Knowing the Unknown: Experimentation under Delayed Success *

Shraman Banerjee † Soumik Kumar Saha ‡

November 26, 2024

Abstract

We consider a continuous-time dynamic principal-agent model with experimentation under a mixed-news framework. In an exponential two-armed bandit framework, the agent can achieve success and failure from both known and unknown risky arms. We frame success to be conclusive i.e. a single success truly reveals its type and inconclusive breakdowns for the unknown arm. Although success arrives at a lower rate in the unknown risky arm than the known arm, it generates higher revenue for the experimenting agent. To motivate the agent to experiment in the unknown arm in spite of delayed success, the principal proposes a mechanism which involves rewarding failures and punishing fake successes, in order to incentivize the agent to experiment in the unknown risky arm. We derive the optimal deadline for experimentation which ensures full honesty from the agent.

JEL Classification: D82, D83, D86, O32

Keywords: Dynamic moral hazard, Continuous-time principal-agent model, Experimentation, Bandit Models, Differential-Difference equation.

*An earlier version was circulated under the title "Knowing the Unknown: Experimentation in Low Rewarding Projects". We are extremely grateful to Kaustav Das, Nicolas Klein and Sven Rady for their many detailed and helpful comments and suggestions. We are thankful to the seminar participants at Shiv Nadar University for their helpful comments and feedback.

†Shiv Nadar University, Uttar Pradesh 201314, India Email: shraman.banerjee14@gmail.com;
‡Shiv Nadar University, Uttar Pradesh 201314, India Email: ss212@snu.edu.in;

1. Introduction

Adopting a new experimental approach can lead to greater initial losses from failures and may result in significant delays before achieving success. However, despite these setbacks, adopting such technology can ultimately provide stakeholders with greater long-term benefits.

Therefore, it is crucial to incentivize researchers and experimental agencies to avoid reverting to safer options. Many firms implement various incentive schemes to promote such exploration. For example, Google X, Procter & Gamble, and Tata allow employees the freedom to pursue low-yield projects and even reward failures. The European Research Council (ERC) also funds groundbreaking research without expecting immediate technological returns[1].

In this paper, we develop a continuous-time strategic experimentation model in a two-armed bandit setup. Arm 1 represents the technology or research path under investigation—the new and unknown avenue for research or technology development. In contrast, Arm 0 represents the safer option, the existing and known research paths. Both arms generate lump-sum rewards and costs at the jump times of the Poisson process.

We consider a mixed news framework integrating the conclusive success as in Keller et al.(2005) and the inconclusive failure in Keller and Rady (2015). Arm 0 generates success (good news) and failure (bad news) with a higher success rate. Arm 1 is unexplored. It can be either good or bad and generates both success (with a lower rate than Arm 0) and failure if it is good and can generate failure (with a higher intensity than if the arm is good) if it is bad.

A single success on Arm 1 makes the belief jump to full certainty and conclusively reveals its type, irrespective of the belief held before. In addition, the arrival of failure and the absence of news make the agent more pessimistic (with a discrete downward jump in case of failure). To simplify our analysis, we restrict ourselves to Markov strategies, with the posterior belief as a state variable. Given the discrete downward adjustment of beliefs, payoff functions solve first-order ordinary differential-difference equations (ODDEs) as in Keller and Rady(2010,2015).

In a principal-agent framework, we demonstrate how the principal can motivate the agent (e.g., a researcher) to experiment with an unknown project (Arm 1). We propose a mechanism that rewards failures and punishes fake successes. We derive the optimal mechanism, which encompasses a reward and punishment scheme along with an optimal deadline, as part of the principal's wage minimization problem.

The agent has one unit of resource to allocate between the two projects. The

[1]According to the data, over 44% of ERC-funded projects have generated research that was subsequently cited by patent applications.

arrival rate of news—whether good or bad—is assumed to be proportional to the resources allocated to each project. The agent receives a reward (or incurs a cost) whenever a good (or bad) signal is realized. While the principal cannot observe the agent's choice of projects (or arms), she can monitor the timing of when good and bad signals occur. The lower success rate and, consequently, the lesser rewards associated with the unexplored project may lead the agent to cheat by switching to the known project, which offers a higher success rate.

Thus to prevent the agent from cheating and using Arm 0, the principal designs an incentive scheme. First, he optimally chooses a deadline for experimentation. At the deadline, the principal collects all good and bad news reported by the agent till that time. While choosing an exogenous threshold (convex combination of the ratio of failure rate to success rate from Arm 0 and Arm 1, see page 18) as a parameter of honesty, the principal calculates the ratio of bad news and good news the agent has reported. Now if the calculated number exceeds the chosen threshold, then the principal compensates the agent for his reported bad news. Note that a relatively higher number of bad news is more likely to occur in Arm 1 rather than Arm 0 since the arrival rate of bad news relative to good news is higher in Arm 1 than in Arm 0. But in case, the calculated ratio does not exceed the chosen threshold, the principal punishes the agent for his reported good news.

As a benchmark, we determine the agent's behavior in the absence of the principal's incentive scheme. Proposition 1 shows that without any incentives, there exists some threshold belief below which the agent cheats by switching to the known project.

In Proposition 2 and Proposition 3, we show that after the implementation of the incentive scheme, there still exists some threshold belief below which the agent switches to the safer option, but this threshold is lower than the former one. The incentive scheme now motivates the agent to spend more time on the unexplored project. Anticipating the agent's behavior, the principal optimally chooses a deadline along with minimizing wage payments that ensures the full honesty of the agent. Proposition 4 shows the optimal stopping time when the principal stops the experimentation and pays the agent depending on its outcome. In Proposition 5 upon setting an optimal deadline, we focus on the arm identification by the principal. Since a single success reveals the unknown state of the project to be good, the principal declares the unknown arm (Arm 1) to be good if the sequence submitted by the agent contains at least one good piece of news. On the contrary, upon receiving no good news till the optimal deadline, the principal identifies the unknown arm (Arm 1) to be bad with a higher probability under the incentive scheme compared to the situation when there is no incentive.

The contribution of our paper is twofold. First, contrary to Klein (2016) and

Hidir (2019), we frame our model with the hybrid news framework by integrating the conclusive success of Keller et al. (2005) and the inconclusive failure introduced by Keller, Rady (2015), i.e., our bandit arms can produce good news and bad news both[2]. Secondly, because the action of the agent is not observable by the principal, hence it is difficult to identify whether reported successes and failures are obtained through cheating or honest experimentation. Thus we propose a novel mechanism by the principal, which involves rewarding occasional failures, and punishing occasional successes, to incentivize the agents to experiment in the unknown Arm 1.

Related Literature: Our work contributes to the ongoing literature on the Principal-agent problem with bandit games, intending to maintain the agent's truthfulness. In this regard, our work is closest to Klein (2016). Considering three-armed bandit Klein (2016) demonstrates how a principal motivates an agent to experiment even when he has the option to cheat. It considers a "good news" model where the agent has two options to work along with a safe option; one arm (Arm 0) is known (where the arrival of success is lower) and another arm (Arm 1) is unknown, and where the arrival rate of success is high if it is good and no success otherwise. Being a "good news" model, it provides an optimal mechanism to be depending on the number of good news. In contrast, we consider a hybrid news (success & failure) framework by integrating conclusive success proposed by Keller et al. (2005) and inconclusive failure introduced by Keller, and Rady (2015). The known risky arm (Arm 0) offers a higher rate of success than failure. On the other hand, the unknown arm (Arm 1), with delayed success, generates a higher rate of failure than success if it is good and no success if it is bad. Thus, it is difficult to verify the agent's honesty by looking into the reported news. Hence, in contrast to Klein (2016), our paper derives the optimal mechanism which rewards honest experimentation and punishes fake successes, based on the ratio of failures and successes.

In terms of the specification of the unknown arm (Arm 1), our paper also relates to Hidir (2019). It considers the agent to have an outside option to shirk (we call it a safe option) along with experimenting with the unknown arm with hybrid news (produces success and failure if the arm is good and generates failure otherwise). In contrast to the framework by Hidir(2019), our model considers one arm to be a known risky arm and another arm to be an unknown risky arm that can be of good quality or bad quality respectively, and like Hidir(2019) we have not considered safe arm to our model. Therefore, the principal is unable to confirm if the agent is telling

[2]The feature that good risky arm can produce bad news is shown in Keller and Rady (2015)

the truth or not. Our paper focuses on this particular question.

In terms of motivation, our paper is somewhat related to Kuvalekar and Ravi (2019). It reports the dynamic principal-agent problem and establishes an optimal incentive scheme to incentivize the agents by rewarding them even if they receive breakdowns in their experimentation. In a similar line, our paper extends our optimal incentive scheme to reward occasional failures and punish fake successes.

In the strategic experimentation literature, several papers consider incentivizing agents in a dynamic principal-agent framework. For example, Bergemann and Hege (1998, 2005) consider the dynamic principal-agent framework where a venture capitalist (the principal) looks for optimal financing for an unknown project conducted by the entrepreneur (the agent) in the context of moral hazard. Horner and Samuelson (2013) also analyze optimal incentive schemes in a related framework. Halac, Kartik, and Liu (2016) consider optimal contracts in the presence of moral hazard and adverse selection in a similar exponential bandit framework. Guo (2016) considers a similar dynamic principle-agent framework where the principal delegates the agent for experimentation.

The paper progresses as follows. Section 2 describes the model; Section 3 derives the agent's problem and proceeds with two sub-sections; Subsection 3.1 looks into the agent's problem without the incentive scheme, whereas Subsection 3.2 focuses on the agent's problem with the incentive scheme; Section 4 considers the principal's problem and consists of two subsections; Subsection 4.1 formally derives the optimal deadline, whereas Subsection 4.2 looks into the principal's problem of identification of unknown arm (Arm 1) and Section 5 concludes the paper. Appendix A includes the proofs of the agent's problem and mathematical details are given in Appendix B.

2. Model

We frame our principal-agent problem in a continuous time ($t \in [0, \infty)$) two-armed exponential bandit framework. One arm ("**Arm 0**") is known to generate success as well as failure whenever played. Another arm ("**Arm 1**") is unknown involving uncertainty and can either be good or bad. If the arm is good, then it produces good news as well as bad. It never generates success if the arm is bad. Each arm generates a lump-sum payoff (Reward or Punishment) whenever pulled.

Arm Specification:

Arm 0 produces lump-sum rewards that arrive according to a Poisson process with the parameter $\gamma_g > 0$. These lump-sum rewards are drawn from a time-invariant

distribution on $\mathbb{R}_{++}$ with mean s_g; therefore it is equivalent to a constant flow payoff of $R_0 = \gamma_g s_g$. In addition to, Arm 0 produces lump-sum costs that arrive according to a Poisson process with the parameter $\gamma_b > 0$. These lump-sum costs are drawn from a time-invariant distribution on $\mathbb{R}^-$ with mean s_b; therefore it is equivalent to a constant flow payoff of $C_0 = \gamma_b s_b$.

Similar to Arm 0, Arm 1 produces lump-sum rewards that arrive according to a Poisson process with the parameter $\lambda_g^G > 0$ (if the time-invariant state of the world $\theta = G$ i.e. Good). These lump-sum rewards are drawn from a time-invariant distribution on $\mathbb{R}_{++}$ with mean h_g; therefore it is equivalent to a constant flow payoff of $R_1 = \lambda_g^G h_g$. Also, Arm 1 produces lump-sum costs that arrive according to a Poisson process with the parameter $\lambda_b^G > 0$ (if the time-invariant state of the world $\theta = G$ i.e. Good) and $\lambda_b^B > 0$ respectively (if the state is $\theta = B$ i.e. Bad). These lump-sum costs are drawn from a time-invariant distribution on $\mathbb{R}^-$ with mean h_b. Table 1 can be referred to for a more concise representation of the arm specifications.

Assumption 1. *We assume that good news arrives at a lower frequency in Arm 1 compared to Arm 0 i.e.* $\gamma_g > \lambda_g^G$ *but generates a higher lump-sum payoff than Arm 0 i.e.* $h_g > s_g$. *In addition, the net flow payoff from Arm 1 is higher than Arm 0, i.e.* $R_1 > R_0$.

