# The Diversity Paradox

Zeinab Aboutalebi
Department of Economics, The University of Warwick *

**Abstract**

Diversity related reputation is becoming increasingly important for managers in organisations. We study a principal manager career concern relationship where manager and principal may not have an identical bias toward diversity. In such a setting, the misaligned manager faces the following trade-off; while hiring minorities reduces his utility, not hiring them may cost him his career. We show that when the success of employees depends on their ability and manager's effort, with low reputation, a positive bias of the principal induces sabotage of minority groups. If the principal has no bias toward diversity, diversity marginally improves. However, if the principal has a positive bias toward diversity, the misaligned manager improves reputation by hiring more from minority groups but sabotages them. We define this, diversity paradox; if there is no positive bias toward diversity, diversity does not improve much. However, if there is, diversity improves at the cost of increased sabotage. We show that minorities in low productivity jobs are more likely to be sabotaged.

**Keywords:** Discrimination, Sabotage, Reputation, Persuasion

**JEL Classification Numbers:**

*Email: Z.Aboutalebi@warwick.ac.uk. Tel: +44 (0) 7787203822. Address: Department of Economics, The University of Warwick, Coventry CV4 7AL, UK. This work was supported by the Economic and Social Research Council [grant number ES/J500203/1].I would like to thank Motty Perry, Debraj Ray, Robert Akerlof, Jacob Glazer, Ilan Kremer, Costas Cavounidis, Ayush Pant and Federico Trombetta for their time and insightful comments. All errors are my own.

# 1 Introduction

Information asymmetry exists within hierarchies in organisations. The managers of different ranks possess extra information about employees and their performance. How this information shapes the promotion and hiring decisions of the lower-ranked managers is essential for shareholders. In such environment reputation for diversity and non-discrimination becomes very important both for the managers and shareholders. To see this, let us look at an example. 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 induced to promote diversity and counter discrimination. Consider a firm who wants to improve diversity. Such a 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. Strategies of such manager 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, such a 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 both the short run and the long 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 of which only the skin colour (or any minority groups' affiliation) is observable. We assume the 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.

We get three key results; firstly for a discriminator with a high degree of aversion from 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 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 discriminator. If the discriminator wants to build a reputation, he will employ more blacks, but by sometime not helping, 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.

Secondly, we find "Diversity Paradox". 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 toward diversity is the main derive of sabotage.

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

Many studies support our result. A good example is The Female FTSEBoard Report. The report has been monitoring trends in women's representation as executive directors on the corporate boards of the UK's top 100 companies from 1999. In the most recent report, the percentage of women representation on the corporate boards was close to the target set, but the report identifies their representation as a "tick box" attitude. The reason is that the report shows on average women are less likely to get promoted than their male counterpart, and their average tenure is half men. They identify the improvement in the percentage as just a tick box attitude toward women or putting it in

the context of our paper, its mostly to gain reputation. The lower tenure length 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 harassment. If we assume harassment as a weak measure of sabotage, this is in line with our result. Small productivity jobs are more prone to sabotage because even slight aversion from minorities will make sabotage optimal.

## 2 Literature Review

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

The most related strand of literature to this work is the literature on discrimination. The literature is divided into three categories: taste-based discrimination, statistical discrimination and invisibility hypothesis. Central to taste base discrimination starting from Becker (1971) is the assumption that some employers dislike members of minority groups. [1]. 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). Milgrom and Oster (1987) suggest that the main reason for discrimination is the fact that the members of minority groups are less observable by the employees. In most of the literature, discrimination is modelled at the market level and how the potentially present discriminatory attitudes 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 term of implication, the closest to our paper is Coate and Loury (1993). However, our primary focus is 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 for discrimination. They show that in the presence of such

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

expectation even if the principal has no bias toward minorities, he will discriminate against them to avoid demotivation of 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 managers private information, he makes a promotion decision. She 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 with it 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 through some time sabotaging the black worker. To the best of our knowledge 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 through sabotaging them can gain more reputation.

