementing with a higher belief or experimenting again. Therefore, the supervisor can do better by lying instead when he observes $\theta_1 = \ell$ when he is expected to be informative. □

Proof of Proposition 1.3

Proof. We prove the proposition in two parts.

Part 1: To show that if $\sigma_1 = 1$ is an equilibrium for $\beta_1^{NI} \leq \beta_1 < \beta_2^{NI}$, then it must be an equilibrium for all $\beta_2^{NI} \leq \beta_1 < 1$.

Consider the region of priors $\beta_2^{NI} \leq \beta_1 < \beta_3^{NI}$. We check whether $\hat{\sigma}_1 = 1$ is an equilibrium. Here, the supervisor has an incentive to reveal the truth about $\theta_1 = \ell$ if the expected probability of success by sending $m_1 = \ell$ is higher than that from sending the message $m_1 = \hbar$. If he sends a message $m_1 = \ell$, the agent at most experiments two more times - consulting the supervisor after one (which is at β_2^{ℓ} where the supervisor is again honest given the premise) and not doing so after the other. Therefore, the expected probability of success by sending $m_1 = \ell$ is

$$\beta_2^{\ell} q + (1 - \beta_2^{\ell} q)(\beta_3^{\ell})^2 (q + (1 - q)k)^2$$

By lying the supervisor convinces the agent that her idea has a high potential

to succeed ($\hat{\lambda}_2^{\hbar} = 1$)and that she is of ability q ($\hat{\beta}_2^{\hbar} = 1$). She then exerts $e = 1$ to implement her idea. However, the supervisor has an updated belief of β_2^{ℓ} knowing that $\theta_1 = \ell$. Thus, expected probability of success by sending $m_1 = h$ is $\beta_2^{\ell} k$.

The supervisor has an incentive to reveal the truth at this stage if

$$\beta_2^{\ell} k \leq \beta_2^{\ell} q + (1 - \beta_2^{\ell} q)(\beta_3^{\ell})^2 (q + (1-q)k)^2.$$

It is easy to check that the above condition is always holds with a strict inequality sign under Assumption (A). So, $\sigma_1 = 1$ is an equilibrium for $\beta_2^{NI} \leq \beta_1 < \beta_3^{NI}$.

Now, for any $\beta_1 \geq \beta_3^{NI}$ the supervisor can make the agent experiment (and if any of the following ideas has a high potential to succeed make them exert $e = 1$ on it) at least three more times by honestly revealing $\theta = \ell$. Given Assumption (A), this should always be an equilibrium.

Part 2: To show that $\sigma_1 = 1$ is an equilibrium for $\beta_1^{NI} \leq \beta_1 < \beta_2^{NI}$ if and only if $c \geq \frac{\kappa k - (\kappa k)^2}{2}$ where $\kappa \equiv \frac{k}{(q+(1-q)k)^2}$ and $k < (q + (1-q)k)^2$.

Suppose $c \geq \frac{\kappa k - (\kappa k)^2}{2}$ where $\kappa \equiv \frac{k}{(q+(1-q)k)^2}$ and $k < (q + (1-q)k)^2$. Now consider the conjectured strategy $\hat{\sigma}_1 = 1$ for $\beta_1^{NI} \leq \beta_1 < \beta_2^{NI}$. When the supervisor observes $\theta_1 = \ell$, his expected probability of success by sending message $m_1 = \ell$ is

$$(\beta_2^{\ell})^2 (q + (1-q)k)^2.$$

Following $m_1 = \ell$, the agent experiments once more but does not consult the supervisor thereafter. Thus, she implements her idea of unknown potential by exerting effort $e = \beta_2^{\ell}(q + (1-q)k)$. On the other hand by sending a message $m_1 = \hbar$ when the gent expects supervisor to be honest leads her to exert $e = 1$ in implementing a $\theta_1 = 1$ idea. This is so because she believes in the supervisor's message, $\hat{\lambda}_2^{\hbar} = 1$ and $\hat{\beta}_2^{\hbar} = 1$. The expected probability of success is then $\beta_2^{\ell} k$.

Truth-telling is an equilibrium if

$$\begin{aligned} &(\beta_2^{\ell})^2 (q + (1-q)k)^2 \geq \beta_2^{\ell} k \\ \implies &\beta_1 \geq \frac{k}{qk + (1-q)(q + (1-q)k)^2} := \beta^{\text{truth}}. \end{aligned}$$

$\hat{\sigma}_1 = 1$ is an equilibrium if for all $\beta_1 \in [\beta_1^{NI}, \beta_2^{NI})$, it is also the case that $\beta_1 \geq \beta^{\text{truth}}$. This can happen iff $\beta^{\text{truth}} \leq \beta_1^{NI}$. This condition then be rearranged given

β^{truth} from above and $\beta_0^{NI} = \left(\frac{2c}{(q+(1-q)k)^2-k^2}\right)^{\frac{1}{2}}$ (from Lemma 1.1), and using the fact that $\beta_1^{NI} = \frac{\beta_0^{NI}}{1-q(1-\beta_0^{NI})}$. This gives us

$$c \geq \frac{\kappa k - (\kappa k)^2}{2}$$

where $\kappa \equiv \frac{k}{(q+(1-q)k)^2}$ and we need that $k < (q+(1-q)k)^2$. But this is also our premise. Thus, $\sigma_1 = 1$ is an equilibrium.

Alternately, suppose $\sigma_1 = 1$ is an equilibrium for $\beta_1^{NI} \leq \beta_1 < \beta_2^{NI}$. Then it must be the case that $\beta_1 \geq \beta^{\text{truth}}$ for all $\beta_1 \in [\beta_1^{NI}, \beta_2^{NI})$. Specifically, it must be that $\beta_1^{NI} \geq \beta^{\text{truth}}$. This condition can then be rearranged to yield

$$c \geq \frac{\kappa k - (\kappa k)^2}{2}$$

where $\kappa \equiv \frac{k}{(q+(1-q)k)^2}$ and with an added constraint $k < (q+(1-q)k)^2$. □

Proof of Lemma 1.4

Proof. It is immediate to see that an increase in belief from β to β' such that

1. $\beta_0^{FI} \leq \beta < \beta' < \beta_0^{NI}$ is welfare improving. This is because $\frac{(\beta k)^2}{2} > \frac{(\beta' k)^2}{2}$ which we get by replacing the optimal effort $e = \beta k$ in the expected utility function.

2. $\beta_0^{NI} \leq \beta < \beta' < \beta_1^{NI}$ is welfare improving. This is because $\frac{(\beta(q+(1-q)k))^2}{2} > \frac{(\beta'(q+(1-q)k))^2}{2}$ which we get by replacing the optimal effort $e = \beta(q+(1-q)k)$ in the expected utility function.

Now consider an increase in belief from β to β' such that $\beta_j^{NI} \leq \beta < \beta' < \beta_{j+1}^{NI}$ such that $j > 1$. Denote the ex-ante expected utility or welfare of the agent at prior β by $W(\beta)$. We have that

$$\begin{aligned} W(\beta) =& \beta\frac{q}{2}[1+(1-q)+\cdots+(1-q)^{j-1}] - \beta c[1+(1-q)+\cdots+(1-q)^j] \\ &+ \beta(1-q)^j[Ke - \frac{e^2}{2}] - (1-\beta)[(j+1)c + \frac{e^2}{2}] \end{aligned}$$

where $K = q+(1-q)k$. Similarly, we can write $W(\beta')$ keeping in mind that the maximum number of attempts is still $j+1$.

Now, comparing term-by-term, it is obvious that everything other than the comparison of $\beta'(1-q)^j Ke' - ((1-\beta') + \beta'(1-q)^j)\frac{e'^2}{2}$ with $\beta(1-q)^j Ke - ((1-\beta) +$

$\beta(1-q)^j\frac{e^2}{2}$ in $W(\beta')$ is greater than that in $W(\beta)$. Thus it is sufficient to show that

$$\beta'(1-q)^jKe' - ((1-\beta') + \beta'(1-q)^j)\frac{e'^2}{2} > \beta(1-q)^jKe - ((1-\beta) + \beta(1-q)^j)\frac{e^2}{2}$$

which can be rearranged to

$$\beta'(1-q)^jKe' - (1-\beta'(1-(1-q)^j))\frac{e'^2}{2} > \beta(1-q)^jKe - (1-\beta(1-(1-q)^j))\frac{e^2}{2}$$

where $e = K\beta_{j+1}^{\ell}$ and $e' = K\beta_{j+1}^{\prime\ell}$.

Now it is easy to check that $Ke - \frac{e^2}{2}$ is increasing in beliefs. So that

$$\begin{aligned}
&Ke' - \frac{e'^2}{2} > Ke - \frac{e^2}{2} \\
\implies &Ke' - (1-\beta'(1-(1-q)^j))\frac{e'^2}{2} > Ke - (1-\beta(1-(1-q)^j))\frac{e^2}{2} \\
\implies &\beta'(1-q)^jKe' - (1-\beta'(1-(1-q)^j))\frac{e'^2}{2} > \beta(1-q)^jKe - (1-\beta(1-(1-q)^j))\frac{e^2}{2}
\end{aligned}$$

where in the second step the inequality is preserved because a greater number is added to the LHS than the RHS. And in the third step the inequality is again preserved because Ke' (which is greater than Ke) on the LHS is multiplied with a greater number than Ke in the RHS. Hence, the welfare has increased. □

Proof of Lemma 1.5

Proof. Using the language introduced in Lemma 1.4, we can write $W(\beta)$ and $W(\beta')$ where $\beta < \beta_0^{NI}$ and $\beta_0^{NI} \leq \beta < \beta_1^{NI}$ as

$$W(\beta) = \frac{(\beta k)^2}{2} \text{ and } W(\beta') = \frac{(\beta' K)^2}{2} - c$$

Now, if the agent finds herself in $[\beta_0^{NI}, \beta_1^{NI})$ then Condition (C1) must be slack. This means

$$\frac{(\beta' K)^2}{2} - c > \frac{(\beta' k)^2}{2} > \frac{(\beta k)^2}{2}$$

where the second inequality follows from the fact that $\beta' > \beta$. Hence, the welfare has increased. □

Proof of Lemma 1.6

Proof. We show here the proof of how an increase in belief from $\beta = \beta_1^{NI} - \varepsilon$ to $\beta' = \beta_1^{NI}$ is welfare improving. The general proof of an increase in the prior from $\beta_{j+1}^{NI} - \varepsilon$ to β_{j+1}^{NI} follows the same argument.

We can write the ex-ante expected welfare in the two cases as follows:

$$W(\beta_1^{NI} - \varepsilon) = (\beta_1^{NI} - \varepsilon)Ke - \frac{e^2}{2} - c$$
$$W(\beta_1^{NI}) = \beta_1^{NI}\frac{q}{2} + \beta_1^{NI}(1-q)Ke' - (c + \frac{e'^2}{2})(1 - \beta_1^{NI}q) - c$$

where $e = (\beta_1^{NI} - \varepsilon)K$ and $e' = \beta_0^{NI}K$.

Now, if $W(\beta_1^{NI}) > W(\beta_1^{NI} - \varepsilon)$, then substituting for e and e', letting $\varepsilon \to 0$, and simplifying the inequality by using $\beta_0^{NI} = \frac{(1-q)\beta_1^{NI}}{1-q\beta_1^{NI}}$ gives

$$\beta_1^{NI}\frac{q}{2} - c(1 - \beta_1^{NI}q) - \frac{(\beta_1^{NI}K)^2}{2} > -\frac{K^2}{2}(1-q)\beta_1^{NI}\beta_0^{NI}.$$

If the above inequality holds, then we are done.

Let $2c < q(1 - K^2)$. Under this assumption, Condition (C2) must hold in a way that k is replaced with K as

$$\beta_1^{NI}\frac{q}{2} - c > \frac{(\beta_1^{NI}K)^2}{2} - (1 - \beta_1^{NI}q)\frac{(\beta_0^{NI}K)^2}{2}.$$

Now, the inequality is preserved if the c on the LHS is reduced. Then rearranging gives

$$\beta_1^{NI}\frac{q}{2} - (1 - \beta_1^{NI}q)c - \frac{(\beta_1^{NI}K)^2}{2} > -(1 - \beta_1^{NI}q)\frac{(\beta_0^{NI}K)^2}{2}.$$

It is now straightforward to verify that $(1 - \beta_1^{NI}q)\frac{(\beta_0^{NI}K)^2}{2} = \frac{K^2}{2}(1-q)\beta_1^{NI}\beta_0^{NI}$, so that our original hypothesis on welfare comparison holds. □

Proof of Proposition 1.4

Proof. From Lemma 1.4 and 1.5, it is immediate that an increase in belief of up to, but not including the level β_1^{NI} is welfare improving. Now, from Lemma 1.6, an epsilon increase in belief that pushes the agent in to experimentation with supervision is also welfare improving. Finally, from Lemma 1.4, any increase in belief of up to but not including the level β_2^{NI} is welfare improving. This reasoning can then be extended for any increase in belief. □

Proof of Proposition 1.5

Proof. To prove the statement, we consider two particular situations, and show how in each the welfare at the correct and overconfident beliefs differ. Let $W(\beta; b)$ be the ex-ante expected utility of the agent when the common prior is β but the correct belief

is b.

Part 1: Showing that overconfidence can be welfare improving

Let $\beta = \beta_1^{NI}$ but $b = \beta_1^{NI} - \varepsilon$. The two expected utility functions can be written as

$$W(b;b) = \frac{(bK)^2}{2} - c$$
$$W(\beta_1^{NI};b) = \frac{bq}{2} + b(1-q)\beta_0^{NI}K^2 - (1-bq)(c + \frac{(\beta_0^{NI}K)^2}{2}) - c$$

We need to show if $W(\beta_1^{NI};b) > W(b;b)$. In order to do so, first observe that $\frac{bq}{2} > \frac{(bK)^2}{2}$. This follows immediately from Assumption (A). So if we are able to show that

$$b(1-q)\beta_0^{NI}K^2 - (1-bq)(c + \frac{(\beta_0^{NI}K)^2}{2}) \geq 0$$

then we are done. Rearranging the above and recognizing that $\frac{b(1-q)}{1-bq} = \beta_0^{NI} - \varepsilon'$ where $\varepsilon' \neq \varepsilon$, we need that

$$\frac{(\beta_0^{NI}K)^2}{2} \geq \varepsilon'\beta_0^{NI}K^2 + c$$

But we know from Condition (C1) that

$$\frac{(\beta_0^{NI}K)^2}{2} = \frac{(\beta_0^{NI}k)^2}{2} + c.$$

Therefore, it is possible to find an ε' (and consequently ε) such that welfare improves under overconfidence. This requires $\varepsilon' \leq \beta_0^{NI}\frac{k^2}{2K^2}$.

Part 2: Showing that overconfidence can be welfare reducing

Let $\beta = \beta_0^{NI}$ but $b < \beta_0^{NI}$. The two expected utility functions can be written as

$$W(b;b) = \frac{(bk)^2}{2}$$
$$W(\beta_0^{NI};b) = b\beta_0^{NI}K^2 - \frac{(\beta_0^{NI}K)^2}{2} - c$$

This time we need to show that $W(\beta_0^{NI};b) < W(b;b)$. Again using Condition (C1) to substitute for $-\frac{(\beta_0^{NI}K)^2}{2} - c$ in $W(\beta_0^{NI};b)$, we can reduce the above to

$$b < \beta_0^{NI}(\frac{2K^2}{k^2} - 1),$$

which must always be true because $\frac{2K^2}{k^2} - 1 > 1$. □

Proof of Proposition 1.6

Proof. Let $\phi_S = \phi_A = 1$. Consider the supervisor who has seen a $\theta_{t-1} = \ell$ and reveals it honestly to the agent. His value function is given by

$$\mathcal{V}_S^{\ell}(\beta_t) = \max\left\{\frac{(\beta_t k)^2}{2}\ ,\ \frac{\beta_t q}{2} - c_S + (1-\beta_t q)\mathcal{V}^{\ell}(\beta_{t+1})\right\}.$$

where the first term is the value that the supervisor would get if he gets the low idea implemented and the second term is what he would get if he gets experimentation again. Given his costs, he would then like the agent to continue experimenting for as long as

$$\frac{\beta q}{2} + (1-\beta q)\frac{(\beta' k)^2}{2} \geq \frac{(\beta k)^2}{2} + c_S\ ,$$

which gives a belief threshold β_{S0}^{FI}. However, under an honest strategy, the agent would like to continue experimenting for as long as

$$\frac{\beta q}{2} + (1-\beta q)\frac{(\beta' k)^2}{2} \geq \frac{(\beta k)^2}{2} + c_A\ ,$$

which gives a belief threshold β_{A0}^{FI}.

Now, if $c_S < c_A$ then $\beta_{S0}^{FI} < \beta_{A0}^{FI}$ so that the supervisor would like the agent to experiment beyond β_{A0}^{FI}. The supervisor then fears discouraging the agent through honest revelation for any prior belief that leads the agent to a belief lower than β_{A0}^{FI}. Therefore, the results of Propositions 1.1, 1.2 and 1.3 hold.

On the other hand if $c_S \geq c_A$, then $\beta_{S0}^{FI} > \beta_{A0}^{FI}$. The agent would like to experiment more than β_{S0}^{FI}. Consider a belief $\beta_{S0}^{FI} \leq \beta_1 < \beta_{S1}^{FI}$ and consider the expected strategy of honesty. When the supervisor has seen a low potential idea, then by announcing it truthfully he gets an effort of $e = \beta_2^{\ell} k$ which is also optimal from the point of view of the supervisor because $\phi_S = 1$. This is so because it is an equilibrium strategy for the supervisor to babble tomorrow. So, even though the agent at this stage would like to experiment again but in the absence of honesty tomorrow, and $\beta_2^{\ell} < \beta_{A0}^{NI}$ she prefers to implement. By deviating and calling it a high potential idea, he induces an effort of 1 on a low-potential idea. However, this is suboptimal from his perspective, since he would also the full cost of implementation. Thus, there is no incentive to lie and honesty is an equilibrium. □

Proof of Proposition 1.8

Proof. Consider a prior $\beta_j^{NI} \leq \beta_1 < \beta_{j+1}^{NI}$. In an immediately honest equilibrium strategy, the agent experiments for j rounds with subsequent messages $m = \ell$ before reaching the babbling region so that $\beta_0^{NI} \leq \beta_j^{\ell} < \beta_1^{NI}$. In addition, the agent experiments one extra round without supervision. Now, consider any strategy that reveals some $j' \leq j$ outcomes together. Let the round of eventual revelation be denoted by τ. Now, the agent is induced to experiment a higher number of rounds in this strategy iff $\beta_\tau^{\ell} < \beta_0^{NI} \leq \beta_j^{\ell}$. Say that this is the case. We determine whether such a strategy is an equilibrium.

Observe that at $\beta_\tau^{\ell} < \beta_0^{NI}$ the agent best responds by abandoning experimentation and implementing any one of her low potential ideas with an effort $\beta_\tau^{\ell} k$. If the supervisor is honest, his expected payoff is $(\beta_\tau^{\ell} k)^2$. By deviating, and calling any one of the low potential ideas a high one, the supervisor is able to induce an effort of 1 by the agent on that idea. This gives the supervisor an expected payoff of $\beta_\tau^{\ell} k$. Since the latter is greater than the former, such an eventually honest strategy cannot be an equilibrium. □

B Additional proofs not in the main text

Comparative statics of β_0^{FI}

Lemma 1.7. *β_0^{FI} is increasing in e, increasing in k, and decreasing in q.*

Proof. Consider, first, an exogenous increase in e. It is easy to verify that an increase in e raises the value of the RHS (i.e. of implementing the idea)in condition C2 for every belief level β. This raises the β_0^{FI}.

