The Importance of Being Honest*

Nicolas Klein†

This version: June 15, 2010
Preliminary and Incomplete

Abstract

I analyze the case of a principal who wants to give an agent proper incentives to investigate a hypothesis which can be either true or false. The agent can shirk, thus never proving the hypothesis, or he can avail himself of a known technology to manipulate the data. If the hypothesis is false, a proper investigation never yields a success. I show that if, in the case the hypothesis is true, a proper investigation yields successes with a higher intensity than manipulation would, any optimal wage scheme leads to the first-best amount, and speed, of experimentation. In the opposite case, honest investigation is impossible to implement.

Keywords: Experimentation, Bandit Models, Poisson Process, Bayesian Learning, Principal-Agent Models, Optimal Incentive Scheme.

JEL Classification Numbers: C79, D82, D83, O32.

*I thank Johannes Hörner and Sven Rady for their advice, patience, and encouragement, as well as Dirk Bergemann, Tri-Vi Dang, Federico Echenique, Lucas Maestri, Thomas Mariotti, Andy Skrzypacz, Juuso Välimäki, Tom Wiseman for helpful comments and discussions. I am especially grateful to the Cowles Foundation for Research in Economics at Yale University for an extended stay during which the idea for this paper took shape. Financial support from the National Research Fund of Luxembourg is gratefully acknowledged.

†Munich Graduate School of Economics, Kaulbachstr. 45, D-80539 Munich, Germany; email: kleinnic@yahoo.com.

1 Introduction

Instances abound when a principal, e.g. society, is interested in the investigation of a certain hypothesis. Indeed, important policy decisions may depend on whether, say, there is a causal link between passively inhaling other people's cigarette smoke and the occurrence of cancer, or whether global warming trends are caused by certain emissions related to specific kinds of economic activity. Often, though, it will not be practical for "society" to carry out the necessary research itself; it will rather have to delegate the investigation to a group of scientists, or, as is the case in my model, to a single scientist. The problem with that, of course, is that this scientist will typically have interests of his own, some of which may even be endogenously generated by society's incentive scheme.

As is well known from the principal-agent literature, when an agent's actions cannot easily be monitored, his pay must be made contingent on his performance, so that he have proper incentives to exert effort. Thus, the scientist will only get paid, or will get paid a substantial bonus if, and only if, he proves his hypothesis. While this may well provide him with the necessary incentives to work, unfortunately, it might also give him incentives to fabricate, or manipulate, his data, in order to make it appear as though his hypothesis was proved. In a setting involving Bayesian learning on the agent's part, my model investigates how optimally to achieve the dual objective of providing the agent with the right incentives to work, while also making sure that he not be tempted to engage in manipulations and trickery, even if said manipulations were not verifiable in a court of law, or even completely unobservable. Alternatively, one could interpret my model as a model of technology adoption: An agent is hired expressly to test some new production method, or some new way of doing business, yet the boss cannot monitor whether the successes he observes are really due to the new method, or whether the agent has surreptitiously availed himself of an old established method to produce the observed results.

In my model, the agent can either shirk, in which case he will never have a success, but which gives him some flow benefit, or he can cheat, which gives him an apparent success according to some known distribution, or he can do the risky thing, and be honest. If the hypothesis is incorrect, honesty never yields a success. The principal can only observe if there has been a success or not; he cannot observe the agent's actions, and, in particular, he does not observe if a success has been achieved by honest means or whether it is the result of manipulation.

I show that if even the investigation of a correct hypothesis yields breakthroughs at a lower frequency than manipulation, honesty is not implementable at all. If, however, investigating a correct hypothesis yields breakthroughs at a higher intensity than manipulation, the

optimal incentive scheme I characterize will make sure that the agent is always honest up to the first breakthrough at least, and even leads to the first-best levels and speed of experimentation, so that there will be no deadweight loss from agency. Indeed, the optimal incentive scheme exactly compensates the agent for his outside option of shirking, which is precisely the relevant yardstick in the benchmark case where the principal conducts the investigation himself. Thus, there will be just as much exploration when the principal has to delegate the operation to an agent as there would be if he were in a position to conduct the investigation himself; indeed, both the amount, as well as the speed, of overall experimentation will be efficient.

While actually investigating the hypothesis, the agent increasingly grows pessimistic about the thesis being true as long as no breakthrough arrives. At the first breakthrough, though, all uncertainty is resolved, and the agent will know for sure that the hypothesis is true. Thus, depending on the incentive scheme, this learning aspect might give the agent an *experimentation* motive for using arm 1, i.e. he might be willing to forgo current payoffs in order to gather information which might then potentially be parlayed into higher payoffs come tomorrow. The principal himself has no learning motive as he is only interested in the *first* breakthrough achieved on arm 1; however, when designing the incentive scheme, it will be one of his goals to kindle the agent's experimentation motive, by endogenously making information valuable to him, as a way of providing incentives.