The main reason for sabotage to play such a role in reputation building is the uninformative monitoring of the principal when faced with a failure of the 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) establishes 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 the bad reputation 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 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 not only the type of the agent is uncertain, but also the monitoring mechanism is uncertain. In their work Deb and Ishii (2018) 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 for 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, that is because both the employment choice and effort choice needs 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 introduce sabotage in reputation games without competition. To the best of our knowledge, such structure of monitoring 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 his seminal work discusses sabotage in tournaments. He uses relative performance evaluation when production is interrelated between co-workers. He shows 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 lobbyist improves their chances of success by using their resource 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 to renegotiate contracts and the incentive of 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 to 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 concerns. Nonetheless, in our work, there is no competition or relative evaluation. The manager sabotages the employee (inflicts failure to 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 (2001) uses this setting in a two-period repeated game between a potentially biased expert and 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 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 in need to build a reputation, one has to choose the not optimal action. However, the main point of departure from this literature is that our framework goes beyond communication and inflict 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 3, we present the model. In section 4, we analyse step by step the equilibrium of a three-period game and illustrate the sabotage equilibrium.

# 3 Model

## 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 $\beta$ 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

$$U_t^P = E(X_t) + m_t \beta \tag{1}$$

The manager at each period receives

$$U_t^\theta = \nu\big(E(X_t) + m_t \theta\big) \tag{2}$$

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.

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

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.

$$\mathcal{V}_{f=0}^{P} = \sum_{s=t+1}^{3} \mathbb{E}(U_s^{p}) \tag{4}$$

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

$$\mathcal{V}_{f=1}^{P} = \sum_{s=t+1}^{3} C \tag{5}$$

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:

$$\mathcal{V}_t^{p} = \max_{f_t \in \{0,1\}} \sum_{s=t+1}^{3} \mathbb{E}(U_s^{P}) \tag{6}$$

If the manager is fired, he gets an outside option of $\mathcal{V}_f^{\theta} = -D$. 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 $\pi_0$ 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.

## 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.

# 4 Reputation building and Sabotage

## 4.1 Preliminaries

The repeated game between the principal and the manager is one of the finitely repeated reputation games.[2] The $\delta$ 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 $\pi_{t-1}$, 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 manager$f_{t-1}$. 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 $\delta$ 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

[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 $\delta$ type manager will choose $m_t$ and $e_t$ to avoid reaching that event. In the equilibrium, the realisation would result in firing the $\beta$ type manager, which is not optimal for the principal.

Only for certain specifications of $D$ and $\beta$ 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.

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 $\pi_{t-1}$ 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 $\delta$ 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 $\pi_t$.

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 $t$, $\pi_t$, the strategies are best response to one another and the belief $\pi_t$, 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 $\beta$ 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.

## 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.

### 4.2.1 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 $\beta$ 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 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 1** ***Monotonicity Condition:*** *For all biases of the manager $\delta$, 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 $\pi_{m=1}^s < \pi_{m=1}^f$*
2. *The lowest when $m = 0$ and the employee fails $X = 0$. That is $\pi_{m=0}^s > \pi_{m=0}^f$*

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 $\beta$ type thane when he is $\delta$ type, that is the condition in point 1 of lemma 1 holds if:

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

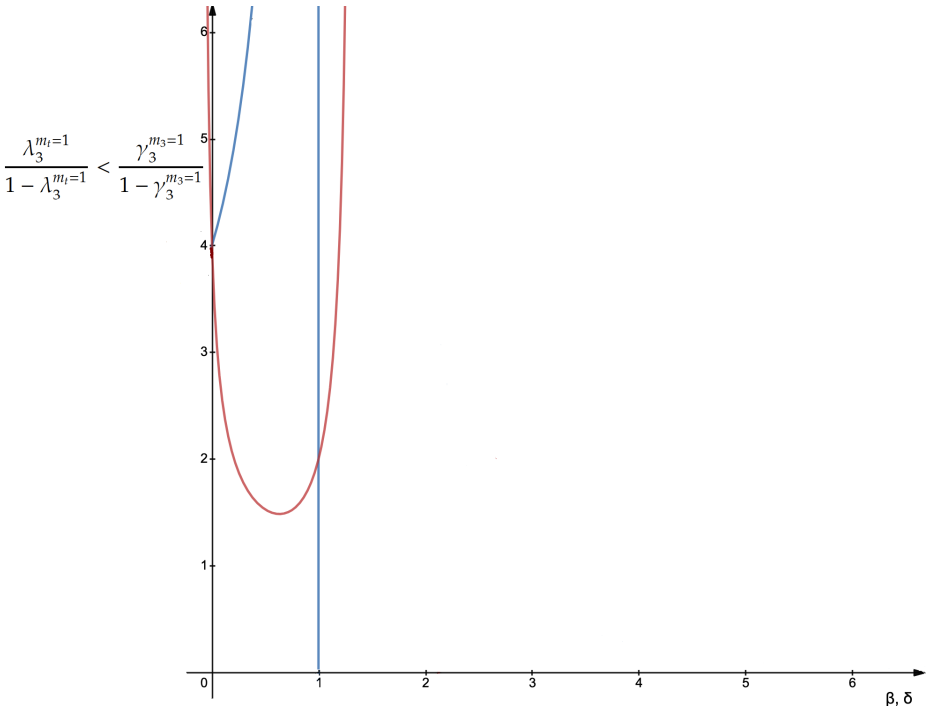


Figure 1: Belief Monotonicity $m_t = 1$

Figure 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 $\delta$ type manager when $m_t = 1$. For point two of the lemma 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 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 $\delta$ type manager when $m_t = 0$.

Lemma 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** *If the manager makes the employment and effort choice without career concern, since the*

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

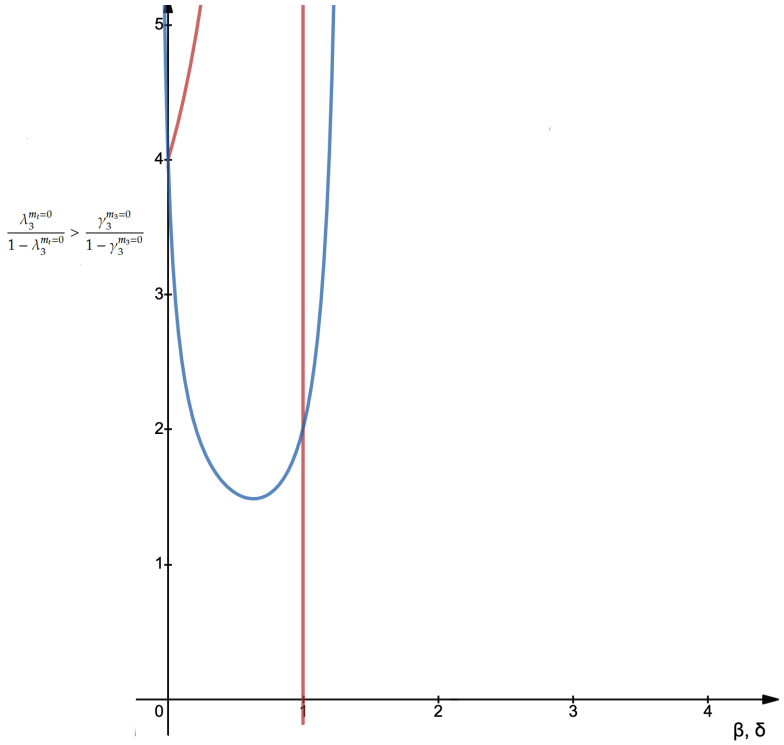


Figure 2: Belief Monotonicity $m_t = 0$

*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 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 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.