Second, consider the effect of an exogenous increase in k.

$$\frac{\partial \text{LHS}}{\partial k^2} = (1-\beta q)\frac{(\beta')^2}{2}$$
$$\frac{\partial \text{RHS}}{\partial k^2} = \frac{\beta^2}{2}.$$

Now, since $\beta > \beta'$ and $1 > \beta q$, $\frac{\partial \text{LHS}}{\partial k^2} < \frac{\partial^2 \text{RHS}}{\partial k^2}$. Thus, the value from implementing increases by more than the value from experimenting, which leads to a higher β_0^{FI}.

Finally, consider an exogenous increase in q. The RHS remains unchanged with an increase in q. For the LHS,

$$\frac{\partial \text{LHS}}{\partial q} = \frac{\beta}{2} - k^2\beta\beta'\left(1-\frac{\beta'}{2}\right).$$

This is positive if $\frac{1}{2} > k^2\beta'\left(1-\frac{\beta'}{2}\right)$, which is true since $\frac{\partial k^2\beta'\left(1-\frac{\beta'}{2}\right)}{\partial \beta'} = k^2(1-\beta') > 0$ and at the limits the inequality holds. As $\beta' \to 0$, we have that $k^2\beta'\left(1-\frac{\beta'}{2}\right) \to 0$ and as $\beta' \to 1$, $k^2\beta'\left(1-\frac{\beta'}{2}\right) \to \frac{k^2}{2}$. □

An exogenous increase in k makes executing a low potential idea more attractive and therefore, leads to a higher β_0^{FI} and reduces the incentives to experiment for long. The agent desires to finish the project with a sufficiently high belief so that he can exert a higher effort in implementing a low potential idea (if need be), thereby maximizing the probability of success even with a poor idea. Finally, an increase in q lowers the belief threshold. This is so because conditional on being of high-ability, a higher q increases the chances of coming up with a high potential idea. Therefore, in a world in which ability is unknown it makes experimentation more attractive and pushes the agent to experiment for longer.

Comparative statics of β_0^{NI}

It is straightforward to derive how β_0^{NI} behaves with a change in parameters. A decrease in the probability of coming up with a high potential idea q or an increase

in the cost of experimentation c has the effect of increasing the threshold. Finally, an increase in k can have a non-monotonic effect on β_0^{NI} depending on the initial value. For $k < \frac{1-q}{2-q}$, an increase in k decreases β_0^{NI}. For $k > \frac{1-q}{2-q}$, an increase in k increases β_0^{NI}. The intuition behind a non-monotonic relation between k and β_0^{NI} is as follows. k measures the success rate (for any given effort level) from a bad idea when the agent is of high-ability. When the agent does not observe the value of θ from experimentation, then she experiments only as a gamble (and this gamble is not worth taking more than once). When k increases from a sufficiently low k to begin with, it makes this gamble more attractive – the agent reasons that even if the gamble fails (i.e. $\theta = \ell$ is the outcome of the gamble), she is more likely to succeed because of a higher k. On the other hand, when k increases further from an already high level, then the gamble becomes less attractive. This is so because the agent already has an outside option $\bar{\theta} = \ell$ available which then becomes relatively more attractive to finish.

C Committed supervisor

A note on the enforcement of commitment

Here we present the case of the supervisor committing to an information policy before the agent starts experimenting with ideas. Before we do so, we should understand how such a commitment may be enforced. An information disclosure policy is a sequence of revelation strategies about the observed potential of ideas produced by the agent to which the supervisor is committed. One may imagine the policy as a sequence of public tests - the supervisor may or may not observe the true potential of the idea but he designs tests that will reveal to the agent (and to the supervisor) the true potential of the idea. Thus, commitment to information disclosure policy is akin to commitment to test designs. This interpretation is in the spirit of Kamenica and Gentzkow (2011) and Smolin (2017).

Another way in which such a commitment may be enforced is through "presentation" of ideas to multiple supervisors. Many co-supervisors rather than one main supervisor may work to discipline each other. This requires that if the optimal disclosure policy involves mixing by the supervisors then they all should agree on such a mixing and then enforce it (say by punishing deviations with full disclosure). Alternately, one supervisor's recommendation may be cross-examined by another supervisor who has also observed the agent's idea. However, these interpretations are not immediate and might not be realistic in many settings. An apprentice working on a project might only be assigned one expert due to cost concerns. It is also not obvious how a supervisor might commit to a test design that reveals his private information to the agent. Because of this limitation, we present the commitment case as an extension of the model in Section 1.4. We consider here only the flavour of an optimal policy by discussing the incentives of the supervisor and the agent, and showing how the supervisor can achieve better outcomes (relative to the equilibrium outcome) for both himself and the agent by committing to information disclosure policies.

Immediate honesty

Consider first the policy in which the supervisor is committed to revealing the true potential of the idea after each round of experimentation. We call this a policy of *immediate honesty*. As illustrated in Lemma 1.2 such a policy induces the agent to experiment with continued low potential ideas all the way down to the belief β_0^{FI}. It is immediate that the agent prefers to experiment more under this policy relative to

the equilibrium outlined in Proposition 1.3. Immediate honesty guarantees maximum possible learning to the agent and in the least cost, which allows the agent to match effort to the true potential of the idea. This helps retain the attractiveness of experimentation insofar as condition (C2) holds. The prior β_1 determines how many more rounds the agent ends up experimenting under this policy relative to the equilibrium.

That the supervisor prefers such a policy is not immediate in the region of beliefs in which the supervisor is honest in equilibrium as well. While on the one hand such a policy induces more experimentation (and therefore, a higher probability of the agent producing a high potential idea), it also depresses the effort of the agent when she does not ever produce a high potential idea. The agent exerts a higher effort in equilibrium on an idea of unknown potential (see Proposition 1.3) because of a higher belief. Let $\beta_1 > \beta_1^{NI}$ such that under both the equilibrium and the immediately honest policy the agent experiments for t rounds until β_1^{NI}, then in equilibrium the agent experiments for one additional round (without supervision) while under the immediately honest policy she does so for t' additional rounds with supervision. Note that t and t' are functions of β_1. The supervisor prefers the immediately honest policy over the equilibrium policy iff

$$\begin{aligned}(\beta_{t+1}^{\ell})^2(q+(1-q)k)^2 \; < \beta_{t+1}^{\ell}q + (1-\beta_{t+1}^{\ell}q)\beta_{t+2}^{\ell}q + \\ + \; (1-\beta_{t+1}^{\ell}q)(1-\beta_{t+2}^{\ell}q)\beta_{t+3}^{\ell}q + \\ + \ldots + (1-\beta_{t+1}^{\ell}q)(1-\beta_{t+2}^{\ell}q)\ldots(1-\beta_{t+t'}^{\ell}q)(\beta_{t+t'+1}^{\ell}k)^2.\end{aligned}$$

Until round t both policies yield the same payoff to the supervisor. The LHS captures the additional payoff from one more round of experimentation in $t+1$. The RHS captures increase in the payoff from t' additional rounds of experimentation with the agent implementing a low potential idea in round $t+t'+1$. A sufficient condition for the above to be satisfied is $q > (q+(1-q)k)^2$, which we know is satisfied from Assumption (A). Lemma 1.8 follows from the above discussion.

Lemma 1.8. *The immediately honest policy is pareto superior to the equilibrium policy.*

Thus, both the supervisor and the agent stand to gain if the supervisor commits to honesty. However, as we show below, the supervisor can do better than immediate honesty.

Delayed honesty

The supervisor's preferred policy is driven by the desire to make the agent experiment more when she has low potential ideas but implement immediately if she

gets a high potential idea. Thus, while on the one hand he wants to be honest with the agent, he also wants the agent to experiment as often as possible. We show how the supervisor can fulfil these two objectives through a delayed disclosure policy which we call *delayed honesty* and quantify the gain attainable over immediate honesty.[21]

A policy is a combination of a disclosure time and what to recommend at that disclosure time. A disclosure timing rule is a mapping from the current belief β_t to a choice of round τ at which the supervisor requires the agent to show her ideas to him (or equivalently the number of rounds the agent is required to experiment). He then makes a comment about each of the τ ideas according to a recommendation policy which is a mapping of $\{\ell, \hbar\}^\tau$ onto itself. A recommendation policy is honest if the supervisor honestly reveals the type of all the ideas that the agent has produced. We restrict attention to honest recommendation policies for the time being and analyse what is the optimal disclosure time τ^*. At the disclosure time τ, the agent and the supervisor update their belief about the ability sequentially according to Bayes' rule. Thus, if the supervisor reveals that any of the ideas are high potential they both update their belief to 1 and otherwise revise their belief downwards by τ times

$$\beta_\tau^\ell = \frac{(1-q)^\tau \beta_1}{1 - q\beta_1 \sum_{t=0}^{\tau-1}(1-q)^t}.$$

Fix a prior $\beta_1 \geq \beta_0^{FI}$ and consider a disclosure policy that requires the agent to experiment at least τ times to receive feedback from the supervisor. We are interested in finding out the *maximum* number of rounds of delay. Let the disclosure policy be such that after the agent discovers all her ideas were of low potential she quits experimentation and implements any one her ideas, i.e. $\beta_{\tau+1}^\ell < \beta_0^{FI}$.[22] We say that such a policy is *implementable* if the agent prefers to experiment τ times and receiving feedback to not experimenting and implementing her outside option idea.[23] This yields the following implementability constraint (IC)

[21] Since we are not focussing on delayed partial disclosure, we will omit any mathematical complexity that comes with it such as that of defining mixed strategies. We will focus on the supervisor using pure strategies.

[22] If there is any implementable delayed policy that leads to a posterior above β_0^{FI}, then the same can be achieved by an immediately honest policy by inducing the same number of rounds of experimentation. We will refer to delayed honesty policy as the one which leads to posteriors below β_0^{FI} so that more number of rounds are induced than in immediately honest policy.

[23] There is no expected benefit to the agent by experimenting less than τ times since given the policy the supervisor does not reveal any information to the agent when this is the case.

$$\overbrace{\frac{1}{2}\beta_1[1-(1-q)^\tau(1-(\beta^\ell_{\tau+1}k)^2)]}^{\text{expected benefit of experimentation}} \geq \overbrace{\underbrace{\frac{(\beta_1 k)^2}{2}}_{\text{opportunity cost}} + \underbrace{\tau c}_{\text{actual cost}}}^{\text{total cost of experimentation}}. \tag{IC}$$

Observe that since the agent is expected to carry out multiple rounds of experimentation without knowing their outcome, she evaluates the possibility of attaining a high potential idea relative to β_1. Conditional on being high-ability, with probability $(1-q)^\tau$ she expects to attain only low potential ideas to implement, and with the remaining probability she expects at least one high potential idea. Therefore, with probability $\beta_1(1-(1-q)^\tau)$ she receives 1/2 and with probability $\beta_1(1-q)^\tau$ she will revise her belief down to $\beta^\ell_{\tau+1}$ after the supervisor honestly reveals all her τ ideas are low potential. At this point, she will implement any one of her low potential ideas to obtain an expected benefit of $\frac{(\beta^\ell_{\tau+1}k)^2}{2}$. Finally, there is no benefit of experimentation if the agent is of low-ability type. This is captured in the LHS of IC condition as the expected benefit of experimentation.

If the agent instead opts for implementing her low potential outside option idea, she expects to receive a payoff of $\frac{(\beta_1 k)^2}{2}$. As illustrated in the RHS, she must forego this expected benefit when she decides to experiment, in addition to paying the cost of experimentation c for τ rounds. The IC condition thus puts a limit on the maximum number of rounds the agent is willing to experiment when she is at a belief β_1 and the supervisor is committed to revealing all the information after those rounds.

We next analyse the supervisor's incentives under such a policy. The supervisor's ex-ante expected payoff from a τ-implementable policy is

$$\beta_1[1-(1-q)^\tau(1-(\beta^\ell_{\tau+1}k)^2)].$$

The supervisor, like the agent, only sees the potential of the ideas once they are presented to him – he evaluates the probability of at least one high potential idea among the τ attempts according to β_1. Does the supervisor benefit from a higher or a lower τ? While on the one hand a higher τ reduces the probability of the agent only producing low potential ideas, but on the other hand it also depresses the effort of the agent in case of such event. The following lemma shows that the first order effect of reduced probability dominates the second order effect of reduced effort so that the supervisor is always better off inducing a higher τ.

Lemma 1.9. *Under assumption (A), the supervisor's payoffs are increasing in the number of rounds the agent experiments τ.*

Proof. Consider the expected probability of success from a τ-implementable policy:

$$\beta_1[1-(1-q)^{\tau}(1-(\beta^{\ell}_{\tau+1}k)^2)] \tag{1.7.C.5}$$

Now consider the expected probability of success from a $\tau+1$-implementable policy:

$$\beta_1[1-(1-q)^{\tau+1}(1-(\beta^{\ell}_{\tau+2}k)^2)] \tag{1.7.C.6}$$

Subtracting equation (1.7.C.5) from (1.7.C.6) and looking at the condition for it being positive, we get

$$q+(1-q)(\beta^{\ell}_{\tau+2}k)^2-(\beta^{\ell}_{\tau+1}k)^2>0$$

This always the case since $q>k$ from Assumption (A), which implies $q>(\beta^{\ell}_{\tau+1}k)^2$. Therefore, the payoff of the supervisor is increasing in the number of rounds of experimentation. □

Supervisor's maximization problem therefore reduces to getting the agent to experiment as many rounds as possible. This is solely determined by the IC condition. It is immediate that the expected benefit of experimentation to the agent under such a policy, although increasing in β_1, is bounded above by 1/2. Consequently, for a higher β_1 the agent should want to experiment more number of rounds but up to a limit. This limit is imposed by the bounded benefits on the one hand, and the increasing cost of experimentation on the other. Our objective is to determine the maximum (β_1,τ) combination that is implementable with such a policy.

For this purpose, fix τ. Now, if there exists a prior belief that makes the IC condition bind, then it must be the *minimum* prior that does so. Define this minimum prior belief by $\bar{\beta}^{\tau}$. So for any belief $\beta_1\geq\bar{\beta}^{\tau}$ the agent finds it optimal to at least experiment τ times. Observe that $\bar{\beta}^{\tau}$ must be increasing in τ since the agent must have a higher belief to induce him to experiment more often by paying a higher cost. Let $\bar{\beta}^{\bar{\tau}}$ be the maximum of this increasing sequence so that $\bar{\tau}$ gives the maximum number of rounds that are implementable and $\bar{\beta}^{\bar{\tau}}$ is the minimum prior that can induce those many rounds. Proposition 1.9 follows from the above discussion.

Proposition 1.9. *The maximum number of rounds τ^* the supervisor can delay honestly*

revealing the outcomes and therefore induce experimentation at prior β_1 is given by

$$\bar{\beta}^{\tau^*} \leq \beta_1 < \bar{\beta}^{\tau^*+1} \text{ if } \beta_1 \leq \bar{\beta}^{\bar{\tau}},$$

and is equal to $\bar{\tau}$ if $\beta_1 > \bar{\beta}^{\bar{\tau}}$.

We end this section with the following observation.

Observation 1.2. *The supervisor weakly prefers a policy of delayed honesty to immediate honesty when delayed honesty is implementable, i.e. when $\beta_1 \leq \bar{\beta}^{\bar{\tau}}$.*

Ali (2017) derives the same result when determining the optimal dynamic disclosure policy in a slightly different environment. In his setting, the agent needs two consecutive successes in order to be successful in the project. The experiments yield success with a positive probability only if the project is of a good type. Ali shows that the more informed party always has an incentive to delay information revelation while the less informed party would prefer early revelation. While we do not solve for the optimal policy here, we showed here delaying may be preferred by the supervisor to immediately revealing the outcome.

For priors above $\bar{\beta}^{\bar{\tau}}$, a combination of immediate honesty and delayed honesty may be preferred by the supervisor. The prospect of finding out the outcome of experimentation immediately after experimenting makes the agent assess future costs probabilistically. Since it might be determined immediately that the last idea had a high potential to succeed, the agent then does not have to bear future costs of experimenting. This reduces the expected cost of experimentation to the agent and makes her willing to experiment. So for higher beliefs, where the agent is not willing to pay a lump sum cost for experimenting with delayed honesty, the supervisor can induce experimentation with immediate honesty. The supervisor can then commit to delayed honesty when the agent reaches a lower belief. However, immediate honesty might provide too much incentive to the agent and the supervisor might do better by committing to a mixed revelation for high beliefs.[24]

[24] We do not consider these policies in this paper as our primary objective is to highlight the tensions when the supervisor does not have commitment power. We merely want to show how the supervisor can do better when there is commitment in the relationship, and what incentives shape a "preferred" policy.

Chapter 2

Diversity Paradox

2.1 Introduction

Information asymmetry exists within hierarchies in organisations. Managers of different ranks possess extra information about employees and their performance. Knowing how this information shapes the promotion and hiring decisions of lower-ranked managers is essential for shareholders. In such an environment, earning a reputation for diversity and non-discrimination becomes very important both for the managers and shareholders.

Given the recent studies on the positive effect of diversity in workplaces and the importance of non-discriminatory behaviour for higher productivity in firms, a large number of firms are motivated to promote diversity and counter discrimination. Consider a firm that wants to improve diversity. The firm will benefit from promoting diversity and therefore is willing to hire less talented members of the minority along with talented ones (Coate and Loury (1993)). Now consider a manager in this firm who dislikes employing minorities. The manager's strategies are contrary to the policy and profit of the firm, and if the firm finds out the real type of the manager, it will fire him. As a result, the manager faces the following trade-off: while hiring from minorities will reduce his utility, not hiring them might cost him his career. This paper aims to look at the manager's problem. We explore how the career concern of a manager with a bias against minorities, will shape the employment and performance of minorities in the long and the short run.

We construct a finitely repeated principal manager career concern model with managers who have either positive or negative bias toward minorities. The manager in each period has to make a hiring decision in which only the skin colour (or any minority groups' affiliation) is observable. We assume the applicant's ability of the applicants to

be the manager's private information. Employees are then required to work on a project. The success of the project depends on the ability of the employee and h the help of the manager. The manager can help the employee by putting a costless effort into the project. The principal only observes the outcome of the project and the skin colour of the employee (group affiliation) and decides whether to keep or fire the manager.

Central to this analysis is our assumption about the monitoring structure. In our model, the manager has both a direct and indirect role in the success of a project. While his indirect role is through the choice of the employee, his direct role is through his effort. As a result, for the principal, the outcome of the project is a two-dimensional signal of the manager's type. It is this multidimensionality of the signal that forms a unique monitoring process in the game. This initiative introduces sabotage into career concern models and plays a central role in the reputation building process.

There are three key results; First, for a discriminator with a high degree of aversion to minorities, higher employment of black workers in the initial phase, is followed by sabotage. Using this strategy, the discriminator can build a reputation and cash it in, at later stages of the game. Intuitively, if the principal has a positive bias toward diversity, a manager with the same positive bias will induce higher payoff for the principal. Such a benevolent manager(hereafter, 'benevolent') will hire more black employees relative to the discriminator. Since the benevolent and the principal both gain from hiring blacks, they are keener to hire lower ability black applicants than a discriminator. The implication will be that the blacks hired by a benevolent are more likely to fail than blacks hired by a discriminatory manager (hereafter, 'discriminator'). If the discriminator wants to build a reputation he will employ more blacks, but by occasionally withholding assistance, he induces their failure. The failure will make him more likely to be perceived as a benevolent type. If the game is repeated once, then the discriminator does not require considerable improvements to his reputation, so sabotage becomes too costly. But when the game is repeated for two periods or more, then it is optimal for the discriminator to incur some loss in the initial stage to build a reputation. He will cash in this reputation at later stages.