Assumption 2. *Arm 1 involves a high failure rate, i.e. even if it is good, breakdown arrives at a higher rate relative to Arm 0, i.e.* $\lambda_b^B > \lambda_b^G > \gamma_b$. *In addition, the absolute value of cost incurred from experimenting with Arm 1 is higher than Arm 0, i.e.* $|h_b| > |s_b|$.

Assumption 3. *Arm 0 produces good news at a higher rate than bad news i.e.* $\gamma_g > \gamma_b$.

Assumption 4. *We assume that good news arrives at a higher rate than bad news if the underlying state is "Good" i.e* $\lambda_g^G > \lambda_b^G$. *Additionally, the Arrival rate of bad news from Arm 1 in a "Good" state is bounded above by some threshold* $\bar{\lambda}_b^G$ *to be determined. (See Lemma 1)*

Assumption 5. *We further assume that any news arrives at a higher rate if the state is "good" than if it is "bad" i.e.* $\lambda_g^G + \lambda_b^G > \lambda_b^B$

Payoffs:

The agent is endowed with one unit of perfectly divisible resource per unit of time. If $k_{1,t}$ denotes the fraction of the agent's resources that he devotes to Arm 1 at instant t, then the fraction $(1 - k_{1,t})$ is allocated to Arm 0.

Arm Specification with Poisson process					
Arm	Arm Quality	Types of News	Intensity	Reward	Cost
Arm 0 (Known)	Good	Good News	γ_g	$s_g > 0$	$\times$
		Bad News	γ_b	$\times$	$s_b < 0$
Arm 1 (Unknown)	If Good	Good News	λ_g^G	$h_g > 0$	$\times$
		Bad News	λ_b^G	$\times$	$h_b < 0$
	If Bad	Bad News	λ_b^B	$\times$	

Table 1: Arm Specification

If the agent allocates the fraction $(1 - k_{1,t})$ of his resources to the Arm 0 over an interval of time $[t, t + dt)$, the probability of a breakthrough at some point in the interval is $(1 - k_{1,t})\gamma_g dt$ and the probability of a breakdown at some point in the interval is $(1 - k_{1,t})\gamma_b dt$. Players start with a common prior belief p_0 about the unknown state of the world. Let, p_t be the subjective probability that the agent assigned to Arm 1 being good at time t. The fraction $k_{1,t}$ allocated to Arm 1 produces a breakthrough at some point in the interval with probability $p_t k_{1,t} \lambda_g^G dt$ and causes a breakdown at some point in the interval with probability $k_{1,t}\lambda(p_t)dt$ with

$$\lambda(p_t) = p_t \lambda_b^G + (1 - p_t)\lambda_b^B.$$

Given the agent's action $\{k_{1,t}\}_{t\geq 0}$ the agent's expected experimentation outcome conditional on all available information in time dt is

$$\Big[(1 - k_{1,t})(R_0 + C_0) + k_{1,t} p_t R_1 + k_{1,t}\lambda(p_t)h_b\Big]dt$$

Now, the agent's expected discounted payoff, expressed in per-period units, is

$$\mathbb{E}\left[\int_0^T re^{-rt}\Big[(1 - k_{1,t})(R_0 + C_0) + k_{1,t} p_t R_1 + k_{1,t}\lambda(p_t)h_b\Big]dt\right]$$

where the expectation is now over the stochastic processes $\{k_t\}$ and $\{p_t\}$.
Starting with a prior belief p_0 that the Arm 1 is good, her overall objective is to choose a strategy $\{k_{1,t}\}_{t\geq 0}$ that maximizes

$$\mathbb{E}\left[\int_0^T re^{-rt}\Big[(1 - k_{1,t})(R_0 + C_0) + k_{1,t} p_t R_1 + k_{1,t}\lambda(p_t)h_b\Big]dt\right]$$

which expresses the total payoff in per-period terms.

Evolution of Beliefs:

To derive the law of motion of beliefs, suppose that over the interval of time $[t, t+dt)$ the agent allocates the fraction $k_{1,t}$ of the unit resource to Arm 1. If Arm 1 is good, the probability of having a breakdown is $\lambda_b^G k_{1,t} dt$ [3]; if Arm 1 is bad, the probability of having a breakdown is $\lambda_b^B k_{1,t} dt$. When the agent starts with the common belief p_t and achieves a breakdown in $[t, t+dt)$, therefore, the updated belief at the end of that period is (by Bayes' rule)

$$\begin{aligned} j(p_{t-}) &= \frac{p_{t-}\lambda_b^G k_{1,t} dt}{p_{t-}\lambda_b^G k_{1,t} dt + (1-p_{t-})\lambda_b^B k_{1,t} dt} \\ &= \frac{\lambda_b^G p_{t-}}{\lambda_b^G p_{t-} + (1-p_{t-})\lambda_b^B} < p_{t-} \end{aligned}$$

As the frequency of arrival of a breakdown is higher when the underlying state is bad than when the state is good, the agent postulates that the breakdown may arrived from a bad quality of Arm 1. Hence, intuitively, in the presence of a breakdown, the agent's posterior belief follows a downward jump according to Bayes' rule. Once there is a breakthrough, of course, the posterior belief jumps to 1. Belief is also updated when the agent observes no news from Arm 1. Now, If Arm 1 is good, the probability of having no news is $(1 - k_{1,t}\lambda_g^G dt - k_{1,t}\lambda_b^G dt)$ and if Arm 1 is bad, the probability of having no news is $(1 - k_{1,t}\lambda_b^B dt)$. When the agent starts with the common belief p_t and finds no news in $[t, t+dt)$, therefore, the updated belief at the end of that period is (by Bayes' rule)

$$p_t + dp_t = \frac{p_t(1 - k_{1,t}\lambda_g^G dt - k_{1,t}\lambda_b^G dt)}{p_t(1 - k_{1,t}\lambda_g^G dt - k_{1,t}\lambda_b^G dt) + (1-p_t)(1 - k_{1,t}\lambda_b^B dt)}$$

Simplifying, we see that after having no news, the belief changes by

$$dp_t = k_{1,t} p_t (1-p_t)(\lambda_b^B - \lambda_b^G - \lambda_g^G) dt = -ve$$

Since in the good state, the news arrives at a higher rate than in the bad state ($\lambda_g^G + \lambda_b^G > \lambda_b^B$, by Assumption 5), the absence of any news makes the agent pessimistic and belief falls downwards.

[3] This is up to the terms of order $o(dt)$, which we can ignore here and in what follows

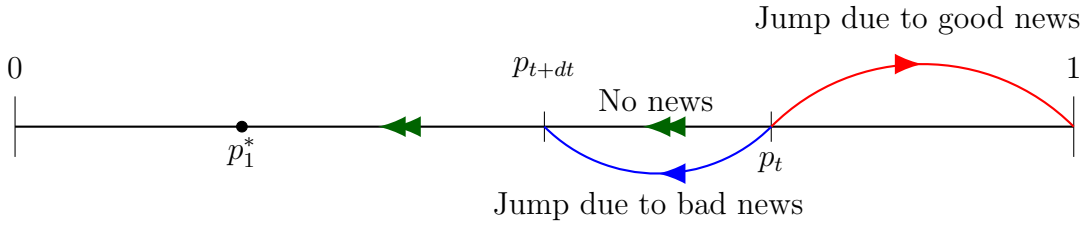

Figure 1: Belief Updation

Principal's Objective:

The principal obtains a benefit of R_P from the agent's experimentation on Arm 1. This consists of the current experimentation payoff generated from the agent's effort on Arm 1 and the future value expected by the principal upon the revelation of the true state of the unknown world (Arm 1).

We assume, $R_P > R_1 > R_0$. Hence the principal receives a higher value from Arm 1 as compared to the agent, the agent tends to offer lower effort than he does when the principal incentivizes him to do so. Having a lucrative option of grabbing higher benefits, the principal is more interested in involving the agent in experimenting on Arm 1 and accessing information regarding success and failures obtained through Arm 1.

The principal monitors all successes and failures and the time they arrive but is unaware at which arm these events have been achieved. Additionally, the agent knows the arm he pulls and generates success and failure. More formally, define the point process $\{(N_g)_t\}_{0\leq t\leq \bar{T}}$ and $\{(N_b)_t\}_{0\leq t\leq \bar{T}}$ by the number of success and failure obtained till the time t respectively. Additionally, the point process $\{N_t\}_{0\leq t\leq \bar{T}}$ is defined by $N_t := (N_g)_t + (N_b)_t$ for all t. Also, we define $\{(N_g)_t\}_{0\leq t\leq \bar{T}}$ and $\{(N_b)_t\}_{0\leq t\leq \bar{T}}$ by $(N_g)_t := (N_g^1)_t + (N_g^0)_t$ and $(N_b)_t := (N_b^1)_t + (N_b^0)_t$ where $(N_g^i)_t$ and $(N_b^i)_t$ represents the number of successes and failures arrived on arm i till and including the time t respectively. Furthermore, let $\xi{:=}\{\xi_t\}_{0\leq t\leq \bar{T}}$ and $\{\xi_t^N\}_{0\leq t\leq \bar{T}}$ represents the filtrations generated by the process $\{((N_g^1)_t, (N_g^0)_t), ((N_b^1)_t, (N_b^0)_t)\}_{0\leq t\leq \bar{T}}$ and $\{((N_g)_t, (N_b)_t)\}_{0\leq t\leq \bar{T}}$, respectively. The former captures the idea that, at any given time, the agent knows when all past successes and failures have been achieved and can pinpoint the arms he has operated. The latter contains less information, in other words, the principal knows the timing of all successes and failures but can not comment about the arm where they arrived.

To maintain the objective of knowing the type of Arm 1, the principal prevents the agent from switching to a known option by implementing an incentive scheme of "rewarding the bads" and "punishing the goods". (For details please refer to Subsection 3.2). Now, therefore, the principal's objective function includes the principal's

benefit and payment to the agent according to the experimentation outcome. Hence, we formalize the objective function as follows,

$$V_P(T) = R_P + \text{Wage payment}$$

In our specification, the principal's benefit R_P is independent of the time, hence to maximize $V_P(T)$, the principal should optimize T by setting up an appropriate deadline (we denote it by T^*) ensuring the complete honesty of the agent.

3. Agent's Problem

In this section, we solve the agent's problem. Subsection 3.1 solves the agent's problem without the incentive scheme and shows that there exists a threshold belief below which the agent switches to "Arm 0" (i.e. prefers to cheat). In Subsection 3.2 we briefly discuss the "Incentive scheme" and solve the agent's problem under the proposed scheme. We show that under the proposed incentive scheme the threshold belief at which the agent switches to "Arm 0" is lower than the former one.

3.1. Agent's Problem without Incentive

The agent wants to maximize his total expected payoff by choosing the action profile $\{k_{1,t}\}_{t\geq 0}$. This is a dynamic programming problem with the current belief p_t as the state variable.

We now derive the agent's Bellman equation. By the Principle of Optimality, the agent's value function satisfies

$$u(p) = \max_{k_1\in[0,1]} \Big\{ rdt[(1-k_1)(R_0+C_0) + k_1 pR_1 + k_1\lambda(p)h_b] + e^{-rdt}\mathbb{E}[u(p+dp)|p] \Big\}$$

where the first term is the expected current payoff and the second term is the discounted expected continuation payoff.