If honesty is implementable, I show that even though the principal is only interested in the *first* breakthrough the agent achieves, he will reward the agent for the $(m+1)$-st breakthrough, with $m \geq 1$, in order to deter the agent from engaging in manipulation, which otherwise might seem expedient to him in the short term. Now, m will be chosen high enough that even for an off-equilibrium agent, who has achieved his first breakthrough via manipulation, m breakthroughs are so unlikely to be achieved by cheating that he will prefer to be honest after his first breakthrough. This will put the cheating off-equilibrium agent at a distinct disadvantage, as, in contrast to the honest on-path agent, he will not have had a discontinuous jump in his belief. This difference in beliefs between on-equilibrium and off-equilibrium agents in turn can be leveraged by the principal, who enjoys full commitment power; thus, the principal can induce investigation of the hypothesis by endogenously creating a high value of information for the agent.

To provide adequate incentives in the cheapest way possible, the principal will endeavor to give the lowest possible value to a dishonest agent, given the continuation value he has promised the on-equilibrium agent. While paying only for the $(m+1)$-st breakthrough ensures that off-equilibrium agents will not continue to cheat, they will nevertheless continue to update their beliefs after their first success, and might be tempted to switch to shirking

once they have grown too pessimistic about the hypothesis, a possibility that, as is well known from the literature on strategic experimentation with bandits, gives them a positive option value. In order to reduce this additional option value, an optimal incentive scheme will make sure that the off-equilibrium agent will imitate the on-equilibrium agent to an arbitrarily large extent.

The rest of the paper is set up as follows: section 2 reviews some relevant literature; section 3 introduces the model; section 4 analyzes the provision of a certain continuation value; section 5 characterizes the optimal mechanism before the first breakthrough, section 6 analyzes when the principal will optimally elect to stop the project, and section 7 concludes.

2 Related Literature

Holmström & Milgrom (1991) analyze a case where, not unlike in my model, the agent performs several tasks, some of which may be undesirable from the principal's point of view. The principal may be able to monitor certain activities more accurately than others. They show that in the limiting case with two activities where one activity cannot be monitored at all, incentives will only be given for the activity which can in fact be monitored; if the activities are substitutes (complements) in the agent's private cost function, incentives are more muted (steeper) than in the single task case. While their model could be extended to a dynamic model where the agent controls the drift rate of a Brownian Motion signal,[1] the learning motive I introduce fundamentally changes the basic trade-offs involved. Indeed, in my model, the optimal mechanism extensively leverages the fact that only an honest agent will have had a discontinuous jump in his beliefs.

Bergemann & Hege (1998, 2005), as well as Hörner & Samuelson (2009) examine a venture capitalist's provision of funds for an investment project of initially uncertain quality; the project is managed by an entrepreneur, who might divert the funds for his private ends. The investor cannot observe the entrepreneur's allocation of the funds, so that, off-equilibrium, the entrepreneur may accumulate some private information about the quality of the project. If the project is good, it yields a success with a probability that is proportional to the amount of funds invested in it; if it is bad, it never yields a success. While Bergemann & Hege (2005) and Hörner & Samuelson (2009) analyze the game without commitment, Bergemann & Hege (1998) investigate the problem under full commitment. These papers differ from my model chiefly in that there is no way for the entrepreneur to "fake" a success; any success that is publicly observed will have been achieved by honest means alone.

[1] See Holmström & Milgrom (1987).

Gerardi & Maestri (2008) investigate the case of a principal who, in order to find out about the binary state of the world, has to employ an agent. The agent can decide to incur private costs to exert effort to acquire an informative binary signal, one realization of which is only possible in the good state. As for the principal, he can monitor neither the agent's effort choice nor the realization of the signal. The game ends as soon as the agent announces that he has had conclusive evidence in favor of the good state. They show that the agent needs to be left an information rent because of both the Moral Hazard and the Adverse Selection problems, suggesting there would tend to be a deadweight loss from agency. In my model, by contrast, the game does not end after the first breakthrough; much to the contrary, I show that in my model, in order to give optimal incentives, it is absolutely vital that they be provided via the continuation game that follows the first breakthrough rather than via an immediate transfer. As a matter of fact, this construction makes sure that there will be *no* deadweight loss from agency in my model.