Second, we find a âĂIJDiversity ParadoxâĂİ. We show that the discriminator is only able to sabotage the black employee if the principal has strong preferences toward diversity. When the principal has no or little bias toward diversity, in the equilibrium, the discriminator is unable to sabotage the employee. This forms the diversity paradox: if there is no positive bias toward diversity, diversity does not improve much. But if

there is, the diversity improves at the cost of increased sabotage. The main intuition behind this result is that the benevolent always hires both more high ability and low ability blacks. As the bias increases the possibility of lower ability blacks relative to higher ability blacks being hired increases and black failures becomes more probable with benevolent managers. As a result, the positive bias towards diversity is the main driver of sabotage.

Finally, we show that when the value of the project is not very high, even a slight aversion to black workers is enough to induce sabotage. When the value or productivity of the project is high, it's more costly to sabotage. Therefore, only discriminators with a high degree of aversion towards black workers would find sabotage optimal.

Many studies support our result. A good example is The Female FTSE Board Report. The report has been monitoring the number of women holding the position of executive director on the corporate boards of the UK's top 100 companies since 1999. In the most recent report, the percentage of female representation on the corporate boards was close to the target set, but the report identifies their representation as a "tick box" attitude. The report shows that on average women are less likely to be promoted than their male counterparts, and their average tenure is half that of men. The improved numbers do not reflect an underlying improvement in female status on boards: in the context of our paper, its mostly to gain reputation. The shorter tenure of women relative to men and low promotion rate confirms the sabotage narrative of our paper.

Finally, The US Equal Opportunity Commission report confirms our final result. The report shows lower productivity jobs have more reports of harassments. If we assume harassment to be a weak measure of sabotage, this is in line with our result. Low productivity jobs are more prone to sabotage because even slight aversion towards minorities will make sabotage optimal.

2.2 Literature Review

This paper relates to four strands of literature: finitely repeated reputation games with imperfect monitoring, dynamic persuasion games, sabotage and discrimination.

This work relates most closely to the literature on discrimination. The literature is divided into three categories: taste-based discrimination, statistical discrimination and invisibility hypothesis. Central to taste based discrimination starting with Becker (1971) is the assumption that some employers dislike members of minority groups.[1].

[1]For an excellent survey of discrimination literature refer to Lang and Lehmann (2012)

The next category, statistical discrimination, starting from Phelps (1972) and Arrow (1973), focuses on imperfect information about worker's training and productivity. The leading cause of discrimination in this category is the belief that on average members of minority groups perform worse than other workers. Coate and Loury (1993) in their work, assume both statistical and taste-based discrimination. They show that quota policies like affirmative action, if implemented in one period, might negatively affect the black workers' skill acquisition and cause patronisation.

The last category stems from the invisibility hypothesis by Milgrom and Oster (1987). They suggest that the main reason for discrimination is that members of minority groups are less observable by the employers. In most of the literature, discrimination is modelled at the market level and how discriminatory attitudes might affect the employment patterns and wage differentials between Black and White workers. This work falls in the first category through how it defines discrimination. Since we show that in the presence of taste-based discrimination, the positive bias of the principal may induce sabotage of the black worker, in terms of implication, our paper is closest to Coate and Loury (1993). However, our primary focus is on how the presence of discriminatory attitudes toward some workers (based on race or gender) can affect the relations within an organisation and hierarchy. In this sense, the most closely related works to ours are Shin (2016) and Kamphorst and Swank (2016). Kamphorst and Swank (2016) look at an organisation where there is an expectation of discrimination. They show that in the presence of such expectation even if the principal has no bias toward minorities, he will discriminate against them to avoid demotivating the white worker. Shin (2016) models an agency problem between the owner of a firm and the managers. The model focuses on the information asymmetry between the manager and the owner. In her model, the managers might have a negative bias toward black workers. While the type of manager and the productivity of the workers is the manager's private information, he makes a promotion decision. Shin (2016) characterises the optimal mechanism to induce the manager to promote the minority worker. She shows that the optimal mechanism is for the manager to report all information to the owner, and for the owner to make promotion decisions. While the environment of this work is very close to Shin (2016), it differs on its key premises. Firstly, we consider a repeated game wherein the principal is never able to observe the ability of the subordinate. In our setting the main deriving force is the career concern of the biased manager. Secondly, in her model, the manager plays no role in the success or failure of the workers. In our

framework, the manager can affect the outcome of the task assigned to the worker and can build a reputation by sometime sabotaging the black worker. To the best of our knowledge the role of manager's reputation in shaping the effect of discrimination has not been studied before.

Bénabou, Falk and Tirole (2019) and Bénabou and Tirole (2011) study the implication of reputational concerns in the provision of social goods. More specifically, their work differs from ours in the key findings. They characterise the "Moral Licensing", where the discriminatory type prefers to initially perform some non-discriminatory tasks in order to gain reputation and discriminate in later stages. The key finding of the current paper is in contrast to moral licensing. We find that in order to gain reputation, the manager might hire more from minority groups but improve his reputation by sabotaging them.

The main reason that sabotage plays such a role in reputation building is the uninformative monitoring of the principal when faced with the failure of a project. There is a vast literature in repeated reputation games that focuses on monitoring. The seminal works by Diamond (1991), Fudenberg and Levine (1992), Cripps, Mailath and Samuelson (2004) and Gossner (2011) establish the links between monitoring and reputation formation and how informativeness of the monitoring systems can shape the equilibrium of these games. Ely and Välimäki (2003) and Ely, Fudenberg and Levine (2008) look at bad reputations and how uninformative monitoring induces good types to use the bad type's strategy to build a reputation. In the contractual environment, the seminal career concern work by Holmstorm (1999) focuses on imperfect monitoring and compensation schemes. Halac and Prat (2016) look at two-sided learning. In the non-contractual environment more recently Bar-Issac and Deb (2018) Deb and Ishii (2018) look at uncertainty in monitoring. Bar-Issac and Deb (2018) look at a setting where monitoring is infrequent. They construct a monitoring mechanism in which infrequent monitoring can improve the incentives for the agent to work. Deb and Ishii (2018) consider a setting in which the type of the agent and the monitoring mechanism are uncertain. They build a setting with a new dynamic commitment type and use the consumer's uncertainty about the state of the world (the type of the firm and the monitoring structure) to show how reputation incentives shape the equilibrium. They show that with uncertain monitoring but without the specified type, the Stackelberg payoff cannot be obtained. However, once they assume the dynamic commitment type, they show that Stackelberg payoff is achievable. What makes this work different from

most of these works is the special structure of monitoring in our setting. In our setting, monitoring needs to be two dimensional, because both the employment choice and the effort choice need to be monitored by the principal. Given the fact that only the outcome of the project is publicly observable, the monitoring is imperfect on one dimension and in case of failure completely uninformative on the other dimension. It is this structure of the monitoring that makes this framework novel and opens the ground to introducing sabotage in reputation games without competition. To the best of our knowledge, a structure of monitoring like this has not been considered in the reputation literature.

Another literature related to our work is sabotage. Sabotage appears in a variety of topics in economics. Most extensively sabotage is modelled in tournaments, teams and contests literature.Lazear and Rosen (1981) in their seminal work discuss sabotage in tournaments. They use relative performance evaluation when production is interrelated between co-workers. They show that under such schemes, agents are more interested in reducing the probability of success of their competitors (sabotage) rather than improving their performance. Konrad (2000) models lobbying in the form of contests. He shows that a lobbyist improves their chances of success by using their resources to reduce the effectiveness of the competing lobbyists (sabotage) rather than improving the effectiveness of their lobbying. Auriol and Guido (2000) look at sabotage in teams and show that if contract renegotiation is possible, the agents become less likely to help each other and may engage in sabotage. In their model, sabotage does not arise because of relative performance schemes. It instead comes from the possibility of renegotiating contracts and the incentive for the agents to build a reputation for high productivity. Chalioti (2019) looks at a framework where workers' ability is unknown. She shows that in the presence of a contract renegotiation option, the agent engages in sabotage to bias the learning process of ability in her favour. Among these works, Auriol and Guido (2000) and Chalioti (2019) are closest to our work, as in both, the primary driver of sabotage is reputation concern. Nonetheless, in our work, there is no competition or relative evaluation. The manager sabotages the employee (inflicts failure on himself and the employee) to prove he is a benevolent type. While in all sabotage literature, it is the presence of some form of relative evaluation of reputation and performance that induces competition and inflicts sabotage.

In this context, our work also relates to dynamic persuasion games. Most of this literature focuses on communication games between an informed but potentially

biased agent and an uninformed decision maker. Bénabou and Laroque (1992) build a repeated communication game between a sender and receiver. They show that when the information is noisy, the sender can engage in repeated manipulation of information without being detected. Morris (2001b) uses this setting in a two-period repeated game between a potentially biased expert and an unbiased decision-maker. He shows that in suggesting the optimal policy, the unbiased expert may lie in the first period for reputational reasons. Gentzkow and Shapiro (2006) build a model of media bias wherein the news outlet tends to be perceived as an accurate provider of information. In their setting, the outlet's past reports build a reputation for accuracy. They show that absent ex-post verification sources, in order to build a reputation, the news outlets distort their reports to conform with their prior belief. Most of these works are related to our setting; especially when to build a reputation, one has to choose the non-optimal action. However, the main point of departure from this literature is that our framework goes beyond communication and inflicts costly actions. The communication games are not able to model the situation described in our framework precisely because the sender only engages in cheap talk and not signalling.

The rest of the paper is structured as follows. In section 2.3, we present the model. In section 2.4, we analyse step by step the equilibrium of a three-period game and illustrate the sabotage equilibrium. The conclusion is in section 1.6.

2.3 Model

2.3.1 Environment

We consider a repeated interaction between a principal (she) p and a manager (he) m, where the manager has a hiring and performance responsibility which affects his future job prospect. Both the manager and the principal have some level of sensitivity toward diversity. At each round, the manager decides to employ or promote an employee from a pool of 2 applicants. The employee then works on a success-failure project. The principal, having observed the choice of the manager and the outcome of the project, decides to keep or fire the manager.

The pool of two applicants is diverse $m \in \{0, 1\}$. That is one applicant belongs to minority groups (women, people of colour etc,) $m_t = 1$ and one does not $m_t = 0$. Each applicant has an ability $a_m \sim U[0, 1]$.

The principle and the manager have both some degree of sensitivity toward diversity. The principal has sensitivity $\beta \in [0, 1]$ toward diversity. That is he gets

an extra utility of β if he hires from the minority group (black worker from now on). The manager has two types: benevolent and discriminator, $\theta \in \{\beta, -\delta\}$ respectively. More specifically either his sensitivity is identical to the principal that is $\theta = \beta$ or its misaligned with her, that is $\theta = -\delta$ with $\delta \in [0, 1]$. The manager's sensitivity type is his private information. The principal holds a public prior belief on the manager's type $\pi_0 = pr(\theta = \beta)$ and updates his belief according to Bay's rule.

Each period the manager has to choose an employee m_t from the pool of applicant $M_t = \{a_{m=1}, a_{m=0}\}$ applying for the position. Prior to his choice, the manager privately observes the ability of both applicants. He hires one according to his sensitivity toward diversity, and the applicants' ability. Once the manager makes his hiring decision, he chooses a costless effort level $e_t \in \{0, 1\}$ to exert on the employee/project, which will improve the chances of success. The principal, on the other hand, never observes the ability of the applicant and the effort choice of the manager; she only observes the manager's hiring decision and the outcome of the project.

As mentioned, the probability of success of the project depends on the ability of the employee and the effort choice of the manager. More formally:

$$X_t = \begin{cases} 1 & \text{with probability } e_\theta \sqrt{a_m}, \\ 0 & \text{with probability } (1 - e_\theta \sqrt{a_m}) \end{cases}$$

wherein X_t is the pay off of the project in case of success and failure respectively. Per period payoff of the principal is

$$U_t^P = E(X_t) + m_t \beta \tag{2.1}$$

The manager at each period receives

$$U_t^\theta = \nu\big(E(X_t) + m_t \theta\big) \tag{2.2}$$

wherein and $\nu \in (0, 1]$, is the fraction of output that the manager obtains.

At each round t, the manager chooses the applicant m_t and the effort level e_t that gives him highest present value of all future pay-offs.

$$\mathcal{V}_t^\theta = \max_{m_t, e_t} \sum_{s=t}^{3} \mathbb{E}(U_s^\theta) \tag{2.3}$$

At the end of each round t after observing the outcome of the project X_t and the hiring decision of the manager m_t, the principal decides to keep or fire the manager $f \in \{0, 1\}$. If she keeps the manager, $f = 0$ she gets sum of the present value of all future pay-offs.

$$\mathcal{V}^P_{f=0} = \sum_{s=t+1}^{3} \mathbb{E}(U^p_s) \tag{2.4}$$

If she fires the manager $f = 1$ then the principal gets an outside option of

$$\mathcal{V}^P_{f=1} = \sum_{s=t+1}^{3} C \tag{2.5}$$

So the principal at the end of each round make the choice that gives him the highest present value of all expected future pay-offs:

$$\mathcal{V}^p_t = \max_{f_t \in \{0,1\}} \sum_{s=t+1}^{3} \mathbb{E}(U^P_s) \tag{2.6}$$

If the manager is fired, he gets an outside option of $\mathcal{V}^\theta_f = -D$. Since the manager is fired based on the belief that he is a discriminator, D is assumed to be very large.

We assume no firing at the prior. That is π_0 is always larger or equal to the minimum belief needed to progress to the next stage. We can justify this assumption as there is always at least a chance to hire a new manager with the same initial prior.

Finally we assume that $\theta = \beta$ is a non-strategic benevolent manager type, who always chooses the action that the principal prefers.

2.3.2 Timing

The timing of the game is as follows:

1. At the start of the game, nature chooses the manager's type, and the manager privately observes it.

2. The pool of applicant with their ability is realised. The manager privately observes the ability of each type and makes the hiring decision. The applicants and the employee's type remains private information of the manager throughout the game.

3. Manager after hiring the employee chooses his costless effort e_t

4. Project outcome is realised and the principal observes both the applicant hired $m_t \in \{0, 1\}$ and the project outcome $X_t \in \{0, 1\}$ and updates his beliefs given the

observables, $\pi_t = pr(\theta = \beta | m_t, x_t)$

5. Principal decides to keep or fire the manager. The manager receives$\mathcal{V}_f^\theta = -D$ if he gets fired and the principal gets her outside option of $\mathcal{V}_{f=1}^P$

6. The game finishes if the manager is fired and repeats if the manager is kept.

2.4 Reputation building and Sabotage

2.4.1 Preliminaries

The repeated game between the principal and the manager is one of the finitely repeated reputation games.[2] The δ type manager strategically chooses the employee and effort to avoid being fired by the principal. The solution concept is the manager preferred (perfect) Bayesian Equilibrium.

To define the strategy of players first, we need to define the history for each player when they have to make a decision. The principal starts each period t, with a belief π_{t-1}, which is formed after having observed the manager's choice of employee and the success or failure of the project in the previous period; namely $\{m_{t-1}, X_{t-1}\}$. More specifically a realised history for the principal is the set of all previous employment choices of the manager, the realise outcome of the past projects (including last period's m_{t-1} and x_{t-1}) and the sequence of his past decisions of keeping the managerf_{t-1}. It is apparent that period t will only be reached if $\{f_s\}_{s=0}^{t-1}$=$\{0\}_{t=0}^{t-1}$. For the manager, on the other hand, the realised history includes in addition to the public history observed by the principal, the set of all past realised pool of applicant's ability $\{a_{m_s=0}, a_{m_s=1}\}_{s=0}^{t-1}$ and the history of his past effort choices, including last period e_{t-1}.

For most of the game, we focus on mixed strategy equilibria. Since one type of manager $\theta = \beta$ has no career concerns, his optimisation decision is per period. Therefore a pure strategy equilibria could only be specified in the very extreme case where $\beta = 1$. In all other cases, a pure strategy by the principal would break down in the equilibrium.[3] A strategy for the δ manager,in round t is a mapping from last observed pool of applicant and belief of the principal about his type to a possible mixed decision in employment choice m_t and e_t Furthermore, a mixed strategy for the principal q_t in period t is a mapping from the last observed outcome $\{m_{t-1}, X_{t-1}\}$ and belief of

[2] I acknowledge that some of the proofs in this section are incomplete and need further work.

[3] To be more explicit suppose the principal contingent on observing an event i.e m_t and or x_t always sets $f_t = 0$. Then δ type manager will choose m_t and e_t to avoid reaching that event. In the equilibrium, the realisation would result in firing the β type manager, which is not optimal for the principal.

Only for certain specifications of D and β a pure equilibrium of always firing if $m_t = 0$ and $x_t = 0$ can exist. However, this is not general enough for the analysis.

π_{t-1} to a possible mixed decision of firing the manager. Let

$$q_t^{m_t,\ X_t} = pr(f_t = 1 \mid m_t,\ X_t,\ \pi_{t-1})$$

be the probability of firing following observed past history, and current employment choice and realised output.

Let the $\mathbf{q}^*_{\pi_t}$ denote the conjectured strategy of the principal, and let $m^*_{\pi_t}$ and $e^*_{pi_t}t$ be the conjectured strategy of the δ type manager. Given the conjectured strategy of the manager, the principal updates belief about the type of the manager. It is worth mentioning that, the public history at the beginning of period t can be summarise by the current belief of the principal about the type of the manager π_t.

Having all this in hand we can now describe the notion of the equilibrium in this repeated game of reputation. The conjuncture strategies, $\mathbf{q}^*_{\pi_t}$, $m^*_{\pi_t}$ and $e^*_{pi_t}t$ can be established as equilibrium if given the belief about the type of manager at period t, π_t, the strategies are best response to one another and the belief π_t, is consistent with what the player's conjectured.

Upon observing the realised outcome and the employment choice of the manager the principal updates his belief about the type of the manager. First we define the following probabilities

$$pr(S|\theta=\delta) = \begin{cases} \gamma_t^{m=1} = pr(s \mid m_t = 1,\ e_t,\ \theta = \delta) = E(\sqrt{a_{m=1}} \mid m_t = 1,\ e_t,\ \theta = \delta) & \text{if } m_t = 1, \\ \gamma_t^{m=0} = pr(s \mid m_t = 0,\ e_t,\ \theta = \delta) = E(\sqrt{a_0} \mid m_t = 0,\ e_t,\ \theta = \delta) & \text{if } m_t = 0 \end{cases}$$

Since the β type manager is non strategic these probabilities for him, would change to

$$pr(S|\theta=\beta) = \begin{cases} \lambda_t^{m=1} = pr(s \mid m_t = 1,\ \theta = \beta) = E(\sqrt{a_{m=1}} \mid m_t = 1,\ \theta = \beta) & \text{if } m_t = 1, \\ \lambda_t^{m=0} = pr(s \mid m_t = 0,\ \theta = \beta) = E(\sqrt{a_{m=0}} \mid m_t = 0,\ \theta = \beta) & \text{if } m_t = 0 \end{cases}$$

Now we can define the updated belief of the principal upon observing $X_t = 1$ and m_t

$$\pi_t^{m_t} = \frac{\lambda_t^{m_t} pr(m_t|\theta=\beta)\pi_{t-1}}{\lambda_t^{m_t} pr(m_t|\theta=\beta)\pi_{t-1} + \gamma_t^{m_t} pr(m_t|\theta=\delta)(1-\pi_{t-1})}$$

and the updated belief upon observing $X_t = 0$ and m_t

$$\pi_t^{m_t} = \frac{(1-\lambda_t^{m_t}) pr(m_t|\theta=\beta)\pi_{t-1}}{(1-\lambda_t^{m_t}) pr(m_t|\theta=\beta)\pi_{t-1} + (1-\gamma_t^{m_t}) pr(m_t|\theta=\delta)(1-\pi_{t-1})}$$

Having defined the belief updating of the principal we can now move to analysing the

three-period game. We start with identifying the solution to the last period of the game.