As to the expected continuation payoff, with subjective probability $pk_1\lambda_g^G dt$ a breakthrough occurs and the agent expects a flow payoff of $u(1) = R_1$ in the future; with probability, $k_1\lambda(p)dt$ the breakdown occurs and receives payoff $u(j(p))$, where $j(p)$ is given by

$$j(p) = \frac{\lambda_b^G p}{\lambda_b^G p + (1-p)\lambda_b^B} < p$$

Moreover, with subjective probability $(1 - pk_1\lambda_g^G dt - k_1\lambda(p)dt)$ no news arrives and

his expected payoff is

$$u(p) + u'(p)dp = u(p) + k_1 u'(p)p(1-p)(\lambda_b^B - \lambda_b^G - \lambda_g^G)dt$$

Using $(1 - rdt)$ to approximate e^{-rdt}, we see the agent's discounted expected continuation payoff is

$$(1-rdt)\Big[pk_1\lambda_g^G R_1 dt + k_1\lambda(p)u(j(p))dt + \Big\{1 - pk_1\lambda_g^G dt - k_1\lambda(p)dt\Big\}\Big\{u(p)+ u'(p)k_1 p(1-p)(\lambda_b^B - \lambda_b^G - \lambda_g^G)dt\Big\}\Big]$$

Simplifying, we have the agent's discounted expected continuation payoff

$$u(p)\big(1-rdt\big)+pk_1\lambda_g^G R_1 dt + k_1\lambda(p)u(j(p))dt - k_1 p u(p)\lambda_g^G dt - k_1\lambda(p)u(p)dt + u'(p)k_1 p(1-p)[\lambda_b^B - \lambda_b^G - \lambda_g^G]dt$$

So, his expected total payoff is

$$rdt\Big[(1-k_1)(R_0+C_0)+k_1 pR_1+k_1\lambda(p)h_b\Big]+u(p)\big(1-rdt\big)+pk_1\lambda_g^G R_1 dt+k_1\lambda(p)u(j(p))dt - k_1 p u(p)\lambda_g^G dt - k_1\lambda(p)u(p)dt + u'(p)k_1 p(1-p)[\lambda_b^B - \lambda_b^G - \lambda_g^G]dt$$

Simplifying, the value function of the agent satisfies the Bellman equation

$$u(p) = (R_0 + C_0) + \max_{k_1\in[0,1]} k_1\Big\{[b_s(p,u) + b_f(p,u) + b_n(p,u)]/r - c(p)\Big\}$$

Define

$$b(p,u) = [b_s(p,u) + b_f(p,u) + b_n(p,u)]/r$$

where $b_s(p,u)$ is the expected benefit from the good news (success) from Arm 1

$$b_s(p,u) = p\lambda_g^G[R_1 - u(p)]$$

$b_f(p,u)$ is the expected loss from the bad news (failure) from Arm 1

$$b_f(p,u) = \lambda(p)[u(j(p)) - u(p)]$$

$b_n(p,u)$ is the expected loss from the absence of news (no success and no failure) from Arm 1 is

$$b_n(p,u) = p(1-p)[\lambda_b^B - \lambda_b^G - \lambda_g^G]u'(p)$$

and the opportunity cost of playing Arm 1 is

$$c(p) = (R_0 + C_0) - \Big[pR_1 + \lambda(p)h_b\Big]$$

If the opportunity cost of playing Arm 1 exceeds the total expected benefit $b(p, u)$ from Arm 1 (sum of expected benefit from success and expected loss from failure and absence of news), it is optimal to set $k_1 = 0$ and $u(p) = R_0 + C_0$, otherwise $k_1 = 1$ and u satisfies the following first-order ordinary differential-difference equation (henceforth ODDE)

$$u(p)(r + p\lambda_g^G) + \lambda(p)[u(p) - u(j(p))] - p(1-p)[\lambda_b^B - \lambda_b^G - \lambda_g^G]u'(p) = pR_1 r + r\lambda(p)h_b + p\lambda_g^G R_1 \quad (1)$$

A particular solution to the ODDE is given by

$$u(p) = pR_1 + \lambda(p)h_b - \frac{\lambda_g^G \lambda_b^G h_b}{r + \lambda_g^G} p$$

The option value being able to switch to Arm 0 is captured by $v(p) = (1-p)\Omega(p)^\mu$ for some $\mu > 0$ to be determined and this $u(p)$ constitutes the solution to the homogeneous version of the ODDE. We define

$$\Omega(p) = \frac{1-p}{p}$$

Now,

$$v'(p) = -\frac{\mu + p}{p(1-p)} v(p) \quad \textit{and} \quad v(j(p)) = \frac{\lambda_b^B}{\lambda(p)} \left(\frac{\lambda_b^B}{\lambda_b^G}\right)^\mu v(p)$$

Substituting these into the homogeneous version of the ODDE and simplifying,

$$r + \lambda_b^B - \mu\Delta\lambda = \lambda_b^B \left(\frac{\lambda_b^B}{\lambda_b^G}\right)^\mu \quad (2)$$

where

$$\Delta\lambda = \lambda_g^G + \lambda_b^G - \lambda_b^B$$

The left-hand side of (2) is a negatively sloped straight line that cuts the vertical axis at $r + \lambda_b^B$. The right-hand side is a increasing exponential function which tends to ∞ as $\mu \to \infty$, tends to 0 as $\mu \to -\infty$, and cuts the vertical axis at λ_b^B. Thus the above equation in μ has one positive solution; we write $\mu(\lambda_g^G, \lambda_b^G, \lambda_b^B)$ for the

positive solution. Observe that

$$\mu \in (0, r/\Delta\lambda)$$

Thus the solution to the ODDE for $k_1 = 1$ is given by

$$V_1(p) = pR_1 + \lambda(p)h_b - \frac{\lambda_g^G \lambda_b^G h_b}{r + \lambda_g^G} p + C(1-p)\Omega(p)^\mu$$

where C is a constant of integration.

PROPOSITION **1.** *In the agent's problem, there exists a threshold belief $p_1^* > 0$ such that the optimal policy $k_1^*(p)$ is given by*

$$k_1^*(p) = \begin{cases} 1 \ (\textit{playing "}\textbf{\textit{Arm 1}}\textit{"}) & \textit{if } p \in (p_1^*, 1] \\ 0 \ (\textit{playing "}\textbf{\textit{Arm 0}}\textit{"}) & \textit{if } p \in [0, p_1^*] \end{cases}$$

and the threshold belief satisfies

$$p_1^* = \frac{\mu(R_0 + C_0 - \lambda_b^B h_b)}{R_1(1+\mu) + (1+\mu)\lambda_b^G h_b - \mu\lambda_b^B h_b - (1+\mu)\frac{\lambda_g^G \lambda_b^G h_b}{r+\lambda_g^G} - (R_0 + C_0)} \tag{3}$$

where $\mu > 0$ is the positive solution to

$$r + \lambda_b^B - \mu[\lambda_g^G + \lambda_b^G - \lambda_b^B] = \lambda_b^B \left(\frac{\lambda_b^B}{\lambda_b^G}\right)^\mu$$

and the agent's value function under optimal policy is given by

$$V_1^*(p) = \begin{cases} V_1(p) & \textit{if } p \in (p_1^*, 1] \\ (R_0 + C_0) & \textit{if } p \in [0, p_1^*] \end{cases}$$

where

$$V_1(p) = pR_1 + \lambda(p)h_b - \frac{\lambda_g^G \lambda_b^G h_b}{r + \lambda_g^G} p + \Big[R_0 + C_0 - p_1^* R_1 - \lambda(p_1^*)h_b + \frac{\lambda_g^G \lambda_b^G h_b}{r + \lambda_g^G} p_1^*\Big] \left(\frac{1-p}{1-p_1^*}\right)\left(\frac{\Omega(p)}{\Omega(p_1^*)}\right)^\mu$$

PROOF: See Appendix A.1 ■

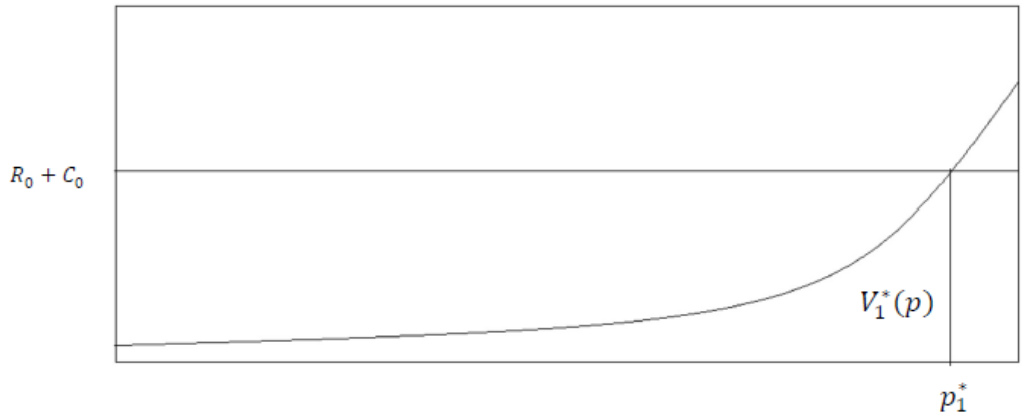

Figure 2: Agent's value functions without incentive

Remark:

From Proposition 1, we have the explicit expression for the threshold p_1^* at which the agent switches to the "Arm 0" from "Arm 1". In Lemma 1 we have already shown that the threshold belief p_1^* exists and it is positive. The value function for the Agent's problem without incentive (V_1^*) and the threshold belief (p_1^*) are illustrated in Figure 1.

To stop the agent from cheating, the principal designs an incentive scheme. Next, we show that under the proposed incentive scheme, the agent has no option to cheat. It can also be possible that under some conditions even if the agent plans to cheat by switching to Arm 0, the threshold will be much lower. The next section follows the discussion of the incentive scheme and looks into the agent's problem with the proposed incentive scheme.

3.2. Agent's Problem with Incentive

B.1. Incentive Scheme

The primary objective of the principal is to prevent the agent from cheating (by playing Arm 0). Every time, when the agent finds any news (breakthrough or breakdown), he reports to the principal. So, the principal has a sequence of news reported by the agent. But interestingly he doesn't know from where the information is coming (from "Arm 0 or Arm 1"). So, the principal asks the agent to come up with a certain number of news containing breakthroughs and breakdowns. The principal calculates the ratio of the number of breakthroughs and breakdowns reported by the agent (henceforth denoted by λ). The principal chooses an exogenous threshold of the ratio (henceforth denoted by $\bar{\lambda}$) between

the number of breakdowns and breakthroughs. Now if the ratio calculated from the reported sequence of breakthroughs and breakdowns exceeds the threshold, the principal surely predicts that the agent has come up with a relatively high number of bad news and more likely pulled "Arm 1" throughout his experiment.[4]So, as a reward for honesty, the principal pays for the bad news he reported. On the other hand, if the calculated ratio from the reported sequence can not exceed the threshold, the principal surely concludes that the agent has come up with a relatively high number of good news and cheated by pulling "Arm 0" throughout his experiment. Hence, the principal now punishes the agent for the good news he reported. More importantly, this incentive is ex-ante information to the agent. So, in his maximization framework, he considers the expected payment obtained from the incentive scheme. Before going to the structural framework, we demonstrate this with a simple example.

EXAMPLE**.** Suppose the principal asks the agent to come up with 10 news consisting of breakthroughs as well as breakdowns. So, we plot these combinations (breakdowns, breakthroughs) in a real line. The texts below the number line denote

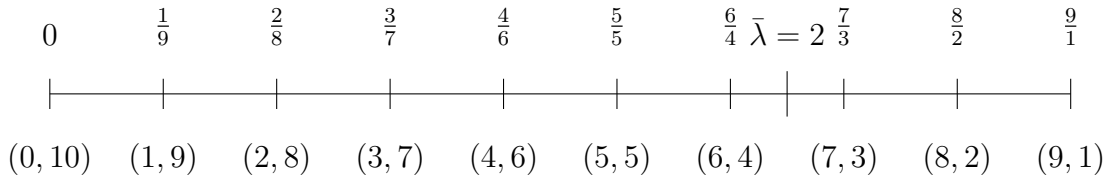

Figure 3

the combinations of news(number of breakdowns, number of breakthroughs), and the text above represents the ratio of the number of breakdowns to breakthroughs. Now, he will choose an exogenous threshold of 2. Now, the principal will punish the agent if his ratio obtained from the reported number of breakdowns and breakthroughs is not enough to cross the threshold of 2, and reward otherwise. Since the information about the incentive scheme is ex-ante to the agent, he will calculate and incorporate the expected reward in his optimization framework. The reward and punishment scheme will be elaborated on in the next section. For now, we represent R as reward and P as punishment.

Interestingly, the number of breakthroughs decreases as we move along the real number line (the opposite holds for the number of breakdowns). Hence we can conclude that the ratio of the number of breakdowns to the number of breakthroughs

[4]Note that relatively higher number of bad news can only be attained from "Arm 1" rather than "Arm 0" since $\gamma_g > \lambda_g^G$ and $\lambda_b^B > \lambda_b^G > \gamma_b$

is unique and for the simplicity of the calculation of the probability, we need to consider the number of breakthroughs only. So, the probability that the ratio will cross the threshold of 2 is the sum of the probability that the number of breakthroughs (g) is 1,2,3 (See Figure 2).