One paper that is close in spirit to mine is Manso (2010), who analyzes a simple, undiscounted two-period, model, where an agent can either shirk, try to produce in some established manner with a known success probability, or experiment with a risky alternative. He shows that, in order to induce experimentation, the principal will optimally not pay for a success in the first period, and might even pay for early failure,[2] while a success in the second period is always rewarded. My continuous-time investigation confirms Manso's (2010) central intuition that it is better to give incentives through later rewards; furthermore, the richer action and signal spaces in my fully-fledged dynamic model yield additional insights into the structure of the optimal incentive scheme. Moreover, the dynamic structure allows me to analyze the principal's optimal stopping time, and to conclude that the overall amount and speed of experimentation will be efficient, whenever honesty is implementable at all.

De Marzo & Sannikov (2008) also incorporate private learning on the agent's part into their model, where current output depends both on the firm's inherent profitability and on the agent's effort, which is unobservable to the principal. Thus, off-equilibrium, the agent's private belief about the firm's productivity will differ from the public belief. Specifically, if the agent withholds effort, this depresses the drift rate of the firm's Brownian motion cash flow. They show that the firm will optimally accumulate cash as fast as it can until it reaches some target level, after which it starts paying out dividends; the firm is liquidated as soon as it runs out of cash. De Marzo & Sannikov (2008) show that one optimal way of providing

[2]This is an artefact of the discrete structure of the model and the limited signal space; indeed, in Manso's (2010) model, early failure can be a very informative signal that the agent has not exploited the known technology, but has rather chosen the risky, unknown alternative. In continuous time, by contrast, arbitrary precision of the signal can be achieved by choosing a critical number of successes that is high enough, as will become clear *infra*.

incentives is to give the agent an equity stake in the firm, which is rescindable in the case of liquidation, and that liquidation decisions are efficient, agency problems notwithstanding.

To capture the learning aspect of the agent's problem, I model it as a bandit problem.[3] Bandit problems have been used in economics to study the trade-off between experimentation and exploitation since Rothschild's (1974) discrete-time single-agent model. The single-agent two-armed exponential model, a variant of which I am using, has first been analyzed by Presman (1990). Strategic interaction among several agents has been analyzed in the models by Bolton & Harris (1999, 2000), Keller, Rady, Cripps (2005), Keller & Rady (2010), who all investigate the case of perfect positive correlation between players' two-armed bandit machines, as well as by Klein & Rady (2010), who investigate the cases of perfect, as well as imperfect, negative correlation. Klein (2010) analyzes the case where bandits have three arms, with the two risky ones being perfectly negatively correlated. While the afore-mentioned papers all assumed that players's actions, as well as the outcomes of their actions, were perfectly publicly observable, Rosenberg, Solan, Vieille (2007), as well as Murto & Välimäki (2009), analyze the case where actions are observable, while outcomes are not. Bonatti & Hörner (2010) analyze the case where actions are not observable, while outcomes are. Bergemann & Välimäki (1996, 2000) consider strategic experimentation in buyer-seller interactions. My contribution to this literature is to introduce the question of optimal incentive provision into a fully-fledged dynamic bandit model.

Rahman (2009, 2010) deals with the question of implementability in dynamic contexts, and finds that, under a full support assumption, a necessary and sufficient condition for implementability is for all non-detectable deviations to be unprofitable under zero transfers. The issue of implementability turns out to be quite simple in my model, and is dealt with in proposition 3.1.

3 The Model

There is one principal and one agent. The agent operates a bandit machine with three arms, i.e. one safe arm yielding the agent a private benefit flow of s, one that is known to yield breakthroughs according to $Po(\lambda_0)$ (arm 0), and arm 1, which either yields breakthroughs according to $Po(\lambda_1)$ (if the time-invariant state of the world $\theta = 1$, which is the case with initial probability $p_0 \in]0,1[$) or never yields a breakthrough (if the state is $\theta = 0$). It is commonly known that $\lambda_1, \lambda_0 > 0$. The principal only observes if, and at what time, there has been a breakthrough; he does not observe on which arm the breakthrough has been

[3] See Bergemann & Välimäki (2008) for an overview of this literature.

achieved. The agent in addition observes on which arm the breakthroughs have occurred. The principal and the agent share a common discount rate r.

The principal, only being interested in the first breakthrough *achieved on arm 1*, chooses and end date $\check{T}(t) \in [t, \overline{T}]$ (where $\overline{T} \in]T, \infty[$ is arbitrary), in case the first breakthrough occurs at time t. Conditional on there having been no breakthrough, the game ends at time $t < \infty$. I the first half of this paper, I take T to be exogenously given. In the second half, the principal optimally chooses the end date T.[4] There, I shall assume that the first breakthrough achieved on arm 1 at time t gives the principal a payoff of $e^{-rt}\Pi$.