### 4.2.2 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 $\pi_1$ to $\pi_2$.

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

*We define $\pi^*$ as the belief at which if $m_t = 0$ and $X_t = 0$ is observed the principals belief is updated to $\underline{\pi}$:*

$$\underline{\pi} = \frac{\pi^*[pr(\sqrt{a_{m=0}} \geq \sqrt{a_{m=1}} + \beta)(1 - \lambda_t^{m=0})}{\pi^*[pr(\sqrt{a_{m=0}} \geq \sqrt{a_{m=1}} + \beta)(1 - \lambda_t^{m=0}) + (1 - \pi^*)[pr(\sqrt{a_{m=0}} \geq \sqrt{a_{m=1}} - \delta)(1 - \gamma_t^{m=0})]} \quad (7)$$

Given the strategy of the principal defined in Lemma 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 $\beta$ 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** *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, 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, 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 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.*

*$\kappa$ is decreasing in $\pi_0$. That is if $\pi_0$ is close to $\pi^*$, $\kappa \to 0$. But if $\pi_0$ 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 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.

### 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 1 and Lemma 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 4** ***Optimality Condition:*** *For every $\nu$, 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 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 1** *If the payoff of the project was $\nu$ for the principal too. For low productivity project's sabotage will always be optimal*

Corollary 1, shows that if the productivity of the projects are low, $\delta^*$ will be very low and sabotage would be optimal more often. This follows from Lemma 4, since the condition specified there is independent of the payoff of the project for the principal.

Once sabotaging becomes an optimal strategy (high $\delta$), 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 1 further specified the belief updating structure.

Recall Lemma 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 $\beta$. However, if the threshold is $\beta$, 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 $\delta$ 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 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 5, specifies condition under which deviation from sabotage is not optimal. If the above condition was not in place, given lemma 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 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 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 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*

$$q_2^{*sab} = \frac{\omega D + 1 + \mathcal{V}_{mix} - (2-\omega)\mathcal{V}_{sab}}{\omega(\mathcal{V}_{sab} + D)} \tag{8}$$

*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*

$$\eta^*_{m_1=1} = \frac{(\beta(\frac{4}{3} - \beta) + \frac{(1-\beta)^4(3+2\beta)}{30})\pi_0(1-\pi*) - [\frac{1}{3}\alpha^2(1-\pi_0)\pi*]}{\frac{2}{3}\alpha^2(1-\pi_0)\pi*} \tag{9}$$

*Where* $\alpha$ *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* $\pi_0$ *not too low and not too high*

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

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

Corollary 2, follows from Proposition 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 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 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 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 $X_1 = 0$.

# 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.

# References


Arrow, K. J. (1973). *The theory of discrimination* (O. Ashenfelter & A. Rees, Eds.). Princeton, NJ: Princeton University press.

Auriol, E., & Guido, F. (2000, August). Career concerns in teams. *working papaer*.

Bar-Issac, H., & Deb, J. (2018, April). Reputation with opportunities for coasting. *working papaer*.

Becker, G. S. (1971). *The economics of discrimination* (2nd ed.). Chicago, IL: Chicago University Press.

Bénabou, R., Falk, A., & Tirole, J. (2019, September). *Narratives, imaperatives and moral persuasion* (Tech. Rep.).

Bénabou, R., & Laroque, G. (1992). Using privileged information to manipulate markets: Insiders, gurus, and credibility. *The Quarterly Journal of Economics*, *107*(3), 921–948.

Bénabou, R., & Tirole, J. (2011). Identity, morals and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, *126*(2), 805–855.

Chalioti, E. (2019). Incentives to help or sabotage co-workers. *Working paper*.

Coate, S., & Loury, G. (1993). Antidiscrimination enforcement and the problem of patronization. *The American Economic Review, Papers and proceedings*, *83*(2), 92–98.

Cripps, M. W., Mailath, G. J., & Samuelson, L. (2004). Imperfect monitoring and impermanent reputations. *Econometrica*, *72*(2), 407–432.

Deb, J., & Ishii, Y. (2018). Reputation building under uncertain monitoring. *Working Paper*.

Diamond, D. W. (1991, August). Monitoring and reputation: The choice between bank loans and directly placed debt. *The Journal of Political Economy*, *99*(4), 689–721.

Ely, J. C., Fudenberg, D., & Levine, D. K. (2008). When is reputation bad? *Games and Economic Behavior*, *63*(2), 498–526.

Ely, J. C., & Välimäki, J. (2003). Bad reputaion. *The Quarterly Journal of Economics*, *118*(3), 785–814.