2.4.2 Reputation building - three period game

We start with a three-period reputation game between the principal and the manager. This preliminary analysis helps in identifying sabotage equilibrium in more than three-period games later on. We will show why the two-period model falls short of capturing the sabotage equilibrium. The main intuition is that in a two-period game as reputation building is only needed to reach the final period, higher than the minimum reputation is redundant. While in a three or more period games reputation building can lead to two or more periods of consecutive maximal discrimination by the discriminator.

Period Three-last period

Starting from the final period, it is straightforward to see that in this period since the game finishes and there is no credible threat of firing by the principal, the unique strategy of the $\theta = \delta$ type manager (the manager henceforth) is to maximise the last period pay off with no reputation (career concern) consideration.

We can therefore define the probability of principal observing a success from each manager type in the following way

$$\gamma_3^{m_3} = \begin{cases} E(\sqrt{a_{m=1}} \mid \sqrt{a_{m=0}} < \sqrt{a_{m=1}}) - \delta) & \text{if } m_3 = 1, e_3 = 1, \\ E(\sqrt{a_{m=0}} \mid \sqrt{a_{m=0}} \geq \sqrt{a_{m=1}}) - \delta) & \text{if } m_3 = 0, e_3 = 1 \end{cases}$$

For the β type manager the probabilities would the be

$$\lambda_3^{m_t} = \begin{cases} E(\sqrt{a_{m=1}} \mid \sqrt{a_{m=0}} < \sqrt{a_{m=1}}) + \beta) & \text{if } m_3 = 1, e_3 = 1, \\ E(\sqrt{a_{m=0}} \mid \sqrt{a_{m=0}} \geq \sqrt{a_{m=1}}) + \beta) & \text{if } m_3 = 0, e_3 = 1 \end{cases}$$

With this the equilibrium can be defined in next proposition

Proposition 2.1. *In the final period of the game, each type of the manager chooses the employee that maximises his last period pay off.*

It is the dominant strategy for both types to set $e_3 = 1$

Each type of manager obtains their maximum payoff and the principal obtains

$$\mathcal{V}_3^P = \mathbb{E}(u_2^P) = \pi_1[pr(\sqrt{a_{m=0}} \geq \sqrt{a_{m=1}} + \beta)\lambda_2^{m=0} + pr(\sqrt{a_{m=0}} < \sqrt{a_{m=1}} + \beta)(\lambda_2^{m=1} + \beta)]$$
$$+(1-\pi_1)[pr(\sqrt{a_{m=0}} \geq \sqrt{a_{m=1}} - \delta)\gamma_2^{m=0} + pr(\sqrt{a_{m=0}} < \sqrt{a_{m=1}}) - \delta)(\gamma_2^{m=1} + \beta)]$$

Proof. Since in the last period threat of firing the manager will not be credible, there will be no career concern consideration for the manager, the equilibrium strategy of the manager is always to choose the applicant and the effort level that maximises his expected pay-off with no reputation concern.

Therefore the manager hiring strategy in the last period is:

$$m_3^{\delta} = \begin{cases} 1 & \text{if } \sqrt{a_{m=0}} < \sqrt{a_{m=1}} - \delta, \\ 0 & \text{if } \sqrt{a_{m=0}} \geq \sqrt{a_{m=1}} - \delta \end{cases}$$

while if the manager was of the benevolent type $\theta = \beta$

$$m_3^{\beta} = \begin{cases} 1 & \text{if } \sqrt{a_{m=0}} < \sqrt{a_{m=1}} + \beta, \\ 0 & \text{if } \sqrt{a_{m=0}} \geq \sqrt{a_{m=1}} + \beta \end{cases}$$

Given the fact that there is no reputation concern in the last period, it is evident that setting $e_2^{\theta} = 1$, is the dominant strategy for both manager types. □

Given the equilibrium strategy of both types in the last period of the game, we can define the belief monotonicity condition.

Lemma 2.1. ***Monotonicity Condition:*** *For all biases of the manager δ, when the manager behaves as if he is in the last period of the game (no career concern strategy), the belief updates is always*

1. *The largest when $m = 1$ and the employee fails $X = 0$. That is $\pi^s_{m=1} < \pi^f_{m=1}$*
2. *The lowest when $m = 0$ and the employee fails $X = 0$. That is $\pi^s_{m=0} > \pi^f_{m=0}$*

In order for belief monotonicity condition to hold it must be that the success to failure ratio of $m = 1$ is lower when the manager is of β type thane when he is δ type, that is the condition in point 1 of lemma 2.1 holds if:

$$\frac{\lambda_3^{m_t=1}}{1-\lambda_3^{m_t=1}} < \frac{\gamma_3^{m_3=1}}{1-\gamma_3^{m_3=1}}$$

Figure 2.1 [4] plots the success to failure ratio for both manager types and shows that the ratio with the last period optimal strategy is always higher for the δ type manager when

[4]The graphs are not discontinues, they converge with a sharp slop toward one

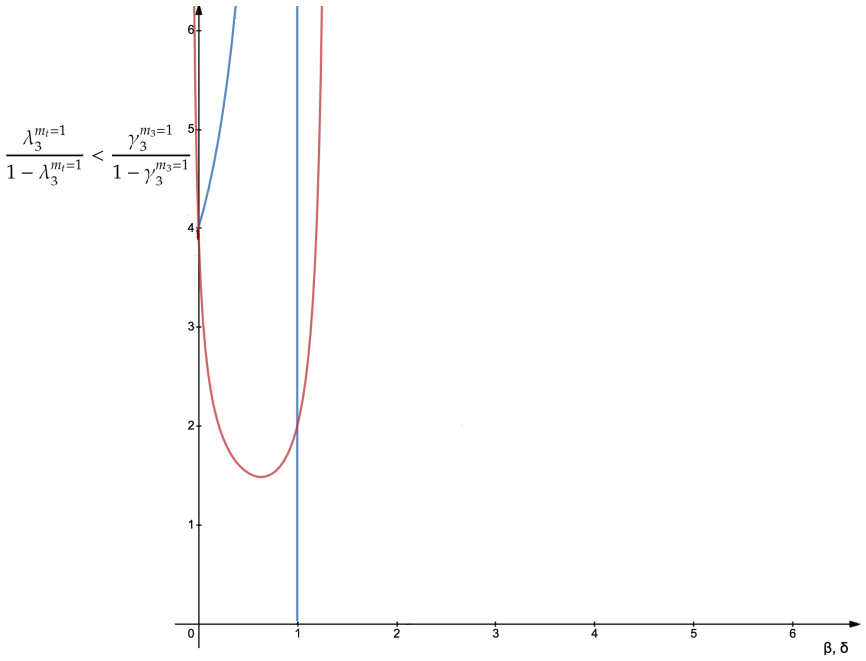


Figure 2.1: Belief Monotonicity-$m_t = 1$

$m_t = 1$. For point two of the lemma 2.1 to hod it must be that the ratio is reversed for $m = 0$

$$\frac{\lambda_3^{m_t=0}}{1-\lambda_3^{m_t=0}} > \frac{\gamma_3^{m_3=0}}{1-\gamma_3^{m_3=0}}$$

Figure 2.2 plots the success to failure ratio for both manager types and shows that the ratio with the last period optimal strategy is always lower for the δ type manager when $m_t = 0$.

Lemma 2.1 shows that since the benevolent prefers employing black workers, he is more likely to hire a lower ability black employee. As a result, if the manager acts without career concern, then the principal believes that the event where a black worker fails is least likely to come from a discriminator. Similarly, for the discriminator, since he dislikes black employees, he is more likely to hire lower ability white employee.Therefore for the principal failure of a white employee is more indicative of the discriminator.

Lemma 2.2. *If the manager makes the employment and effort choice without career concern, since the benevolent manager is more likely to choose the black applicant, beliefs of the principle (for success and failure) is increasing in $m_1 = 1$ and decreasing in $m_1 = 0$*

Lemma 2.2 specifies the updating direction when the manager is behaving without career concern. In this case, a choice of $m = 0$ moves the beliefs of the principal

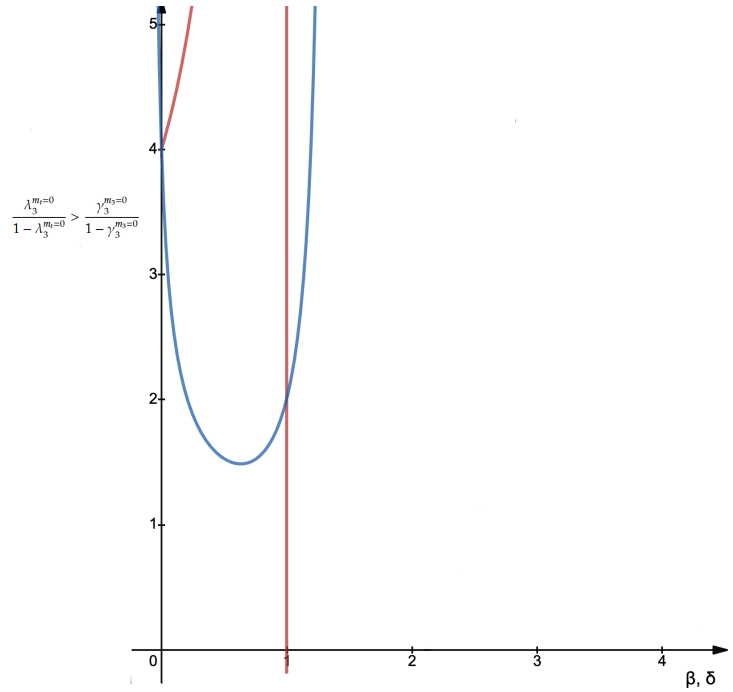


Figure 2.2: Belief Monotonicity$m_t = 0$

away from the benevolent manager. While a choice of $m = 1$ moves the belief of the principal toward the benevolent manager.

We will now proceed to the analysis of one period before the last period and specify the equilibrium in that period.

Period Two

At the end of period two, after observing m_2 and X_2 and given the equilibrium strategy of both types of managers, the principal updates his belief about the manager's type from π_1 to π_2.

Lemma 2.3. *Consider $\underline{\pi}$ as the belief for which $\mathcal{V}_3^P = c$, at the end of the second period, the principal will only keep the manager if $\pi_2 \geq \underline{\pi}$.*

Lemma 2.3, specifies the minimum belief threshold needed for the manager to progress to the last period. Since the threshold depends on the principal's outside option, the minimum belief can be large or small depending on the outside option. Before proceeding to the analysis of the second period, we want to define a belief threshold:

Definition 2.1. *Given the fact that for the manager types defined, if the manager behaves without career concern, uses δ as the hiring threshold and sets $e_t = 1$,the beliefs will always be weakly decreasing in $m = 0$.*

We define π^* as the belief at which if $m_t = 0$ and $X_t = 0$ is observed the

principals belief is updated to $\underline{\pi}$*:*

$$\underline{\pi} = \frac{\pi^*[pr(\sqrt{a_{m=0}} \geq \sqrt{a_{m=1}} + \beta)(1 - \lambda_t^{m=0})}{\pi^*[pr(\sqrt{a_{m=0}} \geq \sqrt{a_{m=1}} + \beta)(1 - \lambda_t^{m=0}) + (1 - \pi^*)[pr(\sqrt{a_{m=0}} \geq \sqrt{a_{m=1}} - \delta)(1 - \gamma_t^{m=0})]} \tag{2.7}$$

Given the strategy of the principal defined in Lemma 2.3, we can now identify the equilibrium strategy of the manager. It is clear that the aim of the manager in period 2, is to reach a belief just above $\underline{\pi}$, further improvements to beliefs is redundant.

Based on the initial assumption of $\pi_0 \geq \underline{\pi}$, we start the analysis with the case where $\underline{\pi} = \pi_1$.

At $\pi_1 = \underline{\pi}$, for the manager to progress to the next round, he needs to improve his reputation or keep it fixed. Therefore he will hire black workers more often. The manager should increase his threshold of hiring $m_1 = 1$ from $-\delta$ toward β till the principal becomes indifferent between keeping or firing him when she observes $m_2 = 0$.

For the principal, the optimal strategy is to mix between firing or keeping the manager when she observes $m_1 = 0$, or more formally when $\pi_2^{m_2=0} = \underline{\pi}$. This makes the manager indifferent between $m_2 = 1$ and $m_2 = 0$ at the optimal threshold. Nonetheless, since the ability is not observable for the principal, the mixing strategy needs to be independent of the ability and only dependent on m_2 and X_2.

Proposition 2.2. *When the prior belief is at its lowest,* $\pi_0 = \underline{\pi}$*, the equilibrium strategy for the principal is mixing strategy. In the equilibrium, she always mixes between firing and keeping the manager, when she observes* $m_2 = 0$*, with the equilibrium probability of firing* $q^* = \frac{\delta+\beta}{U_3^\delta + D}$*. She always keeps the manager if she observes* $m_2 = 1$*.*

The equilibrium strategy of the manager is :

$$m_2 = \begin{cases} 1 & \text{if } \sqrt{a_{m=0}} < \sqrt{a_{m=1}} + \beta, \\ 0 & \text{if } \sqrt{a_{m=0}} \geq \sqrt{a_{m=1}} + \beta \end{cases}$$

The equilibrium effort level of the manager $e_2^{*\theta} = 1$

Proposition 2.2, suggests that in the period before the last period; for a low belief $\underline{\pi} = \pi_0$, the manager in order to progress to the next round, will behave as the benevolent manager behaves. However, the principal still fires him with some positive probability if he chooses $m_2 = 0$.

We now move to a higher range of beliefs and identify the equilibrium strategies for it. Consider the case where $\underline{\pi} < \pi_0 < \pi^*$, if the manager follows the same strategy

of the last period, he will get fired when $m_1 = 0$. As a result, he needs to build a reputation, not to get fired.

We argue that the same strategy of principal in Proposition 2.2, can not be an equilibrium strategy in this range of beliefs by presenting two reasons.

Firstly, since the prior is always higher than $\underline{\pi}$, if in the equilibrium no update occurs, the principal always deviates and keeps the manager. As a result, because fully mimicking of the benevolent will induce no update; this strategy can not be sustained in the equilibrium.

Secondly, any strategy of setting the probability of firing lower, such that the manager uses a lower threshold for hiring, cannot be an equilibrium strategy. The reason is that this strategy induces $\pi^S_{m=0} \neq \pi^F_{m=0}$. Therefore if the principal is indifferent between firing or keeping the manager in one event, she can not be indifferent in the other event and will deviate.

In this case, for the principal, the optimal strategy is to mix between firing and keeping the agent when the manager chooses $m_2 = 0$ and the employee fails $X_2 = 0$. Once again, since the ability is not observable, the probability of the manager getting fired should only depend on m_2 and X_2.

Proposition 2.3. *If the prior belief is not very low, $\underline{\pi} < \pi_0 \leq \pi^*$, then the optimal strategy for the principal is a mixing strategy. She always mixes between firing and keeping the manager, when she observes $m_2 = 0$ and $X_2 = 0$, with the equilibrium probability of firing $q^* = \frac{\kappa}{U_2^\delta + D}$ and to keep the manager in all other cases.*

κ is decreasing in π_0. That is if π_0 is close to π^*, $\kappa \to 0$. But if π_0 is close to $\underline{\pi}$, $\kappa \to U_3^\delta + D$

The equilibrium strategy of the manager is :

$$m_2 = \begin{cases} 1 & \text{if } \sqrt{a_{m=0}} < \frac{\sqrt{a_{m=1}}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa}, \\ 0 & \text{if } \sqrt{a_{m=0}} \geq \frac{\sqrt{a_{m=1}}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa} \end{cases}$$

The equilibrium effort level of the manager $e_2^{*\theta} = 1$

Proposition 2.3, shows for a higher range of beliefs, the manager will progress to the next round if he increases hiring of the black workers. In the equilibrium, the increase will be up to the point where the principal is indifferent between keeping and firing him when she sees a white worker failing. Nonetheless, in the equilibrium, the principal still fires him with some positive probability if he chooses $m_2 = 0$, and the

employee fails $X_2 = 0$.

For beliefs high enough, $\pi_0 > \pi^*$, the manager always behaves as in the last period and always progresses to the next round.

We now proceed to the analysis of the first period of the game where sabotage becomes an optimal strategy for the manager. We show how hiring black workers and not putting effort will help the manager build reputation and minimise diversity to his benefit in the next two periods.

2.4.3 Period one

Moving forward to the analysis of the first period of the game sheds light on the implication of the need to improve reputation. It shows how sabotaging the black worker could help the discriminator build reputation. We define sabotage as exerting no or little effort by the manager in order to make the employee fail in the project. The main argument stems from the implication of Lemma 2.1 and Lemma 2.2. Since the benevolent manager is more likely to hire low ability black workers, failing black workers is indicative of a benevolent manager. A manager who dislikes black applicants can, therefore, hire black workers more often and by sometimes sabotaging them, induce their failure and build a reputation of being benevolent. He then uses this reputation to impose his preferred level of diversity (low diversity) in the future.

We start by arguing that sabotage is not optimal in a two-period game or more concretely in the period before the last. The reason is that the whole aim of the sabotage strategy is to be able to implement a per-period optimal strategy in the periods after sabotage. In the period before the last one, the manager knows that he will be able to implement his optimal strategy in the next period if he progresses. That is because the threat of firing will not be credible in the last period. Therefore, reaching a minimum belief to progress to the next period will be enough for the manager and higher beliefs will not add to his pay off. As a result, sabotage is not an optimal strategy in the period before the last period.

Given the optimality of sabotage, the second-period equilibrium breaks down in the first period. The reason is, given the beliefs of the principal in the equilibrium it is now optimal for the manager to deviate and choose black employees. Then by setting $e_1 = 0$, he can obtain a higher reputation and implement his per period optimal strategy in period 2 and 3. Therefore with the possibility of sabotage, a new equilibrium should emerge in the first period.

To confirm the statement above, we start with a series of Lemmas that specify

the conditions under which sabotage can be an equilibrium strategy in the first period.

The first step is to verify if building a reputation through sabotage increases the present value of future pay-offs.

Lemma 2.4. ***Optimality Condition:*** *For every ν, the manager will only benefit from sabotage, if his bias is large enough, $\delta \geq \delta^*$, that is the improvement in future pay-off from reputation building is so large that it can compensate today's loss, $\mathcal{V}_{sab} > 1 + \mathcal{V}_{mix}$*

Lemma 2.4 shows the condition for the manager to consider sabotage as an optimal path to reputation building. For low biases, since the loss in the second-period pay-off from mixing is not that large, then forgoing first period's pay-off would not be optimal. As the bias of the manager increases, the second-period pay-off shrinks and sabotaging in the first period becomes optimal.