So, the expected payoff can be written as:

$$\mathbb{P}(\lambda \geq 2)R + \mathbb{P}(\lambda < 2)P$$

Now, the $\mathbb{P}(\lambda \geq 2)$ can be written as

$$\sum_{i=1}^{3} \mathbb{P}(g = i)$$

and the $\mathbb{P}(\lambda < 2)$ can be written as

$$\sum_{i=4}^{10} \mathbb{P}(g = i)$$

where g represents the number of breakthroughs. So, the expected payoff can be written like this:

$$\sum_{i=1}^{3} \mathbb{P}(g = i)R + \sum_{i=4}^{10} \mathbb{P}(g = i)P$$

We now move on to the formal incentive structure in the next section.

B.2. Incentive Structure

We list the notations used throughout the paper.

1. **Basic Notations:**

   I. m denotes the number of news asked by the principal.

   II. $V(1), V(0) > 0$ denotes the exogenous payment as a reward and punishment respectively.

   III. N_g and N_b denote the number of breakthroughs and breakdowns reported by the agent.

   IV. λ denotes the ratio of the reported breakdowns and breakthroughs.

First, we construct $\bar{\lambda}$ as

$$\bar{\lambda} = a\frac{\gamma_b}{\gamma_g} + (1-a)\frac{\lambda_b^G}{\lambda_g^G} \quad \text{, where } a \in (0,1) \tag{4}$$

where $\bar{\lambda}$ denotes the exogenous threshold fixed by the principal.

Mechanism. *I. The principal asks the agent to submit a sequence of news including successes and failures, denoted by m.*
II. The principal calculates the ratio of failures and successes achieved by the agent.
III. He chooses an exogenous threshold as an honesty parameter, denoted by $\bar{\lambda}$.
IV. The principal rewards for the bads when the calculated ratio crosses the threshold and punishes for the goods when the agent fails to maintain honesty by not crossing the pre-determined threshold.

We define the **reward** and **punishment scheme** using the notations we have described above. As we have mentioned earlier in the section of the incentive scheme, the principal will check whether the λ calculated from the reported news crosses the threshold $\bar{\lambda}$ or not. If λ crosses $\bar{\lambda}$, the principal surely knows that the agent is honest and operated "Arm 1" and rewards the agent for the breakthroughs he reported. If λ can't cross $\bar{\lambda}$, the principal surely knows that the agent cheated and operated "Arm 0" and punishes the agent for the breakdowns he reported. So the expected payoff for the honest agent under the **reward scheme** is expressed as follows

$$(\lambda - \bar{\lambda})N_b V(1)$$

where, $(\lambda - \bar{\lambda})$ represents how far the agent is from the threshold $\bar{\lambda}$. The expected payoff for the off-path agent(playing "Arm 0") under the **punishment scheme** is expressed as follows

$$(\lambda - \bar{\lambda})N_g V(0)$$

Hence, the **ex-ante expected payoff** for the agent is

$$\mathbb{P}(\lambda \geq \bar{\lambda})(\lambda - \bar{\lambda})N_b V(1) + \mathbb{P}(\lambda < \bar{\lambda})(\lambda - \bar{\lambda})N_g V(0)$$

We need an explicit representation of $\mathbb{P}(\lambda \geq \bar{\lambda})$ and $\mathbb{P}(\lambda < \bar{\lambda})$. As we have mentioned earlier, the $\mathbb{P}(\lambda \geq \bar{\lambda})$ can be represented as a summation of the probability that the agent is receiving some number of breakthroughs. More formally, let us find for a given m and $\bar{\lambda}$, what is the exact number of breakthroughs the agent can at least achieve.

Let, g represent the number of breakthroughs and b represent the number of breakdowns. Additionally, we have $g + b = m$. Hence, we have,

$$\frac{b}{g} = \bar{\lambda} \implies g = \frac{m}{1 + \bar{\lambda}}$$

In the example that we have mentioned earlier, for m=10 and $\bar{\lambda} = 2$, we have $g = 3.33 \approx 3$. So for the calculation of the probability $\mathbb{P}(\lambda \geq \bar{\lambda})$, we need $g = 1, 2, 3$

Figure 4: Diagrammatic Representation of the ratio of reported breakdowns and breakthroughs

and the rest for the opposite calculation. The texts below the number line denote the number of breakthroughs that occurred till time T, and the text above represents the ratio of the number of breakdowns to breakthroughs. From the calculation, we know that the number of breakthroughs corresponding to $\bar{\lambda}$ is $\frac{m}{1+\bar{\lambda}}$. So, from the Figure 3, the calculation of $\mathbb{P}(\lambda \geq \bar{\lambda})$ involves the summation of $\mathbb{P}(g = i)$ where $i = 1, 2, 3, ..., \frac{m}{1+\bar{\lambda}}$ and the rest for the $\mathbb{P}(\lambda < \bar{\lambda})$. Hence,

$$\mathbb{P}(\lambda \geq \bar{\lambda}) = \sum_{i=1}^{\frac{m}{1+\bar{\lambda}}} \mathbb{P}(g = i)$$

$$\mathbb{P}(\lambda < \bar{\lambda}) = \sum_{i=(\frac{m}{1+\bar{\lambda}}+1)}^{m} \mathbb{P}(g = i)$$

$$\implies \mathbb{P}(\lambda < \bar{\lambda}) = \sum_{i=1}^{m} \mathbb{P}(g = i) - \sum_{i=1}^{\frac{m}{1+\bar{\lambda}}} \mathbb{P}(g = i)$$

I. Explicit representation of $\mathbb{P}(g = i)$

Earlier in our model setup, we have mentioned the point process (Poisson) for the number of breakthroughs $\{(N_g)_t\}_{0 \leq t \leq \bar{T}}$ consisting of the number of breakthroughs $\{(N_g^1)_t\}_{0 \leq t \leq \bar{T}}$ coming from the "Arm 1" and the number of breakthroughs $\{(N_g^0)_t\}_{0 \leq t \leq \bar{T}}$ coming from the "Arm 0" with the intensity $\lambda_g^G k_1$ and $\gamma_g(1-k_1)$ respectively. So, $\{(N_g)_t\}_{0 \leq t \leq \bar{T}}$ follows the poison process with the intensity $(1-k_1)\gamma_g + \lambda_g^G k_1$. So, the probability that the number of breakthroughs occurred during the time T can be calculated as follows:

$$\mathbb{P}(g = i) = e^{\{-((1-k_1)\gamma_g + \lambda_g^G k_1)T\}} \frac{\{((1-k_1)\gamma_g + \lambda_g^G k_1)T\}^i}{i!}$$

Let, $K = \{((1-k_1)\gamma_g + \lambda_g^G k_1)T\}$, rewriting the above equation we have

$$\mathbb{P}(g = i) = e^{-K} \frac{K^i}{i!}$$

So, the expressions for $\mathbb{P}(\lambda \geq \bar{\lambda})$ and $\mathbb{P}(\lambda < \bar{\lambda})$ can be written as,

$$\mathbb{P}(\lambda \geq \bar{\lambda}) = \sum_{i=1}^{\frac{m}{1+\bar{\lambda}}} e^{-K} \frac{K^i}{i!}$$

$$\mathbb{P}(\lambda < \bar{\lambda}) = \sum_{i=\frac{m}{1+\bar{\lambda}}+1}^{m} e^{-K} \frac{K^i}{i!}$$

Hence, the **ex-ante expected payoff** for the agent is defined as follows

$$\sum_{i=1}^{\frac{m}{1+\bar{\lambda}}} e^{-K} \frac{K^i}{i!}(\lambda - \bar{\lambda}) N_b V(1) + \sum_{i=\frac{m}{1+\bar{\lambda}}+1}^{m} e^{-K} \frac{K^i}{i!}(\lambda - \bar{\lambda}) N_g V(0)$$

$$\text{where, } K = \{((1-k_1)\gamma_g + \lambda_g^G k_1)T\}$$

Under this proposed incentive scheme, we look into the agent's optimization problem in the next section.

B.3. Agent's value function

Under the ex-ante incentive scheme, the agent wants to maximize his total expected payoff by choosing the action profile $\{k_{1,t}\}_{t\geq 0}$. Given this incentive scheme, the agent will choose a strategy $\{k_{1,t}\}_{t\geq 0}$ that maximizes

$$\mathbb{E}\left[\int_0^T re^{-rt}\Big[(1-k_{1,t})(R_0+C_0) + k_{1,t}\Big(p_t R_1 + \lambda(p_t)h_b\Big)\Big]dt\right] + T_1(k_{1,t}) + T_2(k_{1,t})\Bigg]$$

where, $T_1(k_{1,t})$ and $T_2(k_{1,t})$ is defined by

$$T_1(k_{1,t}) = e^{-rT} \sum_{i=1}^{\frac{m}{1+\bar{\lambda}}} e^{-K} \frac{K^i}{i!}(\lambda - \bar{\lambda}) N_b V(1) \text{ and}$$

$$T_2(k_{1,t}) = e^{-rT} \sum_{i=\frac{m}{1+\bar{\lambda}}+1}^{m} e^{-K} \frac{K^i}{i!}(\lambda - \bar{\lambda}) N_g V(0)$$

and

$$K = \{((1-k_{1,t})\gamma_g + \lambda_g^G k_{1,t})T\}$$

The first term is the payoff generated from the experimentation performed by the agent and the second term is the expected wage payment received by the agent at the end of the experimentation.

In the earlier section, we have derived the agent's Bellman equation. Since the expected wage payment is independent of the prior (as well as posterior) belief, the final term will be considered as a constant and will not be accounted for in the calculation for the Bellman equation. Using the prior calculation, we obtain the agent's value function as follows

$$u(p) = R_0 + C_0 + \max_{k_1 \in [0,1]} \Big[k_1 \Big\{ [b_s(p,u) + b_f(p,u) + b_n(p,u)]/r - c(p) \Big\} + T_1(k_1) + T_2(k_1) \Big]$$

where, the definition of $b_s(p,u), b_f(p,u), b_n(p,u), c(p)$ follows from the Agent's problem without Incentive. (See Subsection 3.1 for more details)

Note that the last term involving the principal's incentive scheme is a non-linear function of k_1. We ensure that under some assumptions imposed on the parameter, we don't have any interior solution. (For detailed discussion please refer to Appendix B.2)

NOTE. We have already shown that under some conditions imposed on the parameter, we don't have any interior solution, hence $k_1 = \{0, 1\}$ solves the agent's optimization problem. We carefully investigate the terms $T_1(k_1)$ and $T_2(k_1)$ at $k_1 = 0, 1$.