Fudenberg, D., & Levine, D. K. (1992, July). Maintaining a reputation when strategies are imperfectly observed. *Review of Economic Studies*, *59*(3), 561–579.

Gentzkow, M., & Shapiro, J. M. (2006). Media bias and reputation. *The Journal of Political Economy*, *114*(2), 280–316.

Gossner, O. (2011). Simple bounds on the value of a reputation. *Econometrica*, *79*(5), 1627–1641.

Halac, M., & Prat, A. (2016). Managerial attention and worker performance. *American Economic Review*, *106*(10), 3104–3132.

Holmstorm, B. (1999). Managerial incentive problems: A dynamic perspective. *The Review of Economic Studies*, *66*(1), 169–182.

Kamphorst, J. J. A., & Swank, O. H. (2016). Don't demotivate, discriminate. *American Economic Journal: Microeconomics*, *8*(1).

Konrad, K. A. (2000). Sabotage in rent-seeking contests. *Journal of Law, Economics, & Organization,*, *16*(1), 155–165.

Lang, K., & Lehmann, J.-Y. K. (2012). Racial discrimination in the labor market: Theory and empirics. *Journal of Economic Literature*, *50*(4), 959–1006.

Lazear, E. P., & Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *The Journal of Political Economy*, *89*(5), 841–864.

Milgrom, P., & Oster, S. (1987). Job discrimination, market forces, and the invisibility hypothesis. *The Quarterly Journal of Economics*, *102*(3), 453–476.

Morris, S. (2001, April). Political correctness. *Journal of Political Economy*, *109*(2), 231–265.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The American Economic Review*, *62*(4), 659–661.

Shin, W. (2016). Discrimination in organizations: Optimal contracts and regulation. *Working Paper*.

# 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 $\beta$ 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.

$\lambda_3^{m_3=1}=$

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

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

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

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

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

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

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

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 $\lambda_3^{m_3=0}$=

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

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

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

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

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

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

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

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

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

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

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

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 $\gamma_3^{m_3=0}=$

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

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

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

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

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

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

**Proof.** Given Lemma 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 $\beta$ and $\delta$, 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 $\beta$ and $\delta$ 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 $\delta$ and $\beta$ are both equal to 1 the argument above always hold.

■

## Proof of Lemma 3

**Proof.** From Proposition 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 and Lemma 1 and the fact that the $\delta$ type manager is less likely to set $m_t = 1$, we can infer that the argument above is increasing in $\pi_2$. 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 $\pi_2$, $\delta$ and $\beta$. 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 $\pi_0$, this condition is always satisfied. ■

## Proof of Proposition 2

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

From Lemma 1 and Lemma 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 $\beta$ 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 3

**Proof.** Suppose$\underline{\pi} < \pi_0 < \pi*$, first it can be verified that the strategy of the principal in Proposition 2, can not be sustained in the equilibrium.

Firstly fully mimicking of the $\beta$ 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 $\sqrt{a_1} - \delta = \sqrt{a_0}$, $\kappa$ can be lower than $\delta$ for $\pi$ 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:

   $\gamma_2^{m_2=1}=$

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

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

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

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

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

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 $\gamma_2^{m_2=0}$=

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

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

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

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

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

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

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 $\kappa > \delta$

$\gamma_2^{m_2=1}=$

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

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

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

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

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

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

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

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

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

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 $\gamma_2^{m_2=0}=$

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

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

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

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

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

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

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 $\kappa$ 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 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 $\mathcal{V}^{sab}$:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

■

## Proof of Lemma 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 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 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 $\beta$ 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 1 and Lemma 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

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

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

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

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

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

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 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 5 and Lemma 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 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 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 $\pi_0$ 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 5 for probability of success given sabotage specifies a lower bound for $\pi_0$. 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 $\pi_0$ 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 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

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

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

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

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 $\alpha$, $\gamma_{m_1=0}=\frac{\frac{2}{3}(1-\alpha^3)}{1-\alpha^2}$ 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 0.A.1 plots $\gamma_{m_1=0}$ and $\lambda_{m_1=0}$, and