Corollary 2.1. *If the payoff of the project was ν for the principal too. For low productivity project's sabotage will always be optimal*

Corollary 2.1, shows that if the productivity of the projects are low, δ^* will be very low and sabotage would be optimal more often. This follows from Lemma 2.4, since the condition specified there is independent of the payoff of the project for the principal.

Once sabotaging becomes an optimal strategy (high δ), one has to check if sabotaging is possible. That is the belief structure is such that failure of the black workers induces the highest increase in reputation. Lemma 2.1 further specified the belief updating structure.

Recall Lemma 2.1 (monotonicity condition), showed, hiring and sabotaging a black applicant always improves the manager's reputation when the principal believes the manager behaves without career concern.

Next, we argue that for sabotage to be possible $\pi_0 > \underline{\pi}$. For $\pi_0 > \underline{\pi}$, sabotage can not happen. The reason is, to progress with this prior, there should be no negative updates in any of the events; this means that the threshold needs to be β. However, if the threshold is β, there is no improvement in belief with $m_1 = 1$ and $X_1 = 0$, so sabotage becomes redundant and not optimal.

Let us now consider the case where sabotage is optimal and possible. That is when δ is large enough and $\pi_0 > \underline{\pi}$. The manager hires more from $m_1 = 1$ and sometimes does not put in the effort, such that $\pi_1^{f,m=1} = \pi^*$. The principal, on the other hand, believes that the manager sometimes sabotages. Therefore her optimal

strategy is to randomise between keeping and firing the manager if she sees a failing white employee in the second period. Using this strategy, the principal makes the manager indifferent between sabotaging and not sabotaging in the first period. The manager too randomises between sabotaging and not sabotaging. The mixing would be such that the principal keeps the agent in all realisations of m_2 and X_2, but $m_2 = 0$ and $X_2 = 0$. In the case of $m_2 = 0$ and $X_2 = 0$, she would be indifferent between firing or keeping the manager and sometimes fires him.

For this strategy to be an equilibrium strategy, it must not be the case that, given the belief of the principal, the manager has an incentive to deviate and not sabotage in the first period.

Lemma 2.5. ***Sabotage Condition:*** *For sabotage to be an equilibrium strategy, it must be that in the equilibrium only the expected update from failure of black employee, makes low diversity viable in the future periods, that is* $\pi^s_{m=1} < \pi^*$ *and* $\pi^f_{m=1} = \pi^*$.

Lemma 2.5, specifies condition under which deviation from sabotage is not optimal. If the above condition was not in place, given lemma 2.1, the manager always had an incentive to deviate from sabotaging. That is because, if he sets $e_1 = 1$, he will still be able to implement his optimal low diversity level in the future periods.

The final condition for the sabotage equilibrium to exist is the updating condition in the first period given the belief of the principal that the manager randomises between sabotaging and not sabotaging:

Lemma 2.6. ***Threshold Condition:*** *If principal believes that the manager will sabotage with positive probability, at the optimum threshold of hiring:*

1. *It must be the case that the improvement in the belief of the principal is large enough when there is no sabotage and* $m_1 = 1$, $X_1 = 0$, *that is* $\pi^{nsab,f}_{m_1=1} > \pi^{nsab,s}_{m_1=1}$ *and* $\pi^{nsab,f}_{m_1=1} > \pi^*$

2. *The manager should not want to deviate from choosing* $m = 0$, $\pi^f_{m_1=0} \geq \underline{\pi}$

The first-period equilibrium given sabotage requires improvements in hiring of the black workers by the manager. That implies a change in the threshold of choosing m_1. Lemma 2.6 specifies further the condition on the belief updating given the new threshold. Since beliefs in case of $m_1 = 1$ and $X_1 = 0$ decreases with sabotage, it must be that the belief without sabotage is big enough to make mixing an optimal strategy for the manager. On the other hand, change in the threshold of hiring must be such that the manager has no incentive to deviate from setting $m_1 = 0$.

Proposition 2.4. *For $\delta > \delta^{*}*$ and $\beta > \beta^{*} \neq 1$, there exists a sabotage equilibrium in the first period that is preferred by the manager to the mixing equilibrium.*

1. *The principal believes that there is a positive probability of sabotage in the first period and mixes between keeping or firing the manager in the second period if she sees $m_1 = 1$ and $X_1 = 0$ in the first period followed by $m_2 = 0$ and $X_2 = 0$ in the second period, with firing probability of*

$$q_2^{*sab} = \frac{\omega D + 1 + \mathcal{V}_{mix} - (2-\omega)\mathcal{V}_{sab}}{\omega(\mathcal{V}_{sab} + D)} \tag{2.8}$$

   where in $\omega = pr(a_{m=0} \geq (\sqrt{a_{m=1}}) - \delta)^2)(1 - \gamma_2^{m=0})$

2. *The manager believes the principal randomises between firing or keeping him in the second period in case of $m_1 = 1$ and $X_1 = 0$ and $m_2 = 0$ and $X_2 = 0$,and randomises between sabotaging and not sabotaging in the first period with*

$$\eta^{*}_{m_1=1} = \frac{(\beta(\frac{4}{3} - \beta) + \frac{(1-\beta)^4(3+2\beta)}{30})\pi_0(1-\pi*) - [\frac{1}{3}\alpha^2(1-\pi_0)\pi*]}{\frac{2}{3}\alpha^2(1-\pi_0)\pi*} \tag{2.9}$$

   Where α is the ability threshold for $m_1 = 0$ above which the manager always sets $m_1 = 0$

3. *In the first period the manager set*

$$m_1 = \begin{cases} 1 & \text{if } \sqrt{a_{m=0}} < \alpha, \\ 0 & \text{if } \sqrt{a_{m=0}} \geq \alpha \end{cases}$$

   and the principal fires the manager in the event of $m_1 = 0$ and $X_1 = 0$ with probability $q_1 = \frac{\kappa}{U_1^{\delta} + D}$ and $\kappa \geq 0$

4. *Sabotage equilibrium exists only for π_0 not too low and not too high*

5. *$e_1^{*} = 1$ is dominant strategy when $m_1 = 0$*

Proposition 2.4 specifies the sabotage equilibrium, wherein the manager is more likely to hire an applicant from minority groups in the first period. Nonetheless, since he gains more reputation form a failing $m = 1$ employee, he will some time sabotage them. The proposition and the conditions show that when the bias of the manager and the principal is large, but not one, and the principal is not too pessimist or too optimist toward the manager, then an equilibrium with sabotage exists. It improves

the manager's reputation to impose his optimal level of diversity and still retain his job in the future.

The principal knowing that there is some chance of sabotage in the first period, some times fires the manager if $m_2 = 0$ and $X_2 = 0$.

It is essential to keep in mind that the manager will only achieve his optimal level of diversity when $m_1 = 1$ and $X_1 = 0$. In all other cases, the mixing equilibrium specified in the two-period game remains the equilibrium of the game.

Corollary 2.2. ***Diversity Paradox:*** *Sabotage will only occurs if the principal has positive and large enough bias toward diversity*

Corollary 2.2, follows from Proposition 2.4, and shows that when the principal has no bias or very low positive bias toward black workers, sabotage can not happen but improvement in the diversity is also minor. When the principal has a significant positive bias toward minorities, the diversity increases at the cost of sabotage.

Finally in the next proposition we identify the equilibrium of the entire three-period game.

Proposition 2.5. *The characterisation of the three-period game's manager preferred equilibrium is:*

1. *For all $m_1 = 0$ in the first period, the next two-period equilibrium would be exactly as in the two-period game equilibrium of $\underline{\pi} < \pi_1 \leq \pi^*$*

2. *For all $m_1 = 1$ and $X_1 = 1$ in the first period, the next two-period equilibrium would be exactly as in the two-period game equilibrium of $\underline{\pi} < \pi_1 \leq \pi^*$*

3. *For all $m_1 = 1$ and $X_1 = 0$, in the first period, the next two period equilibrium would be similar to equilibrium of $\pi_1 > \pi^*$ with the difference that at the end of the second period the principal some time fires the manager with probability q_2, if $m_2 = 0$ and $X_2 = 0$.*

Proposition 2.5 further specifies the equilibrium of a three-period game of reputation building with sabotage. The manager only obtains his optimal diversity level when the black employee fails. Given the specification of the first-period equilibrium in proposition 2.4, this is a more likely event in the first period. The reason for the increase in the likelihood of $m_1 = 1$ and $X_1 = 0$ is two-fold. Primarily the threshold of choosing $m_1 = 1$ has changed. Secondly, due to the positive probability of sabotage, there are higher chances of $m_1 = 1$ and $X_1 = 0$.

2.5 Conclusion

We constructed a model of sabotage in a career concern environment, when the principal and the manager have some bias toward diversity.

We show that an equilibrium with sabotage exists only when both manager and the principal have large biases toward diversity. This forms the diversity paradox. If the principal has no positive bias toward black workers, diversity is minutely improved. However if the principal has large bias toward diversity then diversity is improved but at the cost of sabotage.

We show that when there is chance of sabotage, the principal randomises between keeping or firing the manager when he sees a white employees fail in the period after sabotage. We also show that for the manager it is only optimal to sometime sabotage the black worker and not all the time.

Finally our setting shows that if the productivity of a project is low then managers with slight negative biases are also induce to sabotage. Therefore sabotage is more likely to happen in low productivity jobs.

However the main focus in this paper is finitely repeated environment and more specifically three period-games. It nonetheless shows a further scope in looking at sabotage in infinitely repeated games and to identify conditions under which sabotage equilibrium would be stable.

2.6 Appendix

A Proofs from main text

In the first section we provide the mathematical derivation of the payoffs and probability of success and failure of the employees in the last period of the game.

Mathematical notation for last period of the game

As mentioned earlier we focus attention on the case were β type manager is non-strategic and bases his choices on per period utility. In other words he has no career concern.

Let us now look at the last period of the game, in this period the game finishes after the realisations of the payoffs. Therefore ex-ante threat of firing will not be credible. The implication is that non of the manager types are career concerned in the last period and they base their choice solely on maximisation of their last period pay off. Given the strategy of each manager type, we can calculate the probability of success. If the manager sets $e_3 = 1$ then:

$$pr(S|\theta=\delta) = \begin{cases} \gamma_3^{m_3=1} = pr(s|\sqrt{a_{m_3=0}} < \sqrt{a_{m_3=1}}) - \delta) = E(\sqrt{a_{m_3=1}}|\sqrt{a_{m_3=0}} < \sqrt{a_{m_3=1}}) - \delta) & \text{if } m_3 = 1, \\ \gamma_3^{m_3=0} = pr(s|\sqrt{a_{m_3=0}} \geq \sqrt{a_{m_3=1}}) - \delta) = E(\sqrt{a_{m_3=0}}|\sqrt{a_{m_3=0}} \geq \sqrt{a_{m_3=1}}) - \delta) & \text{if } m_3 = 0 \end{cases}$$

$$pr(S|\theta=\beta) = \begin{cases} \lambda_3^{m_3=1} = pr(s|\sqrt{a_{m_3=0}} < \sqrt{a_{m_3=1}} + \beta) = E(\sqrt{a_{m_3=1}}|\sqrt{a_{m_3=0}} < \sqrt{a_{m_3=1}} + \beta) & \text{if } m_3 = 1, \\ \lambda_3^{m_3=0} = pr(s|\sqrt{a_{m_3=0}} \geq \sqrt{a_{m_3=1}} + \beta) = E(\sqrt{a_{m_3=0}}|\sqrt{a_{m_3=0}} \geq \sqrt{a_{m_3=1}} + \beta) & \text{if } m_3 = 0 \end{cases}$$

Therefore from the principals point of view with $e_3 = 1$ ex-ante probability of success in each case is equal to the expectation of the square root of ability of the employee while ability has a uniform distribution. Let us simplify notation a bit further and call $a_{m_s=1}$, a_1 and $a_{m_s=0}$, a_0 and start deriving probabilities of success and failure. We start by deriving the probability of success and failure of $\theta = \beta$ manager.

$\lambda_3^{m_3=1}=$

$$E(\sqrt{a_1}|\sqrt{a_0} < \sqrt{a_1} + \beta) = \int \sqrt{a_1} f(a_1|\sqrt{a_0} - \sqrt{a_1} < \beta) da_1$$

$$= \int \sqrt{(a_1)} \frac{f(\sqrt{a_0} - \sqrt{a_1} < \beta|a_1) f(a_1)}{f(\sqrt{a_0} - \sqrt{a_1} < \beta)} da_1$$

$$= \frac{1}{p(\sqrt{a_0} - \sqrt{a_1} < \beta)} \int \sqrt{(a_1)} f(a_0 < (\sqrt{a_1} + \beta)^2) f(a_1) da_1$$

$$= \frac{1}{p(\sqrt{a_0} - \sqrt{a_1} < \beta)} \left(\int_0^{(1-\beta)^2} \sqrt{a_1}(\sqrt{a_1} + \beta)^2 da_1 + \int_{(1-\beta)^2}^{1} \sqrt{(a_1)} da_1 \right)$$

$$= \frac{\frac{2}{3} - \frac{(1-\beta)^4(4+\beta)}{15}}{p(\sqrt{a_0} - \sqrt{a_1} < \beta)}$$

$$p(\sqrt{a_0} - \sqrt{a_1} < \beta) = \int_0^{\beta^2} \int_0^1 da_1 d_0 + \int_{\beta^2}^1 \int_{(\sqrt{a_0}-\beta)^2}^1 da_1 da_0 = \frac{1}{2} + \frac{4}{3}\beta - \beta^2 + \frac{\beta^4}{6}$$

$$\lambda_3^{m_3=1} = \frac{\frac{2}{3} - \frac{(1-\beta)^4(4+\beta)}{15}}{\frac{1}{2} + \frac{4}{3}\beta - \beta^2 + \frac{\beta^4}{6}}$$

and the probability of failure would then be

$$1 - \lambda_3^{m_3=1} = 1 - \frac{\frac{2}{3} - \frac{(1-\beta)^4(4+\beta)}{15}}{p(\sqrt{a_0} - \sqrt{a_1} < \beta)}$$

$$= \frac{\beta(\frac{4}{3} - \beta) + \frac{(1-\beta)^4(3+2\beta)}{30}}{\frac{1}{2} + \frac{4}{3}\beta - \beta^2 + \frac{\beta^4}{6}}$$

similarly one can identify $\lambda_3^{m_3=0}$=

$$E(\sqrt{a_0} | \sqrt{a_0} > \sqrt{a_1} + \beta) = \int \sqrt{a_0} f(a_0 | \sqrt{a_0} - \sqrt{a_1} > \beta) da_0$$

$$= \frac{1}{p(\sqrt{a_0} - \sqrt{a_1} > \beta)} \left(\int_{\beta^2}^{(1} \sqrt{a_0}(\sqrt{a_0} - \beta)^2 da_0 \right.$$

$$= \frac{\frac{2}{5} - \beta + \frac{2}{3}\beta^2 - \frac{\beta^5}{15}}{p(\sqrt{a_0} - \sqrt{a_1} > \beta)}$$

$$p(\sqrt{a_0} - \sqrt{a_1} > \beta) = \int_0^{(1-\beta)^2} \int_{(\sqrt{a_1}+\beta)^2}^1 da_0 da_1 = \frac{1}{2}(1-\beta)^3(1+\beta/3)$$

$$\lambda_3^{m_3=0} = \frac{\frac{2}{5} - \beta + \frac{2}{3}\beta^2 - \frac{\beta^5}{15}}{\frac{1}{2}(1-\beta)^3(1+\beta/3)}$$

and the probability of failure would then be

$$1 - \lambda_3^{m_3=0} = 1 - \frac{\frac{2}{5} - \beta + \frac{2}{3}\beta^2 - \frac{\beta^5}{15}}{p(\sqrt{a_0} - \sqrt{a_1} < \beta)}$$

$$= \frac{\frac{1}{10} - \frac{\beta}{3} + \frac{\beta^2}{3} - \frac{\beta^4}{6} + \frac{\beta^5}{15}}{\frac{1}{2}(1-\beta)^3(1+\beta/3)}$$

We now turn to deriving the probability of success and failure when the manager is of δ type.

We start by deriving $\gamma_3^{m_3=1}=$

$$E(\sqrt{a_1}|\sqrt{a_0} < \sqrt{a_1} - \delta) = \int \sqrt{a_1} f(a_1|\sqrt{a_0} - \sqrt{a_1} < -\delta) da_1$$

$$= \frac{1}{p(\sqrt{a_0} - \sqrt{a_1} < -\delta)} (\int_{\delta^2}^{(1} \sqrt{a_1}(\sqrt{a_1} - \delta)^2 da_1$$

$$= \frac{\frac{2}{5} - \delta + \frac{2}{3}\delta^2 - \frac{\delta^5}{15}}{p(\sqrt{a_0} - \sqrt{a_1} < -\delta)}$$

$$p(\sqrt{a_0} - \sqrt{a_1} < -\delta) = \int_0^{(1-\delta)^2} \int_{(\sqrt{a_0}+\delta)^2}^1 da_1 da_0 = \frac{1}{2}(1-\delta)^3(1+\delta/3)$$

$$\gamma_3^{m_3=1} = \frac{\frac{2}{5} - \delta + \frac{2}{3}\delta^2 - \frac{\delta^5}{15}}{\frac{1}{2}(1-\delta)^3(1+\delta/3)}$$

and the probability of failure would then be

$$1 - \gamma_3^{m_3=1} = 1 - \frac{\frac{2}{5} - \delta + \frac{2}{3}\delta^2 - \frac{\delta^5}{15}}{p(\sqrt{a_0} - \sqrt{a_1} < -\delta)}$$

$$= \frac{\frac{1}{10} - \frac{\delta}{3} + \frac{\delta^2}{3} - \frac{\delta^4}{6} + \frac{\delta^5}{15}}{\frac{1}{2}(1-\delta)^3(1+\delta/3)}$$

similarly one can identify $\gamma_3^{m_3=0}=$

$$E(\sqrt{a_0}|\sqrt{a_0} > \sqrt{a_1} - \delta) = \int \sqrt{a_0} f(a_0|\sqrt{a_0} - \sqrt{a_1} > -\delta) da_0$$

$$= \frac{1}{p(\sqrt{a_0} - \sqrt{a_1} > -\delta)} (\int_0^{(1-\delta)^2} \sqrt{a_0}(\sqrt{a_0} + \delta)^2 da_0 + \int_{(1-\delta)^2}^1 \sqrt{a_0} da_0)$$

$$= \frac{\frac{2}{3} - \frac{(1-\delta)^4(4+\delta)}{15}}{p(\sqrt{a_0} - \sqrt{a_1} > -\delta)}$$

$$p(\sqrt{a_0} - \sqrt{a_1} > -\delta) = \int_0^{\delta^2} \int_0^1 da_0 d_1 + \int_{\delta^2}^1 \int_{(\sqrt{a_1}-\delta)^2}^1 da_0 da_1 = \frac{1}{2} + \frac{4}{3}\delta - \delta^2 + \frac{\delta^4}{6}$$

$$\gamma_3^{m_3=1} = \frac{\frac{2}{3} - \frac{(1-\delta)^4(4+\delta)}{15}}{\frac{1}{2} + \frac{4}{3}\delta - \delta^2 + \frac{\delta^4}{6}}$$

and the probability of failure would then be

$$1 - \gamma_3^{m_3=0} = 1 - \frac{\frac{2}{3} - \frac{(1-\delta)^4(4+\delta)}{15}}{p(\sqrt{a_0} - \sqrt{a_1} > -\delta)}$$

$$= \frac{\delta(\frac{4}{3} - \delta) + \frac{(1-\delta)^4(3+2\delta)}{30}}{\frac{1}{2} + \frac{4}{3}\delta - \delta^2 + \frac{\delta^4}{6}}$$

Having derived the probabilities we can now turn to proving the Lemma's and propositions in the text.