Case I: $(k_1 = 0)$

Suppose the agent pulls Arm 0 throughout his experiment, in that case, $k_1 = 0$. By the assumption of Arm 0 and for a given interval of the time the agent will receive more breakthroughs than breakdowns. Therefore, with a very low probability λ calculated from the agent's reporting can cross the threshold $\bar{\lambda}$. We denote this by $\epsilon_c > 0$, where c denotes cheating. Therefore, the probability that λ is strictly lower than the threshold is $1 - \epsilon_c$.
Now,

$$\begin{aligned} T_1(0) &= e^{-rT} \mathbb{P}(\lambda \geq \bar{\lambda})(\lambda - \bar{\lambda}) N_b V(1) \\ &= e^{-rT} (\lambda - \bar{\lambda}) N_b V(1) * \epsilon_c = \epsilon_0 \end{aligned}$$

Therefore, the value of $T_1(0)$ is very low, we denote it by ϵ_0, where "0" indicates pulling Arm 0. Also,

$$\begin{aligned} T_2(0) &= e^{-rT} \mathbb{P}(\lambda < \bar{\lambda})(\lambda - \bar{\lambda}) N_g V(0) \\ &= e^{-rT} (\lambda - \bar{\lambda}) N_g V(0)(1 - \epsilon_c) \end{aligned}$$

As the probability that λ calculated from the agent's reporting crosses the threshold $\bar{\lambda}$, tends to zero i.e. $\epsilon_c \to 0$, we have,

$$T_1(0) = 0 \text{ and } T_2(0) = e^{-rT}(\lambda - \bar{\lambda})N_g V(0) \tag{5}$$

Case II: $(k_1 = 1)$

Suppose the agent pulls Arm 1 throughout his experiment, in that case, $k_1 = 1$. By the assumption of Arm 1 and for a given interval of the time the agent will receive more breakdowns than breakthroughs. Therefore, the probability that λ is strictly lower than the threshold $\bar{\lambda}$ is very low. We denote this by $\epsilon_h > 0$, where h denotes honesty. Therefore, the probability that λ calculated from the agent's reporting crosses the threshold $\bar{\lambda}$ is $1 - \epsilon_h$.
Now,

$$\begin{aligned} T_2(1) &= e^{-rT}\mathbb{P}(\lambda < \bar{\lambda})(\lambda - \bar{\lambda})N_g V(0) \\ &= e^{-rT}(\lambda - \bar{\lambda})N_g V(0) * \epsilon_h = \epsilon_1 \end{aligned}$$

Therefore, the value of $T_2(1)$ is very low, we denote it by ϵ_1, where "1" indicates pulling Arm 1. Also,

$$\begin{aligned} T_1(1) &= e^{-rT}\mathbb{P}(\lambda \geq \bar{\lambda})(\lambda - \bar{\lambda})N_b V(1) \\ &= e^{-rT}(\lambda - \bar{\lambda})N_b V(1) * (1 - \epsilon_h) \end{aligned}$$

As the probability that λ is strictly lower than the threshold $\bar{\lambda}$, tends to zero i.e. $\epsilon_h \to 0$, we have,

$$T_1(1) = e^{-rT}(\lambda - \bar{\lambda})N_b V(1) \text{ and } T_2(1) = 0 \tag{6}$$

Interestingly, we can easily argue $T_2(0)$ is negative and $T_1(1)$ is positive. In both the expression of $T_2(0)$ and $T_1(1)$, we have $(\lambda - \bar{\lambda})$. Notice, whenever $k_1 = 0$ (Pulling Arm 0) calculated λ is strictly lower than $\bar{\lambda}$ and for $k_1 = 1$ (experimenting with Arm 1) λ always exceeds $\bar{\lambda}$. Hence,

$$T_2(0) = e^{-rT} \underbrace{(\lambda - \bar{\lambda})}_{-ve} N_g V(0) = -ve$$

and

$$T_1(1) = e^{-rT} \underbrace{(\lambda - \bar{\lambda})}_{+ve} N_b V(1) = +ve$$

We rewrite the Agent's maximization problem under the incentive scheme and

it simplifies to the following Bellman equation

$$u(p) = R_0 + C_0 + \max_{k_1 \in [0,1]} \Big[k_1 \Big\{ [b_s(p,u) + b_f(p,u) + b_n(p,u)]/r - c(p) \Big\} + T_1(k_1) + T_2(k_1) \Big]$$

where, the definition of $b_s(p,u), b_f(p,u), b_n(p,u), c(p)$ follows from the Agent's problem without Incentive.

Define $\bar{b}(p,u)$, to be the total expected experimentation benefit under the incentive scheme (including the reward from the principal) and it is given by

$$\bar{b}(p,u) = b(p,u) + T_1(1)$$

and the total opportunity cost of playing Arm 1 (including the punishment from the principal) is given by

$$\bar{c}(p) = c(p) + T_2(0)$$

If the opportunity cost of playing Arm 1, $\bar{c}(p)$ (including exogeneous punishment from the principal) exceeds the total expected benefit from Arm 1, $\bar{b}(p,u)$ (sum of expected benefit from success, and expected loss from failure and absence of news, and exogenous reward from the principal), it is optimal to set $k_1 = 0$ and $u(p) = R_0 + C_0 + T_2(0)$, otherwise $k_1 = 1$ and u satisfies the following first-order ordinary differential-difference equation (henceforth ODDE)

$$u(p)(r + p\lambda_g^G) + \lambda(p)[u(p) - u(j(p))] - p(1-p)[\lambda_b^B - \lambda_b^G - \lambda_g^G]u'(p) = pR_1 r + r\lambda(p)h_b + p\lambda_g^G R_1 + rT_1(1)$$

A particular solution to the ODDE is given by

$$u(p) = pR_1 + \lambda(p)h_b - \frac{\lambda_g^G \lambda_b^G h_b}{r + \lambda_g^G} p - \frac{\lambda_g^G T_1(1)}{r + \lambda_g^G} p$$

Since the homogeneous version of the ODDE is the same as the Agent's problem without incentive, we can proceed with the same μ, as derived in the Subsection 3.1. Thus the solution to the ODDE for $k_1 = 1$ is given by

$$\bar{V}_1(p) = pR_1 + \lambda(p)h_b - \frac{\lambda_g^G \lambda_b^G h_b}{r + \lambda_g^G} p - \frac{\lambda_g^G T_1(1)}{r + \lambda_g^G} p + C(1-p)\Omega(p)^\mu$$

where C is a constant of integration.

The Proposition 2 & Proposition 3 stated below describes the optimal policy

for the agent and the switching belief at which he prefers not to work. Most interestingly, this threshold belief ($\bar{p}_1$) is lower than the threshold belief without incentive (p_1^*). This shows that under the incentive scheme, the agent sticks to "Arm 1" for a higher range of belief.

PROPOSITION **2.** *In the agent's problem under the incentive scheme, there exists* $\bar{p}_1 \geq 0$ *such that the optimal policy* $\bar{k}_1(p)$ *is given by*

$$\bar{k}_1(p) = \begin{cases} 1 & \text{if } p \in (\bar{p}_1, 1] \\ 0 & \text{if } p \in [0, \bar{p}_1] \end{cases}$$

and the threshold belief satisfies

$$\bar{p}_1 = \frac{\mu\Big[R_0 + C_0 - \lambda_b^B h_b + T_2(0) - T_1(1)\Big]}{R_1(1+\mu) + (1+\mu)\lambda_b^G h_b - \mu\lambda_b^B h_b - (1+\mu)\frac{\lambda_g^G \lambda_b^G h_b}{r+\lambda_g^G} + \frac{rT_1(1)}{r+\lambda_g^G} - \mu T_1(1) - (R_0 + C_0) - T_2(0)} \tag{7}$$

where $\mu > 0$ *is the positive solution to*

$$r + \lambda_b^B - \mu[\lambda_g^G + \lambda_b^G - \lambda_b^B] = \lambda_b^B \left(\frac{\lambda_b^B}{\lambda_b^G}\right)^{\mu}$$

and the agent's value function under optimal policy is given by

$$\bar{V}_1(p) = \begin{cases} V(p) & \text{if } p \in (\bar{p}_1, 1] \\ R_0 + C_0 + T_2(0) & \text{if } p \in [0, \bar{p}_1] \end{cases}$$

where

$$V(p) = pR_1 + \lambda(p)h_b - \frac{\lambda_g^G \lambda_b^G h_b}{r+\lambda_g^G}p - \frac{\lambda_g^G T_1(1)}{r+\lambda_g^G}p + \Big[R_0 + C_0 + T_2(0) - \bar{p}_1 R_1 + \lambda(\bar{p}_1)h_b - \frac{\lambda_g^G \lambda_b^G h_b}{r+\lambda_g^G}\bar{p}_1 - \frac{\lambda_g^G T_1(1)}{r+\lambda_g^G}\bar{p}_1\Big]\left(\frac{1-p}{1-\bar{p}_1}\right)\left(\frac{\Omega(p)}{\Omega(\bar{p}_1)}\right)^{\mu}$$

and $T_2(0)$, $T_1(1)$ *and* $\bar{\lambda}$ *are respectively defined as follows*

$$T_2(0) = e^{-rT}(\lambda - \bar{\lambda})N_g V(0)$$
$$T_1(1) = e^{-rT}(\lambda - \bar{\lambda})N_b V(1)$$
$$\bar{\lambda} = a\frac{\gamma_b}{\gamma_g} + (1-a)\frac{\lambda_b^G}{\lambda_g^G}$$

PROOF: See Appendix A.2 ■

Till now, we have established the existence of threshold belief $\bar{p}_1$, below which the agent switches to the known option and stops experimenting. Now we will state our main result, which says under the proposed mechanism, the agent now switches to a lower threshold belief, i.e. the agent spends more time experimenting "Arm 1".

PROPOSITION **3.** *The switching threshold belief from the Agent's problem with Incentive ($\bar{p}_1$) is strictly less than of the Agent's problem without Incentive (p_1^*) i.e. $\bar{p}_1 < p_1^*$.*

PROOF: See Appendix A.3 ■

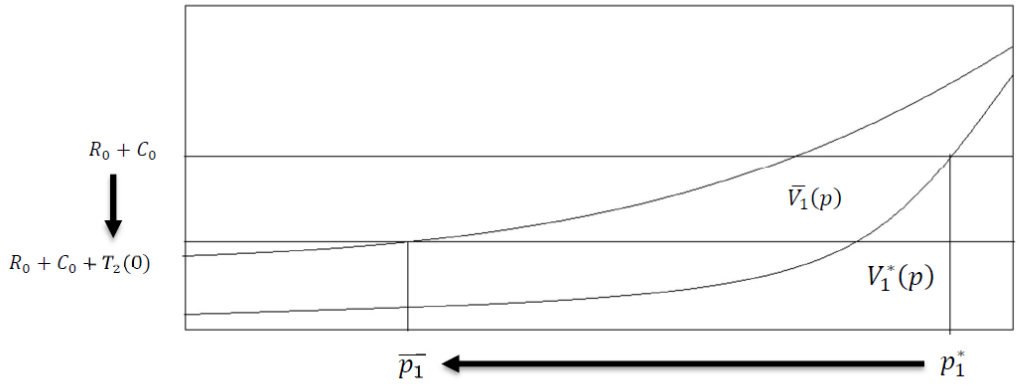

Figure 5: Agent's value functions with and without incentive

Remark: In the Incentive Structure, we have mentioned that the principal rewards an honest agent for the bad news he reports and punishes an off-path agent for the good news he reports. In the Proposition 2 & Proposition 3, we have shown that under this Incentive scheme, the agent has no option to cheat, and even if he plans to cheat, the threshold belief at which he switches to "Arm 0" is lower than the former one. The value function for the Agent's problem under the incentive scheme ($\bar{V}_1$) and the threshold belief ($\bar{p}_1$) are illustrated in Figure 4.

The same result will hold even if we slightly alter our proposed incentive scheme. Instead of rewarding the honest agent with the bad news he reported, if we compensate him by paying for the good news he reported and punishing the off-path agent for the bad news he reported, we can obtain the same result as before.
Formally, if the agent maintains his honesty by pulling the "Arm 1" throughout

his experiment, the principal will compensate for the cost he incurs for several bad news. So, his ex-ante expected reward is

$$e^{-rT}(\lambda - \bar{\lambda})N_g V(1)$$

But, if the agent cheats on the principal throughout his experiment and gains lump-sum payoffs from his reported good news, the principal will punish him for the bad news he reports. So, his ex-ante expected punishment is

$$e^{-rT}(\lambda - \bar{\lambda})(N_b)^{-1}V(0)$$

More specifically, the Proposition 3 holds in the slightly modified version of the proposed incentive scheme.
This concludes the agent's problem. In the next section, we solve the principal's problem. Given the proposed incentive scheme, the principal can prevent the agent from pulling the "Arm 0". To ensure full honesty from the experimentation, he optimally chooses an end date (deadline) at which the experimentation ceases and the principal pays the agent for his effort (See Proposition 4). Next, Proposition 5 looks into the Arm Identification problem for the principal.