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

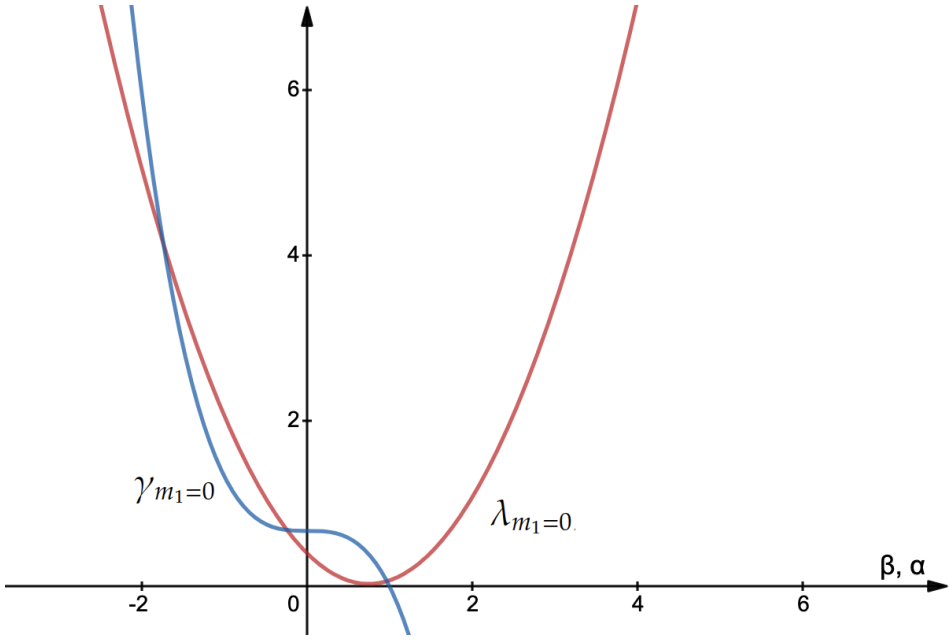


Figure 0.A.1: $\gamma_{m_0=1} > \lambda_{m_0=1}$

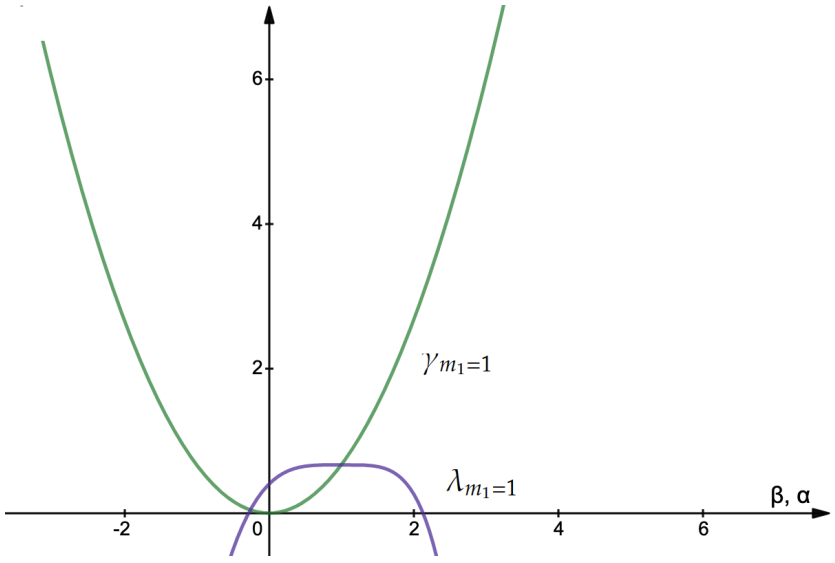


Figure 0.A.2: Lemma 2 with sabotage $\gamma_{m_1=1} < \lambda_{m_1=1}$

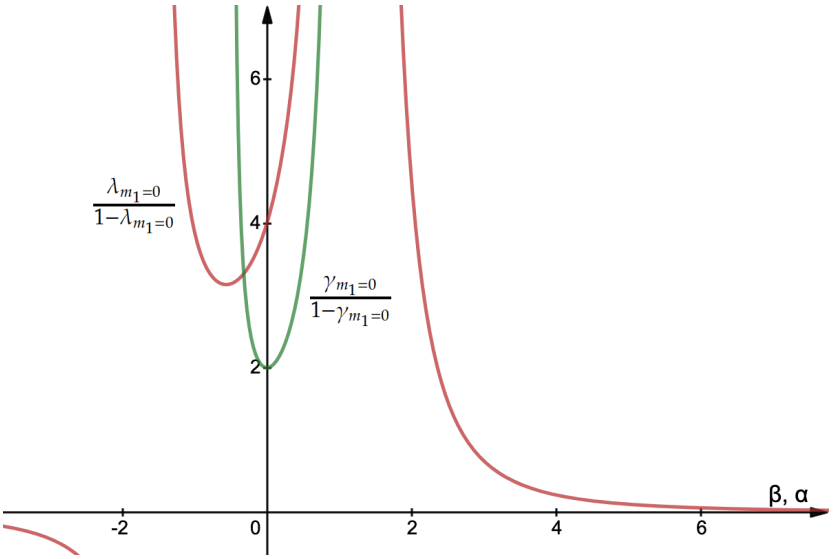


Figure 0.A.3: $\frac{\lambda_{m_1=0}}{1-\lambda_{m_1=0}} > \frac{\gamma_{m_1=0}}{1-\gamma_{m_1=0}}$

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

The two graph show that the condition in Lemma 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 $\delta$, $\alpha$ will be small.

2. The condition in Lemma 1 is also satisfied since

   $1 - \gamma_{m_1=0} = \frac{1}{3} - \alpha^2 + \frac{2}{3}\alpha^3$, Figure 0.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 $\alpha$, that is when $\delta$ is big enough.