Proof of Lemma 2.2

Proof. Given Lemma 2.1, the only thing needed to prove the argument is to show that $\pi^s_{m_3=0} < \pi_0$:

$$\pi^s_{m_3=0} = \frac{\pi_0[pr(\sqrt{a_{m=0}} > \sqrt{a_{m=1}} + \beta)\lambda_t^{m=0}}{\pi_0[pr(\sqrt{a_{m=0}} \geq \sqrt{a_{m=1}} + \beta)\lambda_t^{m=0} + (1-\pi_0)[pr(\sqrt{a_{m=0}} \geq \sqrt{a_{m=1}} - \delta)\gamma_t^{m=0}]} < \pi_0$$

For this to hold it must be that

$$pr(\sqrt{a_{m=0}} > \sqrt{a_{m=1}} + \beta)\lambda_t^{m=0} < pr(\sqrt{a_{m=0}} \geq \sqrt{a_{m=1}} - \delta)\gamma_t^{m=0}$$

This implies

$$\frac{(1-\delta)^4}{15}(\delta+4) + \frac{2}{3}\beta^2 - \beta - \frac{\beta^5}{15} < \frac{4}{15}$$

using the derivatives of the terms with β and δ, we can infer that the minimum of the left hand side is reached at $\delta = 1$, $\beta = 1$ and its maximum at $\delta = 0$ $\beta = 0$'.

At the minimum the left hand side is equal to 0 so the argument is always true. At the maximum its equal to $frac415$. So for all cases were β and δ are both not equal to zero, $\pi^s_{m_3=0}$, is decreasing.

With the same logic we can show that $\pi^s_{m_3=1} > \pi_0$ the argument will be true if

$$\frac{(1-\beta)^4}{15}(\beta+4) + \frac{2}{3}\delta^2 - \delta - \frac{\delta^5}{15} < \frac{4}{15}$$

using the argument above unless both δ and β are both equal to 1 the argument above always hold.

□

Proof of Lemma 2.3

Proof. From Proposition 2.1, we have identified the last period pay off of the principal and the manager. recall

$$\mathcal{V}_3^P = \mathbb{E}(u_3^P) = \pi_2[pr(\sqrt{a_{m_3=0}} \geq \sqrt{a_{m_3=1}} + \beta)\lambda_3^{m_3=0} + pr(\sqrt{a_{m_3=0}} < \sqrt{a_{m_3=1}} + \beta)(\lambda_3^{m_3=1} + \beta)]$$

$$+(1-\pi_2)[pr(\sqrt{a_{m_3=0}} \geq \sqrt{a_{m_3=1}} - \delta)\gamma_3^{m_3=0} + pr(\sqrt{a_{m_3=0}} < \sqrt{a_{m_3=1}}) - \delta)(\gamma_3^{m_3=1} + \beta)]$$

using the result from Lemma2.2 and Lemma 2.1 and the fact that the δ type manager is less likely to set $m_t = 1$, we can infer that the argument above is increasing in π_2. Therefore it is apparent that if $\mathcal{V}_3^P \geq C$, then the manager is kept. This condition can therefore pin down a threshold of beliefs for each C, where in if the manager reaches that, he can always progress to the last period. To identify that threshold let us first plug in the probabilities in to the utility of the principal and obtain an argument with π_2, δ and β. Plunging in the probabilities obtained in the previous section gives us:

$$\mathcal{V}_3^P = \pi_2[\frac{16}{15} - \frac{(1-\beta)^4(4+\beta)}{15} - \frac{\beta}{2} + \frac{2}{3}\beta^2 - \frac{\beta^5}{15} + \frac{4}{3}\beta^2 - \beta^3 + \frac{\beta^5}{6}]$$

$$+(1-\pi_2)[\frac{16}{15} - \frac{(1-\delta)^4(4+\delta)}{15} - \delta + \frac{2}{3}\delta^2 - \frac{\delta^5}{15} + \beta(\frac{1}{2}(1-\delta)^3(1+\delta/3))] > C$$

Therefore we can define $\underline{\pi}$ as the threshold for progress in the following way

$$\underline{\pi} = \frac{C - [\frac{16}{15} - \frac{(1-\delta)^4(4+\delta)}{15} - \delta + \frac{2}{3}\delta^2 - \frac{\delta^5}{15} + \beta(\frac{1}{2}(1-\delta)^3(1+\delta/3))]}{[2\beta^2 + \frac{\beta^5}{10} - \frac{(1-\beta)^4(4+\beta)}{15} - \frac{\beta}{2} - \beta^3] - [+\frac{2}{3}\delta^2 - \frac{(1-\delta)^4(4+\delta)}{15} - \delta - \frac{\delta^5}{15} + \beta(\frac{1}{2}(1-\delta)^3(1+\delta/3))]}$$

For all $\pi \geq \underline{\pi}$ the principal progresses the manager to the next period and for beliefs below $\underline{\pi}$ she fires the manager. It remains to identify condition on C for $\underline{\pi}$ to exists. If

$$C \leq [\frac{16}{15} + 2\beta^2 + \frac{\beta^5}{10} - \frac{(1-\beta)^4(4+\beta)}{15} - \frac{\beta}{2} - \beta^3]$$

then $\underline{\pi} \leq 1$ and therefore progress will be possible. Since the maximum C can be, is hiring a new manager at prior π_0, this condition is always satisfied. □

Proof of Proposition 2.2

Proof. Suppose $\pi_0 = \underline{\pi}$,

From Lemma 2.1 and Lemma 2.2, we know that if the manager behaves without career concern $\pi_{m2=1}^{S,F} > \pi_1$ and $\pi_{m2=0}^{S,F} < \pi_1$. As argued earlier unless $\beta \to 1$, the equilibrium will always involve mixing at least by one of the two players. Since the belief is at the border, the best that the manager can do is to induce no update. That can only be possible if he completely mimics the β manager's strategy both in choice of employee and effort choice. Therefore his criteria of choice should be

$$m_2^\delta = \begin{cases} 1 & \text{if } \sqrt{a_{m=0}} < \sqrt{a_{m=1}} + \beta, \\ 0 & \text{if } \sqrt{a_{m=0}} \geq \sqrt{a_{m=1}} + \beta \end{cases}$$

In this case best response of the principal would be to some time fire the manager if

she sees $m_2 = 0$. This is the best response of the principal because, given that she does not observe the realised ability of the employee and the applicants, if she believes the strategy of the manager is the one above, the manager can easily deviate from it and follow his per period optimal strategy. Therefore the principal needs to set the probability of firing as to make the manager indifferent on the threshold.

$$\sqrt{a_1} - \delta + \mathcal{V}_3^P = \sqrt{a_0} + q_2^{m_2=0}(-D) + (1 - q_2^{m_2=0})(\mathcal{V}_3^P)$$

setting $q_2^{m_2=0} = \frac{\delta+\beta}{D+\mathcal{V}_3^P}$, would make the manager indifferent at the threshold and there will be no incentive to deviate. □

Proof of Proposition 2.3

Proof. Suppose$\underline{\pi} < \pi_0 < \pi*$, first it can be verified that the strategy of the principal in Proposition 2.2, can not be sustained in the equilibrium.

Firstly fully mimicking of the β type will induce no update. Since the prior is always higher than $\underline{\pi}$, in the equilibrium the principal always deviates and keeps the manager. So this can not be the equilibrium strategy.

Secondly as soon as any strategy of setting the probability of firing lower, such that a lower threshold is enforced cannot be equilibrium. The reason is that this strategy induces $\pi^S_{m=0} \neq \pi^F_{m=0}$ so if the principal is indifferent between firing or keeping the manager in one event, she can not be indifferent in the other event and will deviate.

That leaves the principal with the option of mixed strategy when she observes a failure and $m_2 = 0$. In order to do that the principal needs to set the probability of firing in a way to make the manager indifferent between $m_2 = 0$ and $m_2 = 1$ at the threshold that makes $\pi^F_{m_2=0} = \underline{\pi}$

$$\sqrt{a_1} - \delta + \mathcal{V}_3^P = \sqrt{a_0}(1 + \mathcal{V}_3^P) + (1 - \sqrt{a_0})(q_2^{m_2=0}(-D) + (1 - q_2^{m_2=0})(\mathcal{V}_3^P))$$

Setting $q_2^{m_2=0} = \frac{\kappa}{D+\mathcal{V}_3^P}$.

Given the fact that the non-strategic threshold of the manager is $\sqrt{a_1} - \delta = \sqrt{a_0}$, κ can be lower than δ for π close to $\pi*$. This implies that the equilibrium strategy of the manager will be

$$m_1 = \begin{cases} 1 & \text{if } \sqrt{a_{m=0}} < \frac{\sqrt{a_{m=1}}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa}, \\ 0 & \text{if } \sqrt{a_{m=0}} \geq \frac{\sqrt{a_{m=1}}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa} \end{cases}$$

Let us now check if this strategy makes the principal indifferent between firing or not firing: to do so we first need to derive the probability of success and failure when $m_2 = 1$ and when $m_2 = 0$

1. We start with the case where $\kappa < \delta$, as in the previous case:

$\gamma_2^{m_2=1}=$

$$E(\sqrt{a_1}|\sqrt{a_0} < \frac{\sqrt{a_1}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa}) = \int \sqrt{a_1} f(a_1|\sqrt{a_0} < \frac{\sqrt{a_1}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa}) da_1$$

$$= \frac{1}{p(\sqrt{a_0} < \frac{\sqrt{a_1}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa})} (\int_{(\delta-\kappa)^2}^{(1} \sqrt{a_1} (\frac{\sqrt{a_1}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa})^2 da_1)$$

$$= \frac{\frac{2}{3}(\frac{1+\kappa-\delta}{1+\kappa})^2 - (\frac{1+\kappa-\delta}{1+\kappa})^2 \frac{((\delta-\kappa)((\kappa-\delta-2)(\kappa-\delta)+3)+4)}{10}}{p(\sqrt{a_0} < \frac{\sqrt{a_1}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa})}$$

$$p(\sqrt{a_0} < \frac{\sqrt{a_1}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa}) = \int_0^{(\frac{1+\kappa-\delta}{1+\kappa})^2} \int_{(\sqrt{a_0}(1+\kappa)+\delta-\kappa)^2}^1 da_1 da_0 = (\frac{1+\kappa-\delta}{1+\kappa})^2 (1 - \frac{(\kappa-\delta-2)(\kappa-\delta)+3}{6})$$

$$\gamma_2^{m_2=1} = \frac{\frac{2}{3} - \frac{((\delta-\kappa)((\kappa-\delta-2)(\kappa-\delta)+3)+4)}{10}}{1 - \frac{(\kappa-\delta-2)(\kappa-\delta)+3}{6}}$$

and the probability of failure would then be

$$1 - \gamma_2^{m_2=1} = 1 - \frac{\frac{2}{3}(\frac{1+\kappa-\delta}{1+\kappa})^2 - (\frac{1+\kappa-\delta}{1+\kappa})^2 \frac{((\delta-\kappa)((\kappa-\delta-2)(\kappa-\delta)+3)+4)}{10}}{p(\sqrt{a_0} < \frac{\sqrt{a_1}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa})}$$

$$= \frac{\frac{1}{3} - \frac{(\kappa-\delta-2)(\kappa-\delta)+3}{6} + \frac{2}{3}\frac{((\delta-\kappa)((\kappa-\delta-2)(\kappa-\delta)+3)+4)}{10}}{1 - \frac{(\kappa-\delta-2)(\kappa-\delta)+3}{6}}$$

similarly one can identify $\gamma_2^{m_2=0}=$

$$E(\sqrt{a_0}|\sqrt{a_0} > \frac{\sqrt{a_1}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa}) = \int \sqrt{a_0} f(a_0|\sqrt{a_0} > \frac{\sqrt{a_1}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa})) da_0$$

$$= \frac{1}{p(\sqrt{a_0} > \frac{\sqrt{a_1}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa})} (\int_0^{(\frac{1-\delta+\kappa}{1+\kappa})^2} \sqrt{a_0} (\sqrt{a_0}(1+\kappa) + \delta - \kappa)^2 da_0)$$

$$= \frac{\frac{2}{3}((1+2\kappa-\delta)^2 - \frac{2}{5(1+\kappa)^3}[(1+\kappa-\delta)^5] - \frac{\kappa(1+\kappa-\delta)^4}{2(1+\kappa)^3})}{p(\sqrt{a_0} > \frac{\sqrt{a_1}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa})}$$

$$p(\sqrt{a_0} > \frac{\sqrt{a_1}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa}) = \int_0^{\kappa^2} \int_0^1 da_0 d_1 + \int_{\kappa^2}^{(1+2\kappa-\delta)^2} \int_{(\frac{\sqrt{a_1}-\kappa+\delta}{1+\kappa})^2}^1 da_0 da_1$$

$$= (1+2\kappa-\delta)^2 - \frac{(1+\kappa-\delta)^3(3(1+2\kappa-\delta)+\kappa)}{6(1+\kappa)^2}$$

$$\gamma_2^{m_2=0} = \frac{\frac{2}{3}((1+2\kappa-\delta)^2 - \frac{2}{5(1+\kappa)^3}[(1+\kappa-\delta)^5] - \frac{\kappa(1+\kappa-\delta)^4}{2(1+\kappa)^3})}{(1+2\kappa-\delta)^2 - \frac{(1+\kappa-\delta)^3(3(1+2\kappa-\delta)+\kappa)}{6(1+\kappa)^2}}$$

and the probability of failure would then be

$$1-\gamma_2^{m_2=0} = 1 - \frac{\frac{2}{3}((1+2\kappa-\delta)^2 - \frac{2}{5(1+\kappa)^3}[(1+\kappa-\delta)^5] - \frac{\kappa(1+\kappa-\delta)^4}{2(1+\kappa)^3})}{p(\sqrt{a_0} > \frac{\sqrt{a_1}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa})}$$

$$= \frac{\frac{1}{3}(1+2\kappa-\delta)^2 + \frac{4}{15(1+\kappa)^3}[(1+\kappa-\delta)^5] + \frac{\kappa(1+\kappa-\delta)^4}{3(1+\kappa)^3} - \frac{(1+\kappa-\delta)^3(3(1+2\kappa-\delta)+\kappa}{6(1+\kappa)^2}}{(1+2\kappa-\delta)^2 - \frac{(1+\kappa-\delta)^3(3(1+2\kappa-\delta)+\kappa}{6(1+\kappa)^2}}$$

2. We will now derive the probabilities for $\kappa > \delta$

$\gamma_2^{m_2=1}=$

$$E(\sqrt{a_1}|\sqrt{a_0} < \frac{\sqrt{a_1}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa}) = \int \sqrt{a_1} f(a_1|\sqrt{a_0} < \frac{\sqrt{a_1}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa}) da_1$$

$$= \frac{1}{p(\sqrt{a_0} < \frac{\sqrt{a_1}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa})}(\int_0^1 \sqrt{a_1}(\frac{\sqrt{a_1}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa})^2 da_1)$$

$$= \frac{\frac{2}{3(1+\kappa)^2}}{p(\sqrt{a_0} < \frac{\sqrt{a_1}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa})}\Big[(1+\kappa-\delta)^2 + (\kappa-\delta)^3(1+2(\kappa-\delta) - \frac{3}{2}(\kappa-\delta)(1+2(\kappa-\delta))((1+\kappa-\delta)^2 + (\kappa-\delta)^2)$$

$$-\frac{2}{5}((1-\kappa-\delta)^5 - (\kappa-\delta)^5) - 2(\kappa-\delta)^2[(1+\kappa-\delta)^2 + (\kappa-\delta)(1+\kappa-\delta) + (\kappa-\delta)^2)]\Big]$$

$$p(\sqrt{a_0} < \frac{\sqrt{a_1}}{1+\kappa} + \frac{\kappa-\delta}{1+\kappa}) = \int_0^{(\frac{\kappa-\delta}{1+\kappa})^2} \int_0^1 da_1 da_0 + \int_{(\frac{\kappa-\delta}{1+\kappa})^2}^{(\frac{1+\kappa-\delta}{1+\kappa})^2} \int_{(\sqrt{a_0}(1+\kappa)+\delta-\kappa)^2}^1 da_1 da_0$$

$$= (\frac{1}{1+\kappa})^2(\frac{1}{2} + \frac{4}{3}(\kappa-\delta) + (\kappa-\delta)^2)$$

$$\gamma_2^{m_2=1} = \frac{\frac{2}{3(1+\kappa)^2}}{(\frac{1}{1+\kappa})^2(\frac{1}{2} + \frac{4}{3}(\kappa-\delta) + (\kappa-\delta)^2)}\Big[(1+\kappa-\delta)^2$$

$$+(\kappa-\delta)^3(1+2(\kappa-\delta) - \frac{3}{2}(\kappa-\delta)(1+2(\kappa-\delta))((1+\kappa-\delta)^2 + (\kappa-\delta)^2)$$

$$-\frac{2}{5}((1-\kappa-\delta)^5 - (\kappa-\delta)^5) - 2(\kappa-\delta)^2[(1+\kappa-\delta)^2 + (\kappa-\delta)(1+\kappa-\delta) + (\kappa-\delta)^2)]\Big]$$

and the probability of failure would then be

$$1-\gamma_2^{m_2=1}=1-\frac{\frac{2}{3(1+\kappa)^2}}{(\frac{1}{1+\kappa})^2(\frac{1}{2}+\frac{4}{3}(\kappa-\delta)+(\kappa-\delta)^2)}\Big[(1+\kappa-\delta)^2$$

$$+(\kappa-\delta)^3(1+2(\kappa-\delta)-\frac{3}{2}(\kappa-\delta)(1+2(\kappa-\delta))((1+\kappa-\delta)^2+(\kappa-\delta)^2)$$

$$-\frac{2}{5}((1-\kappa-\delta)^5-(\kappa-\delta)^5)-2(\kappa-\delta)^2[(1+\kappa-\delta)^2+(\kappa-\delta)(1+\kappa-\delta)+(\kappa-\delta)^2)]\Big]$$

similarly one can identify $\gamma_2^{m_2=0}=$

$$E(\sqrt{a_0}|\sqrt{a_0}>\frac{\sqrt{a_1}}{1+\kappa}+\frac{\kappa-\delta}{1+\kappa})=\int\sqrt{a_0}f(a_0|\sqrt{a_0}>\frac{\sqrt{a_1}}{1+\kappa}+\frac{\kappa-\delta}{1+\kappa}))da_0$$

$$=\frac{1}{p(\sqrt{a_0}>\frac{\sqrt{a_1}}{1+\kappa}+\frac{\kappa-\delta}{1+\kappa})}(\int_{(\frac{\kappa-\delta}{1+\kappa})^2}^{(\frac{1-\delta+\kappa}{1+\kappa})^2}\sqrt{a_0}(\sqrt{a_0}(1+\kappa)+\delta-\kappa)^2da_0)$$

$$=\frac{\frac{2}{3}(1-\frac{1}{(1+\kappa)^3}[\frac{2}{5}+\frac{3}{2}(\kappa-\delta)+2(\kappa-\delta)^2+(\kappa-\delta)^3])}{p(\sqrt{a_0}>\frac{\sqrt{a_1}}{1+\kappa}+\frac{\kappa-\delta}{1+\kappa})}$$