4. Principal's Problem

4.1. Optimal stopping time

We derive the objective function of the principal. The principal gains a higher private benefit of R_P from "Arm 1" which includes the current benefit from the agent's experimentation on Arm 1 and the future valuation upon the revelation of the true state of the unknown project. Hence, his objective function considers the experimentation value as well as payment to the agent according to the experimentation outcome. So the objective function of the principal is given by

$$V_P(T) = R_P + T_1(1) + T_2(0) = R_P + e^{-rT}(\lambda - \bar{\lambda})N_b V(1) + e^{-rT}(\lambda - \bar{\lambda})N_g V(0) \quad (8)$$

Now, the principal wants the agent to experiment by honest means, i.e. the λ crosses the threshold $\bar{\lambda}$. Then, in equilibrium, the component in the principal's objective function $T_2(0)$ will be zero. Hence, the principal's objective function is given by

$$V_P(T) = R_P + e^{-rT}(\lambda - \bar{\lambda})N_b V(1) \quad (9)$$

On the equilibrium path characterized by the Markov perfect equilibrium, the principal will optimize his payoff function by optimally choosing the deadline for the experimentation and denoted by T^*

As discussed in the previous section, we observe that under the incentive scheme proposed by the principal, the agent will switch himself to Arm 0 at a lower threshold. Keeping this in mind, the principal will choose the end date, cease the experimentation, and pay the agent accordingly. For this purpose, he chooses optimal T^* to ensure that the agent never opts for playing "Arm 0", more formally, $\bar{p}_1 = 0$. The main idea of the proof is that here we want to ensure full honesty from the agent, our condition $\bar{p}_1 = 0$ is equivalent to the fact that the λ calculated from the agent's reported information always exceeds the exogenous threshold $\bar{\lambda}$, more formally $\mathbb{P}(\lambda \geq \bar{\lambda}) = 1$. More specifically, we are only interested in the fact that the honest agent just crosses the threshold $\bar{\lambda}$, hence then the number of good news, N_g (or bad news, N_b) corresponding to his reported λ just next to $\bar{\lambda}$ serves our requirement for the proof.

PROPOSITION **4.** *The principal will stop the game at time T^* when $\bar{p}_1 = 0$ where T^* is given by*

$$T^* = \frac{1}{r} \ln \frac{(\bar{\lambda}+1)(m\bar{\lambda}+1+\bar{\lambda})V(1)}{(R_0 + C_0 - \lambda_b^B h_b)(m-1-\bar{\lambda})} \tag{10}$$

PROOF: Let, the experimentation ceases at T^*. Hence, at T^*, $\bar{p}_1 = 0$
Now, from the expression of $\bar{p}_1$, we have,

$$\bar{p}_1 = \frac{\mu\Big[R_0 + C_0 - \lambda_b^B h_b + T_2(0) - T_1(1)\Big]}{R_1(1+\mu) + (1+\mu)\lambda_b^G h_b - \mu\lambda_b^B h_b - (1+\mu)\frac{\lambda_g^G \lambda_b^G h_b}{r+\lambda_g^G} + \frac{rT_1(1)}{r+\lambda_g^G} - \mu T_1(1) - (R_0 + C_0) - T_2(0)}$$

Set, $\bar{p}_1 = 0$. Since the denominator is positive, $\bar{p}_1 = 0$ implies

$$R_0 + C_0 - \lambda_b^B h_b + T_2(0) - T_1(1) = 0$$

Where $T_2(0)$ represents the ex-ante punishment for the off-path agent for playing "Arm 0" and $T_1(1)$ represents the ex-ante reward for the honest agent for playing "Arm 1".

Now, the condition $\bar{p}_1 = 0$ ensures that the agent never switches to Arm 0, in other words, he never operates on Arm 0. Hence, in this context,

$$T_2(0) = 0 \text{ and } T_1(1) = e^{-rT^*}(\lambda - \bar{\lambda})N_b V(1)$$

From the condition, we have,

$$R_0 + C_0 - \lambda_b^B h_b - e^{-rT^*}(\lambda - \bar{\lambda})N_b V(1) = 0$$

Simplifying, we have,

$$T^* = \frac{1}{r} \ln \frac{(\lambda - \bar{\lambda})N_b V(1)}{R_0 + C_0 - \lambda_b^B h_b}$$

Now, since we are only interested in the situation where the calculated ratio just crosses the exogenous threshold, in that case, $N_b = m - (\frac{m}{1+\bar{\lambda}} - 1) = (\frac{m\bar{\lambda}}{1+\bar{\lambda}} + 1)$ and corresponding $(\lambda - \bar{\lambda})$ can be calculated as follows:

$$\lambda - \bar{\lambda} = \frac{m - (\frac{m}{1+\bar{\lambda}} - 1)}{\frac{m}{1+\bar{\lambda}} - 1} - \bar{\lambda}$$

$$\implies \lambda - \bar{\lambda} = \frac{\bar{\lambda} + 1}{\frac{m}{1+\bar{\lambda}} - 1}$$

And,

$$N_b(\lambda - \bar{\lambda}) = (\frac{m\bar{\lambda}}{1+\bar{\lambda}} + 1)\frac{\bar{\lambda} + 1}{(\frac{m}{1+\bar{\lambda}} - 1)} = \frac{(\bar{\lambda} + 1)(m\bar{\lambda} + 1 + \bar{\lambda})}{m - 1 - \bar{\lambda}}$$

Now, substituting the value of $N_b(\lambda - \bar{\lambda})$ in the expression of T^*, we have,

$$T^* = \frac{1}{r} \ln \frac{(\bar{\lambda} + 1)(m\bar{\lambda} + 1 + \bar{\lambda})V(1)}{(R_0 + C_0 - \lambda_b^B h_b)(m - 1 - \bar{\lambda})}$$

Hence, proved. ■

In this section, We have obtained the optimal deadline that ensures the full honesty of the agent. The question now remains to solve is whether the unknown arm is good or bad. In the next section, we identify the type of the unknown arm.

4.2. Identification of Arm-1

Since, the optimal deadline T^* ensures full honesty, the problem of identification is straightforward if the sequence of good-bad news reported by the agent contains at least one good news. In this case, we directly conclude that the unknown arm is of good type. Now if the agent's reported sequence does not contain any good news, in that case, with a high probability the principal can conclude that the unknown arm is of bad type. We will argue this formally in Proposition 5.

PROPOSITION **5.** *Arm Identification:*

A. If at least one good news arrives till the optimal deadline T^*, given the mechanism the principal can identify that Arm 1 is good with certainty.

B. If no good news has arrived till the optimal deadline T^*, given the mechanism the principal identifies the Arm 1 to be bad with a higher probability than that without the mechanism.

PROOF: A. Given the mechanism, in equilibrium, the agent plays only Arm 1. Since a bad Arm 1 can not produce any good news, therefore at least one good news arrives, the principal is certain that it comes from a good Arm 1.
B. We know bad news arrives from a Poisson process with intensity λ_b^G and λ_b^B whenever Arm 1 is good and bad respectively. Also, Arm 0 produces bad news through a Poisson process with intensity γ_b. The principal ceases the experimentation at time T^* and asks to submit m news including breakthroughs and breakdowns. When the agent reports no breakthrough, it is equivalent to reporting m breakdowns.

Without any mechanism, since bad news can be achieved from both types (good and bad) of Arm 1 and Arm 0, the probability of Arm 1 being bad given the information that m bad news is achieved is given by,

$$\mathbb{P}_{W/O}(\textit{Arm 1 is bad}|m \textit{ bad news})$$

Define $R_G = \textit{Arm 1 is good}$ and $R_B = \textit{Arm 1 is bad}$

$$= \frac{\mathbb{P}(m \textit{ bad news}|R_B)\mathbb{P}(R_B)}{\mathbb{P}(m \textit{ bad news}|R_G)\mathbb{P}(R_G) + \mathbb{P}(m \textit{ bad news}|R_B)\mathbb{P}(R_B) + \mathbb{P}(m \textit{ bad news}|\textit{Arm 0})\mathbb{P}(\textit{Arm 0})}$$

Now, the probability of m-bad news achieved from the Arm 1 of good, bad type and Arm 0 are given by

$$\mathbb{P}(m \textit{ bad news}|R_B)\mathbb{P}(R_B) = e^{-\lambda_b^B T^*}(\lambda_b^B)^m \frac{T^{*m}}{m!}$$
$$\mathbb{P}(m \textit{ bad news}|R_G)\mathbb{P}(R_G) = e^{-\lambda_b^G T^*}(\lambda_b^G)^m \frac{T^{*m}}{m!}$$
$$\mathbb{P}(m \textit{ bad news}|\textit{Arm 0}) = e^{-\gamma_b T^*}(\gamma_b)^m \frac{T^{*m}}{m!}$$

Hence, $\mathbb{P}_{W/O}(\textit{Arm 1 is bad}|m \textit{ bad news})$ is given by

$$= \frac{(1-p_{T^*})e^{-\lambda_b^B T^*}(\lambda_b^B)^m \frac{T^{*m}}{m!}}{p_{T^*}e^{-\lambda_b^G T^*}(\lambda_b^G)^m \frac{T^{*m}}{m!} + (1-p_{T^*})e^{-\lambda_b^B T^*}(\lambda_b^B)^m \frac{T^{*m}}{m!} + e^{-\gamma_b T^*}(\gamma_b)^m \frac{T^{*m}}{m!}}$$
$$= \frac{(1-p_{T^*})e^{-\lambda_b^B T^*}(\lambda_b^B)^m}{p_{T^*}e^{-\lambda_b^G T^*}(\lambda_b^G)^m + (1-p_{T^*})e^{-\lambda_b^B T^*}(\lambda_b^B)^m + e^{-\gamma_b T^*}(\gamma_b)^m}$$

Now, our mechanism restricts bad news arriving from Arm 0. Hence, the probability

of Arm 1 being bad given the information that m bad news is achieved is given by

$$\begin{aligned}\mathbb{P}_W(\textit{Arm 1 is bad}|m \textit{ bad news}) =& \frac{\mathbb{P}(m \textit{ bad news}|R_B)\mathbb{P}(R_B)}{\mathbb{P}(m \textit{ bad news}|R_G)\mathbb{P}(R_G) + \mathbb{P}(m \textit{ bad news}|R_B)\mathbb{P}(R_B)} \\ &= \frac{(1-p_{T^*})e^{-\lambda_b^B T^*}(\lambda_b^B)^m \frac{T^{*m}}{m!}}{p_{T^*}e^{-\lambda_b^G T^*}(\lambda_b^G)^m \frac{T^{*m}}{m!} + (1-p_{T^*})e^{-\lambda_b^B T^*}(\lambda_b^B)^m \frac{T^{*m}}{m!}} \\ &= \frac{(1-p_{T^*})e^{-\lambda_b^B T^*}(\lambda_b^B)^m}{p_{T^*}e^{-\lambda_b^G T^*}(\lambda_b^G)^m + (1-p_{T^*})e^{-\lambda_b^B T^*}(\lambda_b^B)^m}\end{aligned}$$

Hence, it is easy to see that $\mathbb{P}_{W/O}(\textit{Arm 1 is bad}|m \textit{ bad news})$ is less than $\mathbb{P}_W(\textit{Arm 1 is bad}|m \textit{ bad news})$ i.e. our proposed mechanism identifies the Arm 1 to be bad more precisely.
Hence proved. ■

5. Conclusion

Our paper focuses on the principal who wants to engage an agent for a low-rewarding and unknown project. For this purpose, we design an optimal incentive scheme in a continuous-time dynamic principal-agent framework with a two-armed bandit (Arm 0 and Arm 1). We frame our model in a hybrid news environment, where breakthroughs and breakdowns can be achieved in both known (Arm 0) and unknown (Arm 1) arms. Hence, it is difficult for the principal to distinguish whether an agent achieves the reported breakthroughs and breakdowns through cheating or honest experimentation. For this purpose, we construct a threshold as a parameter of honesty and design an optimal incentive scheme through which the principal compensates the agent for the honest experimentation and punishes him for the fake successes he achieves. However, we show that our incentive scheme works well in such a hybrid news framework and ensures the full honesty of the agent. In addition, along the equilibrium path, we also obtain an optimal deadline that supports our foremost requirement of honest experimentation on low-rewarding projects.

Our result does not depend on the risk-neutrality of the agent. The threshold at which the agent chooses to switch to the known arm continues to be the same even if we consider the risk aversion in our framework. Although under the risk-aversion, we see the changes in the value function as compared to the risk-neutrality of the agent.

Appendix

A. Detailed Proofs for Agent's Problem

A.1 Proof of Proposition 1

The proof is by a standard verification argument. The expression for p_1^* and the constant of integration are obtained by imposing $V_1^*(p_1^*) = (R_0 + C_0)$ (value matching) and $(V_1^*)'(p_1^*) = 0$ (smooth pasting). The existence of p_1^* (i.e. $p_1^* > 0$) can be established from the following Lemma 1.
We first state the Lemma and establish the existence of p_1^*, i.e. the threshold belief is positive.