It remains to check if given the new threshold $pr(m_1 = 1|\theta = \delta)$, the conditions in Lemma 1 and Lemma 2 are satisfied. In the previous section Lemma 2 was shown to be satisfied so it remain to check Lemma 1, since $1 - \gamma_{m_1=1} = \frac{1}{3}$ then it 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 0.A.4, shows that for all $\alpha$ if $\beta$ 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 $\delta$ is high.

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

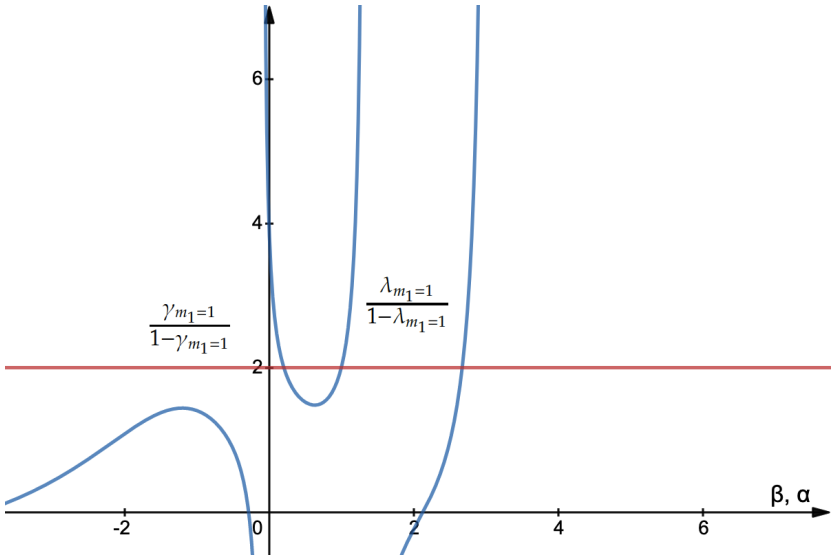


Figure 0.A.4: $\frac{\lambda_{m_1=1}}{1-\lambda_{m_1=1}} < \frac{\gamma_{m_1=1}}{1-\gamma_{m_1=1}}$

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

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

■