$$p(\sqrt{a_0}>\frac{\sqrt{a_1}}{1+\kappa}+\frac{\kappa-\delta}{1+\kappa})=\int_0^1\int_{(\frac{\sqrt{a_1}-\delta+\kappa}{1+\kappa})^2}^1 da_0da_1$$

$$=1-(\frac{1}{1+\kappa})^2(\frac{1}{2}+\frac{4}{3}(\kappa-\delta)+(\kappa-\delta)^2)$$

$$\gamma_2^{m_2=1}=\frac{\frac{2}{3}(1-\frac{1}{(1+\kappa)^3}[\frac{2}{5}+\frac{3}{2}(\kappa-\delta)+2(\kappa-\delta)^2+(\kappa-\delta)^3])}{1-(\frac{1}{1+\kappa})^2(\frac{1}{2}+\frac{4}{3}(\kappa-\delta)+(\kappa-\delta)^2)}$$

and the probability of failure would then be

$$1-\gamma_2^{m_2=0}=1-\frac{\frac{2}{3}(1-\frac{1}{(1+\kappa)^3}[\frac{2}{5}+\frac{3}{2}(\kappa-\delta)+2(\kappa-\delta)^2+(\kappa-\delta)^3])}{p(\sqrt{a_0}>\frac{\sqrt{a_1}}{1+\kappa}+\frac{\kappa-\delta}{1+\kappa})}$$

$$=\frac{\frac{1}{3}+\frac{2}{3(1+\kappa)^3}[\frac{2}{5}+\frac{3}{2}(\kappa-\delta)+2(\kappa-\delta)^2+(\kappa-\delta)^3]-\frac{1}{6(1+\kappa)^2}[3+8(\kappa-\delta)+6(\kappa-\delta)^2]}{1-(\frac{1}{1+\kappa})^2(\frac{1}{2}+\frac{4}{3}(\kappa-\delta)+(\kappa-\delta)^2)}$$

given these probabilities we now need to verify if κ exists. Consider the case $e_2=1$

$$\pi_2^{m_2=0,X_2=0}=\frac{\pi_1[pr(\sqrt{a_{m=0}}\geq\sqrt{a_{m=1}}+\beta)(1-\lambda_t^{m=0})]}{\pi_1[pr(\sqrt{a_{m=0}}\geq\sqrt{a_{m=1}}+\beta)(1-\lambda_t^{m=0})]+(1-\pi_1)[pr(\sqrt{a_0}>\frac{\sqrt{a_1}}{1+\kappa}+\frac{\kappa-\delta}{1+\kappa})(1-\gamma_t^{m=0})]}$$

$$\pi_2^{m_2=0,X_2=0} =$$

$$\frac{p[\frac{1}{10}-\frac{\beta}{3}+\frac{\beta^2}{3}-\frac{\beta^4}{6}+\frac{\beta^5}{15}]}{p[\frac{1}{10}-\frac{\beta}{3}+\frac{\beta^2}{3}-\frac{\beta^4}{6}+\frac{\beta^5}{15}]+(1-p)[\frac{1}{3}+\frac{2}{3(1+\kappa)^3}[\frac{2}{5}+\frac{3}{2}(\kappa-\delta)+2(\kappa-\delta)^2+(\kappa-\delta)^3]-\frac{1}{6(1+\kappa)^2}[3+8(\kappa-\delta)+6(\kappa-\delta)^2]]}$$

□

Proof of Lemma 2.4

Proof. Based on probabilities we derived in the previous section we can now characterise $\mathcal{V}_2^{\pi_1,Mix}$ and compare it with $\mathcal{V}^{sab}$ and establish the condition under which sabotage is optimal. To be concrete lets first define $\mathcal{V}_2^{\pi_1,Mix}$ and $\mathcal{V}^{sab}$:

$$\mathcal{V}^{sab} = 2\mathcal{V}_3^{\delta} = 2\nu\big(\gamma_3^{m_3=0}p(\sqrt{a_0}-\sqrt{a_1}>-\delta)+p(\sqrt{a_0}-\sqrt{a_1}<-\delta)(\gamma_3^{m_3=1}-\delta)\big)$$

$$\mathcal{V}^{sab} = 2\nu\big(\frac{2}{5}-\delta+\frac{2}{3}\delta^2-\frac{\delta^5}{15}-\delta(\frac{1}{2}(1-\delta)^3(1+\delta/3))+\frac{2}{3}-\frac{(1-\delta)^4(4+\delta)}{15}\big)$$

$$\mathcal{V}_2^{\pi_1,Mix} = p(\sqrt{a_0}<\frac{\sqrt{a_1}}{1+\kappa}+\frac{\kappa-\delta}{1+\kappa})(\nu(\gamma_2^{m_2=1}-\delta)+\mathcal{V}_3^{\delta})+p(\sqrt{a_0}>\frac{\sqrt{a_1}}{1+\kappa}+\frac{\kappa-\delta}{1+\kappa})((\nu+\mathcal{V}_3^{\delta})\gamma_2^{m_2=0}-\kappa(1-\gamma_2^{m_2=0}))$$

For π closer to π^*, $\kappa<\delta$ Therefore:

$$\mathcal{V}_2^{\pi_1,Mix} = \nu(\frac{2}{3}-\frac{((\delta-\kappa)((\kappa-\delta-2)(\kappa-\delta)+3)+4)}{10})-\nu\delta((\frac{1+\kappa-\delta}{1+\kappa})^2(1-\frac{(\kappa-\delta-2)(\kappa-\delta)+3}{6})$$

$$+\frac{2\nu}{3}\big((1+2\kappa-\delta)^2-\frac{2}{5(1+\kappa)^3}[(1+\kappa-\delta)^5]-\frac{\kappa(1+\kappa-\delta)^4}{2(1+\kappa)^3}\big)$$

$$-\kappa(\frac{1}{3}(1+2\kappa-\delta)^2+\frac{4}{15(1+\kappa)^3}[(1+\kappa-\delta)^5]+\frac{\kappa(1+\kappa-\delta)^4}{3(1+\kappa)^3}-\frac{(1+\kappa-\delta)^3(3(1+2\kappa-\delta)+\kappa}{6(1+\kappa)^2})$$

$$+\mathcal{V}_3^{\delta}((\frac{1+\kappa-\delta}{1+\kappa})^2(1-\frac{(\kappa-\delta-2)(\kappa-\delta)+3}{6})+(\frac{2}{3}((1+2\kappa-\delta)^2-\frac{2}{5(1+\kappa)^3}[(1+\kappa-\delta)^5]-\frac{\kappa(1+\kappa-\delta)^4}{2(1+\kappa)^3}))$$

For π closer to $\underline{\pi}$, $\kappa>\delta$. Therefore the expression changes to:

$$\mathcal{V}_2^{\pi_1,Mix} = \frac{2\nu}{3(1+\kappa)^2}(\Big[(1+\kappa-\delta)^2+(\kappa-\delta)^3(1+2(\kappa-\delta)-\frac{3}{2}(\kappa-\delta)(1+2(\kappa-\delta))((1+\kappa-\delta)^2+(\kappa-\delta)^2)$$

$$-\frac{2}{5}((1-\kappa-\delta)^5-(\kappa-\delta)^5)-2(\kappa-\delta)^2[(1+\kappa-\delta)^2+(\kappa-\delta)(1+\kappa-\delta)+(\kappa-\delta)^2)]\Big])-\nu\delta(\frac{1}{1+\kappa})^2(\frac{1}{2}+\frac{4}{3}(\kappa-\delta)+(\kappa-\delta)^2)$$

$$+\frac{2\nu}{3}(1-\frac{1}{(1+\kappa)^3}[\frac{2}{5}+\frac{3}{2}(\kappa-\delta)+2(\kappa-\delta)^2+(\kappa-\delta)^3])$$

$$-\kappa(\frac{1}{3}+\frac{2}{3(1+\kappa)^3}[\frac{2}{5}+\frac{3}{2}(\kappa-\delta)+2(\kappa-\delta)^2+(\kappa-\delta)^3]-\frac{1}{6(1+\kappa)^2}[3+8(\kappa-\delta)+6(\kappa-\delta)^2])$$

$$+\mathcal{V}_3^{\delta}(\frac{1}{(1+\kappa)^2}(\frac{1}{2}+\frac{4}{3}(\kappa-\delta)+(\kappa-\delta)^2)+\frac{2}{3}(1-\frac{1}{(1+\kappa)^3}[\frac{2}{5}+\frac{3}{2}(\kappa-\delta)+2(\kappa-\delta)^2+(\kappa-\delta)^3]))$$

For each of the two cases we need to show that $1+\mathcal{V}_2^{\pi_1,Mix}<\mathcal{V}^{sab}$.

□

Proof of Lemma 2.5

Proof. Proof by Contradiction: suppose the argument does not hold, that is when there is positive probability of sabotage $\pi^s_{m_1}>\pi*$ and $\pi^f_{m_1}=\pi*$, then the manager has always an incentive to deviate and set $e_1=1$ and never sabotage. Therefore for the sabotage equilibrium to exist it must be that $\pi^s_{m_1}<\pi*$ and $\pi^f_{m_1}=\pi*$ □

Proof of Lemma 2.6

Proof.

1. Suppose the first argument does not hold, then the manager is always better off deviating and setting $e_1=1$ in either cases and the sabotage equilibrium breaks down.

2. Suppose the second argument fails, then the manager would always want to deviate and set $m_1=1$. But this breaks down the equilibrium.

So for sabotage equilibrium to exist, it must be the case that both of the conditions in the Lemma are met. □

Proof of Proposition 2.4

Proof. To start the proof, we should emphasize that the sabotage equilibrium would only be possible if $\pi>\underline{\pi}$. When $\pi=\underline{\pi}$, the manager's strategy should either induce no update or upward update of beliefs. While we will show that the equilibrium strategy of sabotage will induce some downward belief update when $m_1=0$ is observed by the principal.

As described in the text sabotage is a reputation building strategy in so long as the principal believes sabotage is not happening with certainty. That is, it's never optimal for the manager to sabotage with probability one. The reason is that if he always sabotages then realisation of a success with $m_1=1$ will only come from a β type manager. Since this implies both higher current and future payoff, the manager will always deviate from sabotaging and sets $e_1=1$. So the manager will only sabotage

if the principal believes sabotage is happening with some positive probability and not with certainty. Now that we have established sabotage being a mixed strategy and not a pure one, we need to identify the optimal sabotage strategy of the manager. Let us look at the strategy of the manager where he sabotages the $m_1 = 1$ with high enough probability such that the principal belief upon observing $m_1 = 1$ and $X_1 = 0$ reaches $\pi*$. Given this belief update the principal in the second period will be indifferent between firing or keeping the manager if she observes $m_2 = 0$ and $X_2 = 0$. Therefore her best response will be to randomise between keeping and firing the manager if she observes $m_2 = 0$ and $X_2 = 0$, such that the manager will be indifferent between sabotaging and not sabotaging in the first period i.e. $\pi_1^{F,sab,a_1} = \pi^*$

Let us check if this is an equilibrium strategy for both principal and the manager. Given the randomisation strategy of the manager, at the end of period one $\pi_1^{m=1,X=0} > \underline{\pi}$ so the principal has no incentive to deviate and fire the manager. Also in the second period given Lemma 2.1 and Lemma 2.2, the principal has no incentive to deviate and fire the manager if she does not observe $m_2 = 0$ and $X_2 = 0$. If she does observe $m_2 = 0$ and $X_2 = 0$, she is indifferent between firing or keeping the manager so there is no incentive to deviate.

Deviation is not optimal for the manager too. Given the mixing strategy of the principal, he gets same sum of present value of future and current pay off, so he has no incentive to deviate from his sabotage equilibrium.

It remains to characterise the equilibrium probabilities and check if the equilibrium is sustained in the entire sub game.

To specify the equilibrium probabilities we start with sabotage probability. Let us specify the utility of the manager from sabotage

$$\mathcal{V}_1^{sab} = 2U_3^{\delta}\Big(pr(\sqrt{a_0} < \sqrt{a_1} - \delta) + \gamma_2^{m_2=0} pr(\sqrt{a_0} > \sqrt{a_1} - \delta)\Big) + (1 - \gamma_2^{m_2=0}) pr(\sqrt{a_0} > \sqrt{a_1} - \delta)\Big[q_2^{sab}(-D) + (1 - q^{sab})U_3^{\delta}\Big]$$

For each realization of $a_1 \sim u[0, 1]$, the manager's utility from not sabotaging would be

$$\mathcal{V}_1^{\delta} = \sqrt{a_1}(1 + \mathcal{V}_2^{Mix}) + (1 - \sqrt{a_1})\Bigg[2U_3^{\delta}\Big(pr(\sqrt{a_0} < \sqrt{a_1} - \delta) + \gamma_2^{m_2=0} pr(\sqrt{a_0} > \sqrt{a_1} - \delta)\Big)$$

$$+ (1 - \gamma_2^{m_2=0}) pr(\sqrt{a_0} > \sqrt{a_1} - \delta)\Big[q_2^{sab}(-D) + (1 - q^{sab})U_3^{\delta}\Big]\Bigg]$$

Define $\omega = (1 - \gamma_2^{m_2=0}) pr(\sqrt{a_0} > \sqrt{a_1} - \delta)$

The principal will set q_2^{sab} such that $\mathcal{V}_1^{\delta} = \mathcal{V}_1^{sab}$.

In the equilibrium $q_2^{*sab} = \frac{\omega D+1+\mathcal{V}_{mix}-(2-\omega)\mathcal{V}_{sab}}{\omega(\mathcal{V}_{sab}+D)}$

For q_2^{*sab} to exist

1. $\omega D + 1 + \mathcal{V}_{mix} - (2-\omega)\mathcal{V}_{sab} > 0$, Since D is assumed to be big, this condition is fulfilled.

2. $\omega D + 1 + \mathcal{V}_{mix} - (2-\omega)\mathcal{V}_{sab} < \omega(\mathcal{V}_{sab} + D)$, this is also satisfied as long as optimality condition in Lemma 2.4 is satisfied.

We now need to characterise the equilibrium probability of sabotage, recall, for sabotage to be an equilibrium strategy it must be that Lemma 2.5 and Lemma 2.6 are satisfied. We know that sabotage should push up the beliefs of the principal after observing $m_1 = 1$ and $X_1 = 0$ to $\pi*$. That means

$$\pi* = \frac{(1-\lambda_{m_1=1})pr(m_1=1|\theta=\beta)\pi_0}{(1-\lambda_{m_1=1})pr(m_1=1|\theta=\beta)\pi_0 + [\eta+(1-\eta)(1-\gamma_{m_1=1})]pr(m_1=1|\theta=\delta)(1-\pi_0)}$$

This implies that

$$\eta^*_{m_1=1} = \frac{(\beta(\frac{4}{3}-\beta) + \frac{(1-\beta)^4(3+2\beta}{30})\pi(1-\pi*) - [(1-\gamma_{m_1=1})pr(m_1=1|\theta=\delta)(1-\pi_0)\pi*]}{(\gamma_{m_1=1})pr(m_1=1|\theta=\delta)(1-\pi_0)\pi*}$$

Lemma 2.5 also specifies that

$$\pi* > \frac{(\lambda_{m_1=1})pr(m_1=1|\theta=\beta)\pi_0}{(\lambda_{m_1=1})pr(m_1=1|\theta=\beta)\pi_0 + [(1-\eta)(\gamma_{m_1=1})pr(m_1=1|\theta=\delta)(1-\pi_0)]}$$

As mentioned earlier for $\eta^*_{m_1=1}$ to exist it must be that the conditions below are satisfied

1. $(1-\lambda_{m_1=1})pr(m_1=1|\theta=\beta)(1-\pi*)\pi_0 > (1-\gamma_{m_1=1})pr(m_1=1|\theta=\delta)(1-\pi_0)\pi*$, This condition is only satisfied when Lemma 2.5 is satisfied. This will pin down maximum $pr(m_1=1|\theta=\delta)$. We will further specify the existence of this condition once we solve for the entire game and $pr(m_1=1|\theta=\delta)$is characterised.

2. $(1-\lambda_{m_1=1})pr(m_1=1|\theta=\beta)(1-\pi*)\pi_0 - [(1-\gamma_{m_1=1})pr(m_1=1|\theta=\delta)(1-\pi_0)\pi*] < (\gamma_{m_1=1})pr(m_1=1|\theta=\delta)(1-\pi_0)\pi*$.

   This condition can be simplified in to

$$\pi_0(1-\pi*)(1-\lambda_{m_1=1})pr(m_1=1|\theta=\beta) < \pi*(1-\pi_0)pr(m_1=1|\theta=\delta)$$

   Given $pr(m_1=1|\theta=\delta)$ specified in point 1, this point defines an upper bound

for π_0 such that

$$\pi_0 \leq \frac{pr(m_1 = 1|\theta = \delta)}{\pi^* pr(m_1 = 1|\theta = \delta) + (1 - \lambda_{m_1=1}) pr(m_1 = 1|\theta = \beta)(1 - \pi*)}$$

We will further specify this upper bound once we solve for the entire game and $pr(m_1 = 1|\theta = \delta)$is characterised.

3. Finally the condition in Lemma 2.5 for probability of success given sabotage specifies a lower bound for π_0. For the condition to hold it must be that

$$(1-\lambda_{m_1=1})pr(m_1 = 1|\theta = \beta)(1-\pi*)\pi_0 - [(1-\gamma_{m_1=1})pr(m_1 = 1|\theta = \delta)(1-\pi_0)\pi*]$$

$$> (\gamma_{m_1=1})pr(m_1 = 1|\theta = \delta)(1 - \pi_0)\pi * -[\lambda_{m_1=1} pr(m_1 = 1|\theta = \beta)(1 - \pi*)\pi_0]$$

This further can be simplified in to

$$\pi_0(1 - \pi*)pr(m_1 = 1|\theta = \beta) > \pi * (1 - \pi_0)pr(m_1 = 1|\theta = \delta)$$

Once again, given $pr(m_1 = 1|\theta = \delta)$ specified in point 1, this point defines a lower bound for π_0 such that

$$\pi_0 > \frac{\pi * pr(m_1 = 1|\theta = \delta)}{\pi^* pr(m_1 = 1|\theta = \delta) + (1 - \pi*)pr(m_1 = 1|\theta = \beta)}$$

We will further specify the lower bound on prior belief once we solve for the entire game and $pr(m_1 = 1|\theta = \delta)$is characterised.