Lemma **1.** *For the Agent's Problem without Incentive, the threshold belief p_1^* exists and it is positive.*

Proof: The expression for p_1^* and the constant of integration are obtained by imposing $V_1^*(p_1^*) = (R_0 + C_0)$ (value matching) and $(V_1^*)'(p_1^*) = 0$ (smooth pasting). Thus the expression for p_1^* is as follows

$$p_1^* = \frac{\mu(R_0 + C_0 - \lambda_b^B h_b)}{R_1(1+\mu) + (1+\mu)\lambda_b^G h_b - \mu\lambda_b^B h_b - (1+\mu)\frac{\lambda_g^G \lambda_b^G h_b}{r+\lambda_g^G} - (R_0 + C_0)}$$

From the assumption (A.2), we have

$$\begin{aligned} \lambda_b^G > \gamma_b &\implies \lambda_b^G |h_b| > \gamma_b |h_b| > \gamma_b |s_b| \\ &\implies -\lambda_b^G h_b > -\gamma_b s_b = -C_0 \\ &\implies C_0 - \lambda_b^G h_b > 0 \end{aligned}$$

Since, R_0 and μ is positive, $\mu(R_0 + C_0 - \lambda_b^B h_b)$ is positive. Hence, the numerator is positive.

Now, it remains to check for the sign of the denominator for p_1^*. The expression for the denominator is as follows

$$R_1(1+\mu) + (1+\mu)\lambda_b^G h_b - \mu\lambda_b^B h_b - (1+\mu)\frac{\lambda_g^G \lambda_b^G h_b}{r+\lambda_g^G} - (R_0 + C_0)$$

Rearranging, we have

$$(1+\mu)(R_1 - R_0) + (1+\mu)\left[\frac{r\lambda_b^G h_b}{r+\lambda_g^G} - C_0\right] + \mu(R_0 + C_0 - \lambda_b^B h_b) \tag{11}$$

From the assumption (A.1), we have $R_1 - R_0$ is positive, and we have shown that $\mu(R_0 + C_0 - \lambda_b^B h_b)$ is positive.
Define a function,

$$f(\lambda_b^G) := \frac{r\lambda_b^G h_b}{r + \lambda_g^G} - C_0$$

Since h_b is negative, $f'(p)$ is negative. Hence $f(p)$ is a monotonically decreasing function. Hence there exists a $\bar{\lambda}_b^G$ such that $f(\bar{\lambda}_b^G) = 0$, also $f(\lambda) > 0$ for all $\lambda \in (0, \bar{\lambda}_b^G)$, where, $\bar{\lambda}_b^G$ is expressed as follows

$$\bar{\lambda}_b^G = \frac{\gamma_b s_b(r + \lambda_g^G)}{rh_b} = \frac{C_0(r + \lambda_g^G)}{rh_b}$$

Now, from the assumption (A.4), $\lambda_b^G \in (0, \bar{\lambda}_b^G)$, implies $f(\lambda_b^G) > 0$. So the expression (11) is positive and hence, p_1^* exists and positive.

We have established the existence of the threshold belief p_1^*. V_1^* solves the Bellman equation, which can easily be verified.
It is easy to check $b(p, V_1^*)$ falls short of $c(p)$ to the left of p_1^*, coincides at p_1^* and exceeds right to the p_1^*. So V_1^* solves the Bellman equation as stated in the proposition. ■

A.2 Proof of Proposition 2

The proof is by a standard verification argument. The expression for $\bar{p}_1$ and the constant of integration are obtained by imposing $\bar{V}_1(\bar{p}_1) = R_0 + C_0 + T_2(0)$ (value matching) and $(\bar{V}_1)'(\bar{p}_1) = 0$ (smooth pasting). The existence of $\bar{p}_1$ (i.e. $\bar{p}_1 \geq 0$) can be established from the following Lemma 2.
We first state the Lemma and establish the existence of $\bar{p}_1$, i.e. the threshold belief is non-negative.

LEMMA **2.** *In the Agent's Problem with Incentive, under the condition* $|T_2(0)| > \mu T_1(1)$ *the threshold belief* $\bar{p}_1$ *exists and it is non-negative.*

PROOF: The expression for $\bar{p}_1$ and the constant of integration are obtained by imposing $\bar{V}_1(\bar{p}_1) = R_0 + C_0 + T_2(0)$ (value matching) and $(\bar{V}_1)'(\bar{p}_1) = 0$ (smooth pasting). Thus the expression for $\bar{p}_1$ is as follows

$$\bar{p}_1 = \frac{\mu\Big[R_0 + C_0 - \lambda_b^B h_b + T_2(0) - T_1(1)\Big]}{R_1(1+\mu) + (1+\mu)\lambda_b^G h_b - \mu\lambda_b^B h_b - (1+\mu)\frac{\lambda_g^G \lambda_b^G h_b}{r+\lambda_g^G} + \frac{rT_1(1)}{r+\lambda_g^G} - \mu T_1(1) - (R_0 + C_0) - T_2(0)}$$

If the following condition holds

$$R_0 + C_0 - \lambda_{1b}^B h_b \geq T_1(1) - T_2(0)$$

the numerator is non-negative. Now It remains to check for the sign of the denominator for $\bar{p}_1$. The expression for the denominator is as follows

$$R_1(1+\mu)+(1+\mu)\lambda_b^G h_b - \mu\lambda_b^B h_b - (1+\mu)\frac{\lambda_g^G \lambda_b^G h_b}{r+\lambda_g^G} + \frac{rT_1(1)}{r+\lambda_g^G} - \mu T_1(1) - (R_0+C_0) - T_2(0) \tag{12}$$

We have proved in Lemma 1, the following expression is positive i.e.

$$R_1(1+\mu) + (1+\mu)\lambda_b^G h_b - \mu\lambda_b^B h_b - (1+\mu)\frac{\lambda_g^G \lambda_b^G h_b}{r+\lambda_g^G} - (R_0 + C_0) > 0$$

Under the condition $|T_2(0)| > \mu T_1(1)$, the following expression is positive, i.e.

$$-\mu T_1(1) - T_2(0) > 0$$

Since, $T_1(1)$ is positive, the expression of the denominator is positive. Hence the threshold belief $\bar{p}_1$ exists and is non-negative.

Now we have established the existence of the threshold belief $\bar{p}_1$. $\bar{V}_1$ solves the Bellman equation can easily be verified. Now, it is easy to check $\bar{b}(p, \bar{V}_1)$ falls short of $\bar{c}(p)$ to the left of $\bar{p}_1$, coincides at $\bar{p}_1$ and exceeds right to the $\bar{p}_1$. So $\bar{V}_1$ solves the Bellman equation as stated in the proposition. ■

A.3 Proof of Proposition 3

We have,

$$\bar{p}_1 = \frac{\mu\Big[R_0 + C_0 - \lambda_b^B h_b + T_2(0) - T_1(1)\Big]}{R_1(1+\mu) + (1+\mu)\lambda_b^G h_b - \mu\lambda_b^B h_b - (1+\mu)\frac{\lambda_g^G \lambda_b^G h_b}{r+\lambda_g^G} + \frac{rT_1(1)}{r+\lambda_g^G} - \mu T_1(1) - (R_0+C_0) - T_2(0)}$$

and

$$p_1^* = \frac{\mu(R_0 + C_0 - \lambda_b^B h_b)}{R_1(1+\mu) + (1+\mu)\lambda_b^G h_b - \mu\lambda_b^B h_b - (1+\mu)\frac{\lambda_g^G \lambda_b^G h_b}{r+\lambda_g^G} - (R_0+C_0)}$$

In the Subsection 3.2, we have argued that $T_2(0)$ is negative and $T_1(1)$ is positive . Define,

$$X = \frac{rT_1(1)}{r + \lambda_{1g}^G} - \mu T_1(1) - T_2(0)$$

Since, $T_1(1)$ is positive and from Lemma 2, we have $-\mu T_1(1) - T_2(0)$ is positive, hence, X is positive. Now,

$$\begin{aligned} p_1^* &= \frac{\mu(R_0 + C_0 - \lambda_b^B h_b)}{R_1(1+\mu) + (1+\mu)\lambda_b^G h_b - \mu\lambda_b^B h_b - (1+\mu)\frac{\lambda_g^G \lambda_b^G h_b}{r+\lambda_g^G} - (R_0 + C_0)} \\ &> \frac{\mu(R_0 + C_0 - \lambda_b^B h_b) + \mu T_2(0) - \mu T_1(1)}{R_1(1+\mu) + (1+\mu)\lambda_b^G h_b - \mu\lambda_b^B h_b - (1+\mu)\frac{\lambda_g^G \lambda_b^G h_b}{r+\lambda_g^G} - (R_0 + C_0)} \\ &> \frac{\mu(R_0 + C_0 - \lambda_b^B h_b) + \mu T_2(0) - \mu T_1(1)}{R_1(1+\mu) + (1+\mu)\lambda_b^G h_b - \mu\lambda_b^B h_b - (1+\mu)\frac{\lambda_g^G \lambda_b^G h_b}{r+\lambda_g^G} - (R_0 + C_0) + X} = \bar{p}_1 \end{aligned}$$

Hence, $\bar{p}_1 < p_1^*$, Proved.

B. Mathematical Appendix

B.1 Assumptions

1. Basic Assumptions:

   (a) $\gamma_g > \gamma_b$

   (b) $R_0, R_1 > 0$ with $R_1 > R_0$

   (c) $s_b, h_b < 0$ with $|h_b| > |s_b|$

   (d) $\lambda_g^G, \lambda_b^G, \lambda_b^B > 0$ with $\lambda_b^B > \lambda_b^G$

   (e) $\lambda_b^G > \gamma_b$

B.2 Check the Sign for the expressions with unknown parameters

I. Sign$(T_1)^2(k_1) + (T_2)^2(k_1)$

We know,

$$T_1(k_1) = e^{-rT}\left(\sum_{i=1}^{\frac{m}{1+\bar{\lambda}}} e^{-[(1-k_1)\gamma_g + \lambda_g^G k_1]T} \frac{[(1-k_1)\gamma_g + \lambda_g^G k_1]^i\, T^i}{i!}\right)(\lambda - \bar{\lambda})N_b V(1)$$

For simplicity of calculation, define, $(1-k_1)\gamma_g + \lambda_g^G k_1 = A(k_1)$, $\frac{m}{1+\bar{\lambda}} = \bar{m}$, $e^{-rT}(\lambda - \bar{\lambda})N_b V(1) = B$
Hence,

$$T_1(k_1) = Be^{-A(k_1)T}\sum_{i=1}^{\bar{m}} \frac{A(k_1)^i T^i}{i!} \tag{13}$$

Now, differentiating w.r.t k_1, we have,

$$(T_1)^1(k_1) = Be^{-A(k_1)T}(-A'(k_1)T)\sum_{i=1}^{\bar{m}} \frac{A(k_1)^i T^i}{i!} + Be^{-A(k_1)T}A'(k_1)T\sum_{i=0}^{\bar{m}-1} \frac{A(k_1)^i T^i}{i!}$$

Simplifying,

$$(T_1)^1(k_1) = -Be^{-A(k_1)T}A'(k_1)T\left[\frac{A(k_1)^{\bar{m}}T^{\bar{m}}}{\bar{m}!} - 1\right] \tag{14}$$

Again, differentiating w.r.t k_1, we have,

$$(T_1)^2(k_1) = Be^{-A(k_1)T}(A'(k_1)T)^2\left[\frac{A(k_1)^{\bar{m}}T^{\bar{m}}}{\bar{m}!} - 1\right] - Be^{-A(k_1)T}(A'(k_1)T)^2\left[\frac{A(k_1)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!}\right]$$

Simplifying,

$$(T_1)^2(k_1) = Be^{-A(k_1)T}(A'(k_1)T)^2\left[\frac{A(k_1)^{\bar{m}}T^{\bar{m}}}{\bar{m}!} - \frac{A(k_1)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!} - 1\right] \tag{15}$$

Also,

$$(T_2)(k_1) = e^{-rT}\left(\sum_{i=\bar{m}+1}^{m} e^{-[(1-k_1)\gamma_g+\lambda_g^G k_1]T}\frac{[(1-k_1)\gamma_g + \lambda_g^G k_1]^i\, T^i}{i!}\right)(\lambda-\bar{\lambda})N_g V(0)$$

For simplicity of calculation, let's define, $C = e^{-rT}(\lambda-\bar{\lambda})N_g V(0)$
Hence,

$$(T_2)(k_1) = Ce^{-A(k_1)T}\sum_{i=\bar{m}+1}^{m}\frac{A(k_1)^i T^i}{i!} \tag{16}$$

differentiating w.r.t k_1,

$$(T_2)^1(k_1) = Ce^{-A(k_1)T}(-A'(k_1)T)\sum_{i=\bar{m}+1}^{m}\frac{A(k_1)^i T^i}{i!} + Ce^{-A(k_1)T}A'(k_1)T\sum_{i=\bar{m}}^{m-1}\frac{A(k_1)^i T^i}{i!}$$

Simplifying,

$$(T_2)^1(k_1) = -Ce^{-A(k_1)T}A'(k_1)T\left[\frac{A(k_1)^m T^m}{m!} - \frac{A(k_1)^{\bar{m}}T^{\bar{m}}}{\bar{m}!}\right] \tag{17}$$

Again, differentiating w.r.t k_1, we have,

$$(T_2)^2(k_1) = Ce^{-A(k_1)T}(A'(k_1)T)^2\left[\frac{A(k_1)^m T^m}{m!} - \frac{A(k_1)^{\bar{m}}T^{\bar{m}}}{\bar{m}!}\right] - Ce^{-A(k_1)T}(A'(k_1)T)^2\left[\frac{A(k_1)^{m-1}T^{m-1}}{m-1!} - \frac{A(k_1)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!}\right]$$

Simplifying,

$$(T_2)^2(k_1) = Ce^{-A(k_1)T}(A'(k_1)T)^2\left[\left(\frac{A(k_1)^m T^m}{m!} - \frac{A(k_1)^{m-1}T^{m-1}}{(m-1)!}\right) - \left(\frac{A(k_1)^{\bar{m}}T^{\bar{m}}}{\bar{m}!} - \frac{A(k_1)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!}\right)\right] \quad (18)$$

Summarizing $(T_1)^2(k_1)$ and $(T_2)^2(k_1)$

$$(T_1)^2(k_1) = Be^{-A(k_1)T}(A'(k_1)T)^2\left[\frac{A(k_1)^{\bar{m}}T^{\bar{m}}}{\bar{m}!} - \frac{A(k_1)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!} - 1\right] \quad (19)$$

$$(T_2)^2(k_1) = Ce^{-A(k_1)T}(A'(k_1)T)^2\left[\left(\frac{A(k_1)^m T^m}{m!} - \frac{A(k_1)^{m-1}T^{m-1}}{(m-1)!}\right) - \left(\frac{A(k_1)^{\bar{m}}T^{\bar{m}}}{\bar{m}!} - \frac{A(k_1)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!}\right)\right] \quad (20)$$

Define,

$$\left[\frac{A(k_1)^{\bar{m}}T^{\bar{m}}}{\bar{m}!} - \frac{A(k_1)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!}\right] = I$$

and

$$\left[\frac{A(k_1)^m T^m}{m!} - \frac{A(k_1)^{m-1}T^{m-1}}{(m-1)!}\right] = II$$

Hence, we have,

$$(T_1)^2(k_1) = Be^{-A(k_1)T}(A'(k_1)T)^2[I-1]$$

and

$$(T_2)^2(k_1) = Ce^{-A(k_1)T}(A'(k_1)T)^2[II-I]$$

Now, we need to determine the sign of B and C.
Recall that the term B is the reward for the honest agent for pulling "Arm 1", throughout his experiment and denoted by $T_1(1)$ and the term C is the punishment for an off-path agent for not experimenting, and denoted by $T_2(0)$. Also, in Subsection 3.2, we have argued that $T_1(1)$ is positive and $T_2(0)$ is negative. Hence, we can conclude that B is positive and C is negative.

$$B = e^{-rT}(\lambda - \bar{\lambda})N_b V(1) = +ve$$

$$C = e^{-rT}(\lambda - \bar{\lambda})N_g V(0) = -ve$$

Now,

$$(T_1)^2(k_1) + (T_2)^2(k_1) = \underbrace{Be^{-A(k_1)T}(A'(k_1)T)^2}_{+ve}[I-1] + \underbrace{Ce^{-A(k_1)T}(A'(k_1)T)^2}_{-ve}[II-I] \tag{21}$$

So, for $(T_1)^2(k_1) + (T_2)^2(k_1)$ to be $+ve$, we need,

$$I - 1 > 0 \implies I > 1$$

$$II - I < 0 \implies I > II$$

A. Condition for $I > 1$

Given the assumption

$$\gamma_g > \lambda_g^G$$

for any $k_1 \in (0, 1)$, we have

$$\gamma_g > (1-k_1)\gamma_g + k_1\lambda_g^G \implies \gamma_g > A(k_1)$$

and

$$(1-k_1)\gamma_g + k_1\lambda_g^G > \lambda_g^G \implies A(k_1) > \lambda_g^G$$

Thus, we have,

$$\frac{(\gamma_g)^{\bar{m}}T^{\bar{m}}}{\bar{m}!} > \frac{A(k_1)^{\bar{m}}T^{\bar{m}}}{\bar{m}!} \tag{22}$$

and,

$$\frac{A(k_1)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!} > \frac{(\lambda_g^G)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!}$$

$$\implies -\frac{(\lambda_g^G)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!} > -\frac{A(k_1)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!} \tag{23}$$

Adding equation (22) and (23), we have

$$\frac{(\gamma_g)^{\bar{m}}T^{\bar{m}}}{\bar{m}!} - \frac{(\lambda_g^G)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!} > \frac{A(k_1)^{\bar{m}}T^{\bar{m}}}{\bar{m}!} - \frac{A(k_1)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!} = I > 1 \tag{24}$$

Hence, for $I > 1$, we must have

$$\frac{(\gamma_g)^{\bar{m}}T^{\bar{m}}}{\bar{m}!} - \frac{(\lambda_g^G)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!} > 1 \tag{25}$$

B. Condition for $I > II$

From the equation (24), we have,

$$\frac{(\gamma_g)^{\bar{m}}T^{\bar{m}}}{\bar{m}!} - \frac{(\lambda_g^G)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!} > \frac{A(k_1)^{\bar{m}}T^{\bar{m}}}{\bar{m}!} - \frac{A(k_1)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!} = I > 1$$

Now, proceeding as above, we have,

$$\frac{A(k_1)^m T^m}{m!} > \frac{(\lambda_g^G)^m T^m}{m!} \tag{26}$$

and,

$$\frac{(\gamma_g)^{m-1}T^{m-1}}{(m-1)!} > \frac{A(k_1)^{m-1}T^{m-1}}{(m-1)!}$$

$$\implies -\frac{A(k_1)^{m-1}T^{m-1}}{(m-1)!} > -\frac{(\gamma_g)^{m-1}T^{m-1}}{(m-1)!} \tag{27}$$

Adding the equations (26) and (27), we have,

$$II = \frac{A(k_1)^m T^m}{m!} - \frac{A(k_1)^{m-1}T^{m-1}}{(m-1)!} > \frac{(\lambda_g^G)^m T^m}{m!} - \frac{(\gamma_g)^{m-1}T^{m-1}}{(m-1)!} \tag{28}$$

Now, from the equations (24) and (28), we have,

$$\frac{(\gamma_g)^{\bar{m}}T^{\bar{m}}}{\bar{m}!} - \frac{(\lambda_g^G)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!} > I > II > \frac{(\lambda_g^G)^m T^m}{m!} - \frac{(\gamma_g)^{m-1}T^{m-1}}{(m-1)!}$$

Hence, for $I > II$, we have,

$$\frac{(\gamma_g)^{\bar{m}}T^{\bar{m}}}{\bar{m}!} - \frac{(\lambda_g^G)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!} > \frac{(\lambda_g^G)^m T^m}{m!} - \frac{(\gamma_g)^{m-1}T^{m-1}}{(m-1)!} \tag{29}$$

Now, conditions for $I > 1$ and $I > II$ are stated as follows

$$\frac{(\gamma_g)^{\bar{m}}T^{\bar{m}}}{\bar{m}!} - \frac{(\lambda_g^G)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!} > 1 \tag{30}$$

$$\implies \frac{(\gamma_g)^{\bar{m}}T^{\bar{m}}}{\bar{m}!} - \frac{(\lambda_g^G)^{\bar{m}-1}T^{\bar{m}-1}}{(\bar{m}-1)!} > \frac{(\lambda_g^G)^m T^m}{m!} - \frac{(\gamma_g)^{m-1}T^{m-1}}{(m-1)!} \tag{31}$$

Hence, under the conditions (30) and (31), the expression $(T_1)^2(k_1) + (T_2)^2(k_1)$ is positive.

B.3 Conditions for Interior Solution

Define,

$$f(k_1) = (1-k_1)(R_0+C_0)+k_1\Bigg\{pR_1+\lambda(p)h_b+\frac{1}{r}\Bigg[p\lambda_g^G[R_1-u(p)]+\lambda(p)[u(j(p))-u(p)]+ p(1-p)[\lambda_b^B-\lambda_b^G-\lambda_g^G]u'(p)\Bigg]\Bigg\}+T_1(k_1)+T_2(k_1)$$

For simplicity define,

$$R_0 + C_0 = A$$

and,

$$pR_1+\lambda(p)h_b+\frac{1}{r}\Bigg[p\lambda_g^G[R_1-u(p)]+\lambda(p)[u(j(p))-u(p)]+p(1-p)[\lambda_b^B-\lambda_b^G-\lambda_g^G]u'(p)\Bigg] = B$$

Hence,

$$f(k_1) = (1-k_1)A + k_1B + T_1(k_1) + T_2(k_1) \tag{32}$$

Taking differentiation w.r.t k_1, we get,

$$f^1(k_1) = -A + B + (T_1)^1(k_1) + (T_2)^1(k_1) \tag{33}$$

$$f^2(k_1) = (T_1)^2(k_1) + (T_2)^2(k_1) \tag{34}$$

where the superscript $i = 1, 2$ denotes the i-th derivative of the function.

Now, check for the sign of $(T_1)^2(k_1) + (T_2)^2(k_1)$. (See Appendix B.2). From the detailed discussion from Appendix B.2, we have that under the conditions (30) and (31), $f(k_1)$ is strictly convex and no interior solution exists. Hence proved.

References

[1] Dirk Bergemann and Ulrich Hege. Venture capital financing, moral hazard, and learning. *Journal of Banking & Finance*, 22(6-8):703–735, August 1998.

[2] Dirk Bergemann and Ulrigh Hege. The Financing of Innovation: Learning and Stopping. *RAND Journal of Economics*, 36(4):719–752, Winter 2005.

[3] Yingni Guo. Dynamic Delegation of Experimentation. *American Economic Review*, 106(8):1969–2008, August 2016.

[4] Marina Halac, Navin Kartik, and Qingmin Liu. Optimal Contracts for Experimentation. *Review of Economic Studies*, 83(3):1040–1091, 2016.

[5] Sinem Hidir. Contracting for Experimentation and the Value of Bad News. *Working paper*.

[6] Johannes Horner and Larry Samuelson. Incentives for Experimenting Agents. *RAND Journal of Economics*, 44(4):632–663, Winter 2013.

[7] Godfrey Keller and Sven Rady. Strategic experimentation with Poisson bandits. *Theoretical Economics*, 5(2):275–311, 2010.

[8] Godfrey Keller and Sven Rady. Breakdowns. *Theoretical Economics*, 10(1):175–202, 2015.

[9] Godfrey Keller, Sven Rady, and Martin Cripps. Strategic Experimentation with Exponential Bandits. *Econometrica*, 73(1):39–68, January 2005.

[10] Nicolas Klein. The Importance of Being Honest. *Theoretical Economics*, 11(3), September 2016.

[11] Aditya Kuvalekar and Nishant Ravi. Rewarding Failure. *Working paper*.

[12] Statistica.com. Number of smartphone users in the world.