We now turn to specifying $pr(m_1 = 1|\theta = \delta)$, going back to Lemma 2.6, we know that the belief update should be such that the manager chooses $m_1 = 0$ when $a_{m_1=0}$ is large enough. Recall form sabotage condition, the randomisation strategy of the principal is such that the manager's utility from setting $m_1 = 1$ is $\mathbb{U}^{\delta}_{m_1=1} = 1 + \mathcal{V}_2^{Mix}$ and given positive probability of sabotage, it must be that the manager does not want to deviate from setting $m_1 = 0$ when $a_{m_1=0}$ is large enough. The manager will set $m_1 = 0$ when $\mathbb{U}^{\delta}_{m_1=1} < \mathbb{U}^{\delta}_{m_1=0} - \delta$ that is when

$$\sqrt{a_0}(1 + \mathcal{V}^{Mix,\pi^S_{m=0}}) + (1 - \sqrt{a_0})(\mathcal{V}^{\pi^F_{m=0}}) > 1 + \mathcal{V}^{Mix}_{m_1=1} - \delta$$

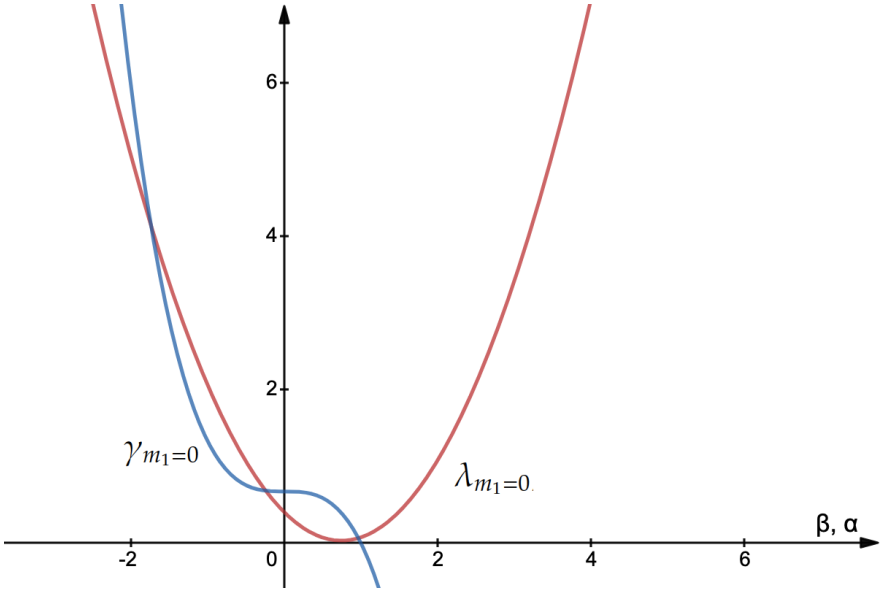


Figure 2.A.1: Lemma 2.2 with sabotage $\gamma_{m_1=0} > \lambda_{m_1=0}$

or more explicitly when

$$\sqrt{a_0} > \frac{1 + \mathcal{V}^{Mix}_{m_1=1} - \delta - \mathcal{V}^{\pi^F_{m=0}}}{1 + \mathcal{V}^{Mix,\pi^S_{m=0}} - \mathcal{V}^{\pi^F_{m=0}}}$$

Define

$$\alpha^2_\delta = \left(\frac{1 + \mathcal{V}^{Mix}_{m_1=1} - \delta - \mathcal{V}^{\pi^F_{m=0}}}{1 + \mathcal{V}^{Mix,\pi^S_{m=0}} - \mathcal{V}^{\pi^F_{m=0}}}\right)^2$$

as the threshold for setting $m_1 = 0$, for the equilibrium to exist two conditions needs to be satisfied

1. $\alpha_\delta < 1$, for this condition to be true it must be that $\mathcal{V}^{Mix}_{m_1=1} - \mathcal{V}^{Mix,\pi^S_{m=0}} < \delta$. Given α, $\gamma_{m_1=0} = \frac{\frac{2}{3}(1-\alpha^3)}{1-\alpha^2}$ and $\gamma_{m_1=1} = \frac{2}{3}$ We can therefore specify :

$$\pi^S_{m=0} = \frac{(\frac{2}{5} - \beta + \frac{2}{3}\beta^2 - \frac{\beta^5}{15})\pi_0}{(\frac{2}{5} - \beta + \frac{2}{3}\beta^2 - \frac{\beta^5}{15})\pi_0 + \frac{2}{3}(1-\alpha^3)(1-\pi_0)}$$

Figure 2.A.1 plots $\gamma_{m_1=0}$ and $\lambda_{m_1=0}$, and

$$\pi^S_{m=1} = \frac{(\frac{2}{3} - \frac{(1-\beta)^4(\beta+4)}{15})\pi_0}{(\frac{2}{3} - \frac{(1-\beta)^4(\beta+4)}{15})\pi_0 + \frac{2}{3}\alpha^2(1-\pi_0)}$$

Figure 2.A.2 plots $\gamma_{m_1=1}$ and $\lambda_{m_1=1}$

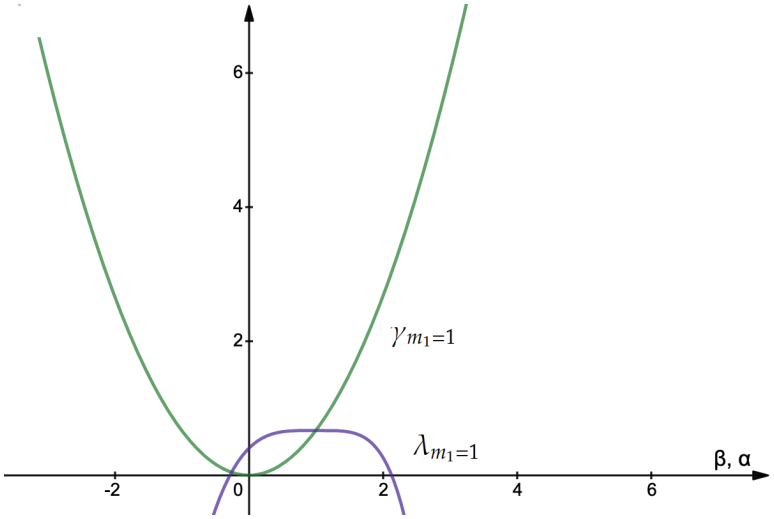


Figure 2.A.2: Lemma 2.2 with sabotage $\gamma_{m_1=1}<\lambda_{m_1=1}$

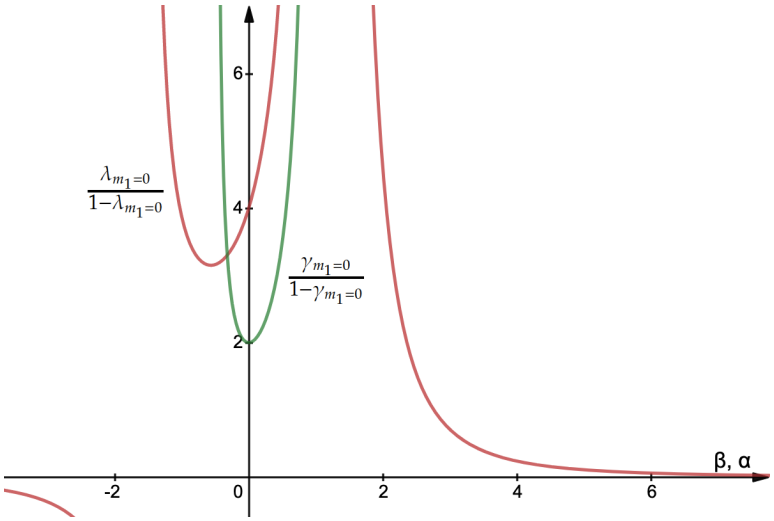


Figure 2.A.3: Lemma 2.1 with sabotage $\frac{\lambda_{m_1=0}}{1-\lambda_{m_1=0}}>\frac{\gamma_{m_1=0}}{1-\gamma_{m_1=0}}$

The two graph show that the condition in Lemma 2.2 is satisfied and $\pi^S_{m=0}<\pi^S_{m=1}$. This implies $\mathcal{V}^{Mix}_{m_1=1} > \mathcal{V}^{Mix,\pi^S_{m=0}}$. Therefore $\exists\ \delta > \mathcal{V}^{Mix}_{m_1=1} - \mathcal{V}^{Mix,\pi^S_{m=0}}$ for which $\alpha_\delta < 1$. It can be observed that for large δ, α will be small.

2. The condition in Lemma 2.1 is also satisfied since

   $1-\gamma_{m_1=0} = \frac{1}{3} - \alpha^2 + \frac{2}{3}\alpha^3$, Figure 2.A.3 plots $\frac{\lambda_{m_1=0}}{1-\lambda_{m_1=0}}$ and $\frac{\gamma_{m_1=0}}{1-\gamma_{m_1=0}}$ and proves that these condition holds for low enough α, that is when δ is big enough.

It remains to check if given the new threshold $pr(m_1 = 1|\theta = \delta)$, the conditions in Lemma 2.1 and Lemma 2.2 are satisfied. In the previous section Lemma 2.2 was shown to be satisfied so it remain to check Lemma 2.1, since $1-\gamma_{m_1=1} = \frac{1}{3}$ then it

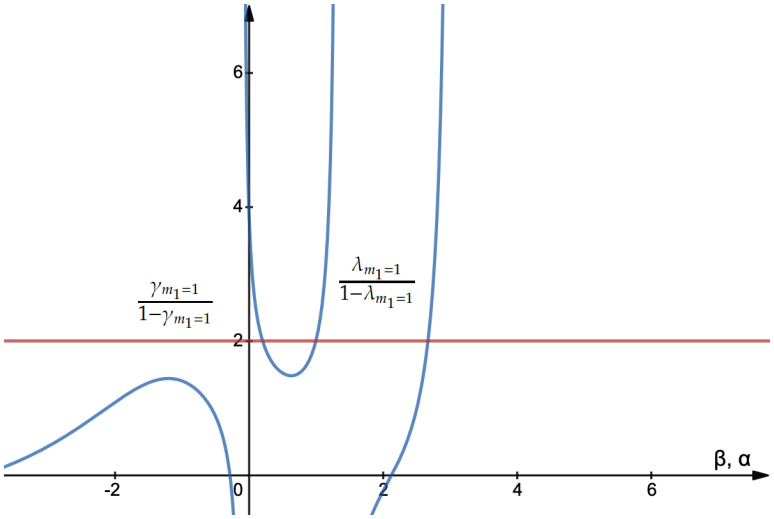


Figure 2.A.4: Lemma 2.1 with sabotage $\frac{\lambda_{m_1=1}}{1-\lambda_{m_1=1}} < \frac{\gamma_{m_1=1}}{1-\gamma_{m_1=1}}$

must be that $\frac{\gamma_{m_1=1}}{1-\gamma_{m_1=1}} = 2$ plotting this with $\frac{\lambda_{m_1=1}}{1-\lambda_{m_1=1}}$ in Figure 2.A.4, shows that for all α if β is large, the condition will hold.

To finish the proof of this Proposition, we will now return to the conditions for probability of sabotage to exist. The three conditions specified there will now be characterised in the following way:

1. The sabotage equilibrium exists if $\alpha < \alpha*$ where $\beta(\frac{4}{3} - \beta) + \frac{(1-\beta)^4(3+2\beta)}{30})(1 - \pi*)\pi_0 = \frac{1}{3}(\alpha*)^2(1-\pi_0)\pi*$ This condition can be satisfied if δ is high.

$$\eta^*_{m_1=1} = \frac{(\beta(\frac{4}{3} - \beta) + \frac{(1-\beta)^4(3+2\beta)}{30})\pi_0(1-\pi*) - [\frac{1}{3}\alpha^2(1-\pi_0)\pi*]}{\frac{2}{3}\alpha^2(1-\pi_0)\pi*}$$

2. This condition will further simplify to

$$\pi_0 \leq \frac{\alpha^2}{\pi^*\alpha^2 + (\beta(\frac{4}{3} - \beta) + \frac{(1-\beta)^4(3+2\beta)}{30})(1-\pi*)}$$

3. The lower bound of belief for sabotage then can be specified as

$$\pi_0 > \frac{\pi^*\alpha^2}{\pi^*\alpha^2 + (1-\pi*)(\frac{1}{2} + \frac{4}{3}\beta - \beta^2 + \frac{\beta^4}{6})}$$

□

Bibliography

Ali, Omer, "Disclosure in Multistage Projects," *Working paper*, 2017.

Arrow, Kenneth J, *The theory of discrimination*, Princeton, NJ: Princeton University press, 1973.

Aumann, Robert J and Sergiu Hart, "Long cheap talk," *Econometrica*, November 2003, *71* (6), 1619–1660.

Auriol, Emmanuelle and Friebel Guido, "Career Concerns in Teams," *working papaer*, August 2000.

Bandura, Albert, "Self-efficacy: toward a unifying theory of behavioral change.," *Psychological review*, 1977, *84* (2), 191.

Bar-Issac, Heski and Joyee Deb, "Reputation with Opportunities for Coasting," *working papaer*, April 2018.

Becker, Gary S, *The Economics of Discrimination*, 2 ed., Chicago, IL: Chicago University Press, 1971.

Bénabou, Roland and Guy Laroque, "Using Privileged Information to Manipulate Markets: Insiders, Gurus, and Credibility," *The Quarterly Journal of Economics*, 1992, *107* (3), 921–948.

_ **and Jean Tirole**, "Self-confidence and personal motivation," *The Quarterly Journal of Economics*, 2002, *117* (3), 871–915.

_ **and** _ , "Identity, Morals and Taboos: Beliefs as Assets," *The Quarterly Journal of Economics*, 2011, *126* (2), 805–855.

_ **, Armin Falk, and Jean Tirole**, "Narratives, Imaperatives and Moral Persuasion," Technical Report September 2019.

Bizzotto, Jacopo, Jesper Rüdiger, and Adrien Vigier, "Dynamic Persuasion with Outside Information," *Working paper*, 2018.

Boleslavsky, Raphael and Tracy R Lewis, "Evolving influence: Mitigating extreme conflicts of interest in advisory relationships," *Games and Economic Behavior*, 2016, *98*, 110–134.

Chalioti, Evangelia, "Incentives to help or sabotage co-workers," *Working paper*, 2019.

Che, Yeon-Koo and Johannes Horner, "Optimal design for social learning," *Working Paper*, 2015.

Coate, Stephan and Glenn Loury, "Antidiscrimination Enforcement and the Problem of Patronization," *The American Economic Review, Papers and proceedings*, 1993, *83* (2), 92–98.

Cripps, Martin W, George J Mailath, and Larry Samuelson, "Imperfect Monitoring and Impermanent Reputations," *Econometrica*, 2004, *72* (2), 407–432.

Deb, Joyee and Yuhta Ishii, "Reputation Building under Uncertain Monitoring," *Working Paper*, 2018.

Diamond, Douglas W, "Monitoring and Reputation: The Choice between Bank Loans and Directly Placed Debt," *The Journal of Political Economy*, August 1991, *99* (4), 689–721.

Ely, Jeffrey C, "Beeps," *The American Economic Review*, 2017, *107* (1), 31–53.

_ **and Juuso Välimäki**, "Bad Reputaion," *The Quarterly Journal of Economics*, 2003, *118* (3), 785–814.

_ **and Martin Szydlowski**, "Moving the Goalposts," *Working paper*, 2017.

_ **, Drew Fudenberg, and David K Levine**, "When is reputation bad?," *Games and Economic Behavior*, 2008, *63* (2), 498–526.

Forges, Françoise and Frédéric Koessler, "Long persuasion games," *Journal of Economic Theory*, 2008, *143* (1), 1–35.

Fuchs, William, "Contracting with repeated moral hazard and private evaluations," *The American Economic Review*, 2007, *97* (4), 1432–1448.

Fudenberg, Drew and David K Levine, "Maintaining a Reputation when Strategies are Imperfectly Observed," *Review of Economic Studies*, July 1992, *59* (3), 561–579.

Gentzkow, Matthew and Jesse M Shapiro, "Media Bias and Reputation," *The Journal of Political Economy*, 2006, *114* (2), 280–316.

Golosov, Mikhail, Vasiliki Skreta, Aleh Tsyvinski, and Andrea Wilson, "Dynamic strategic information transmission," *Journal of Economic Theory*, 2014, *151*, 304–341.

Goltsman, Maria, Johannes Hörner, Gregory Pavlov, and Francesco Squintani, "Mediation, arbitration and negotiation," *Journal of Economic Theory*, 2009, *144* (4), 1397–1420.

Gossner, Olivier, "Simple Bounds on the Value of a Reputation," *Econometrica*, 2011, *79* (5), 1627–1641.

Halac, Marina and Andrea Prat, "Managerial Attention and Worker Performance," *American Economic Review*, 2016, *106* (10), 3104–3132.

Henry, Emeric and Marco Ottaviani, "Research and the Approval Process: The Organization of Persuasion," *American Economic Review*, March 2019, *109* (3), 911–55.

Hirshleifer, David, Angie Low, and Siew Hong Teoh, "Are overconfident CEOs better innovators?," *The Journal of Finance*, 2012, *67* (4), 1457–1498.

Holmstorm, Bengt, "Managerial Incentive Problems: A Dynamic Perspective," *The Review of Economic Studies*, 1999, *66* (1), 169–182.

Honryo, Takakazu, "Dynamic persuasion," *Journal of Economic Theory*, 2018, *178*, 36 – 58.

Hörner, Johannes and Andrzej Skrzypacz, "Learning, experimentation and information design," *Working Paper*, 2016.

Kamenica, Emir and Matthew Gentzkow, "Bayesian persuasion," *The American Economic Review*, 2011, *101* (6), 2590–2615.

Kamphorst, Jurjen J. A and Otto H Swank, "Don't demotivate, discriminate," *American Economic Journal: Microeconomics*, 2016, *8* (1).

Koellinger, Philipp, Maria Minniti, and Christian Schade, "âĂIJI think I can, I think I canâĂİ: Overconfidence and entrepreneurial behavior," *Journal of economic psychology*, 2007, *28* (4), 502–527.

Konrad, Kai A., "Sabotage in Rent-Seeking Contests," *Journal of Law, Economics, & Organization,*, 2000, *16* (1), 155–165.

Kremer, Ilan, Yishay Mansour, and Motty Perry, "Implementing the "Wisdom of the Crowd"," *Journal of Political Economy*, 2014, *122* (5), 988–1012.

Krishna, Vijay and John Morgan, "The art of conversation: eliciting information from experts through multi-stage communication," *Journal of Economic theory*, 2004, *117* (2), 147–179.

Lang, Kevin and Jee-Yeon K Lehmann, "Racial Discrimination in the Labor Market: Theory and Empirics," *Journal of Economic Literature*, 2012, *50* (4), 959–1006.

Lazear, Edward P and Sherwin Rosen, "Rank-order tournaments as optimum labor contracts," *The Journal of Political Economy*, 1981, *89* (5), 841–864.

Lizzeri, Alessandro, Margaret A Meyer, and Nicola Persico, *The incentive effects of interim performance evaluations*, University of Pennsylvania, Center for Analytic Research in Economics and the Social Sciences, 2002.

Milgrom, Paul and Sharon Oster, "Job Discrimination, Market Forces, and the Invisibility Hypothesis," *The Quarterly Journal of Economics*, 1987, *102* (3), 453–476.

Morris, Stephen, "Political Correctness," *Journal of Political Economy*, 2001, *109* (2), 231–265.

_ , "Political Correctness," *Journal of Political Economy*, April 2001, *109* (2), 231–265.

Orlov, Dmitry, "Optimal Design of Internal Disclosure," *Working paper*, 2013.

_ **, Andrzej Skrzypacz, and Pavel Zryumov**, "Persuading the principal to wait," *Working paper*, 2018.

Phelps, Edmund S, "The Statistical Theory of Racism and Sexism," *The American Economic Review*, 1972, *62* (4), 659–661.

Renault, Jérôme, Eilon Solan, and Nicolas Vieille, "Dynamic sender–receiver games," *Journal of Economic Theory*, 2013, *148* (2), 502–534.

_ **,** _ **, and** _ , "Optimal dynamic information provision," *Games and Economic Behavior*, 2017, *104*, 329–349.

Shin, Wiroy, "Discrimination in Organizations: Optimal Contracts and Regulation," *Working Paper*, 2016.

Smolin, Alex, "Dynamic Evaluation Design," *Working Paper*, 2017.