Formally, I consider the point processes $\{N_t^i\}_{0\leq t\leq \overline{T}}$ (for $i \in \{0,1\}$), where N_t^i measures the number of breakthroughs achieved on arm i up to, and including, time t. In addition, I define the point process $\{N_t\}_{0\leq t\leq \overline{T}}$, where $N_t := N_t^0 + N_t^1$ for all t. Moreover, I consider the filtrations $\mathfrak{F}^{N^{0+1}} := \left\{\mathfrak{F}_t^{N^{0+1}}\right\}_{0\leq t\leq \overline{T}}$ and $\mathfrak{F}^N := \left\{\mathfrak{F}_t^N\right\}_{0\leq t\leq \overline{T}}$ generated by the processes $\{(N_t^0, N_t^1)\}_{0\leq t\leq \overline{T}}$ and $\{N_t\}_{0\leq t\leq \overline{T}}$, respectively.

By choosing which arm to pull, the agent affects the probability of breakthroughs on his several arms. Specifically, if he commits a constant fraction k_0 of his unit endowment flow to arm 0 over a time interval of length $\Delta > 0$, the probability of achieving at least one breakthrough on arm 0 in that interval is given by $1 - e^{-\lambda_0 k_0 \Delta}$. If he commits a constant fraction of k_1 of his endowment to arm 1 over a time interval of length $\Delta > 0$, the probability of achieving at least one breakthrough on arm 1 in that interval is given by $\theta\left(1 - e^{-\lambda_1 k_1 \Delta}\right)$.

Formally, a strategy for the agent is a process $\{(k_{0,t}, k_{1,t})\}_t$ which satisfies $(k_{0,t}, k_{1,t}) \in \{(a,b) \in \mathbb{R}_+ : a + b \leq 1\}$ for all t, and is $\mathfrak{F}^{N^{0+1}}$-predictable, where $k_{i,t}$ ($i \in \{0,1\}$) denotes the fraction of the agent's resource that he devotes to arm i at instant t. The agent's strategy space, which I denote by $\mathcal{U}^A$, is given by all the processes $\{(k_{0,t}, k_{1,t})\}_t$ satisfying these requirements.

A *wage scheme* offered by the principal is a non-negative, non-decreasing process $\{\mathcal{W}_t\}_{0\leq t\leq \overline{T}}$ which is $\mathfrak{F}^N$-adapted, where $\mathcal{W}_t$ denotes the discounted time 0 value of the cumulated payments the principal has consciously made to the agent up to, and including, time t. I assume the agent is protected by limited liability; hence $\{\mathcal{W}_t\}_{0\leq t\leq \overline{T}}$ is non-negative and non-decreasing.[5] I furthermore assume that the principal has full commitment power, i.e. he commits to a wage scheme $\{\mathcal{W}_t\}_{0\leq t\leq \overline{T}}$, as well as a schedule of end dates $\{\check{T}(t)\}_{t\in[0,T]}$, at the outset of the game.

[4] I am essentially following Grossman & Hart's (1983) classical approach to principal-agent problems in that I first solve for the optimal incentive scheme given an arbitrary T (sections 4 and 5), and then let the principal optimize over T (section 6).

[5] If the game ends at time $\check{T} < \overline{T}$, we set $\mathcal{W}_{\check{T}+\Delta} = \mathcal{W}_{\check{T}}$ for all $\Delta > 0$.

Over and above the payments he gets as a function of breakthroughs, the agent can secure himself a safe payoff flow of s from the principal by pulling the safe arm; the principal, however, can do nothing about this, and only observes it after the end of the game. The idea is that society cannot observe its scientists shirking in real time, as it were; only after the lab e.g. is shut down, such information might come to light, and society will only learn *ex post* that it has been robbed of the payoff flow of s during the operation of the research lab.

It is the principal's goal to induce the agent to use arm 1 at least up to the first breakthrough, and to do so in the most cost-efficient manner possible. Thus, I shall refer to $\mathcal{K} := \left\{ \{(k_{0,t}, k_{1,t})\}_t \in \mathcal{U}^A : N_t = 0 \Rightarrow k_{1,t} = 1 \right\}$ as the set of *incentive compatible* strategies. Clearly, as it is the principal's goal to get the agent to exert effort in order to achieve a breakthrough, it is never a good idea for him to pay the agent in the absence of a breakthrough; as the principal is only interested in the first breakthrough, the notation can be simplified somewhat. Let $\{\mathcal{W}_t\}_{0 \leq t \leq \overline{T}}$ be the principal's wage scheme, and t the time of the first breakthrough: In the rest of the paper, I shall write $h_t := e^{rt} \left(\mathcal{W}_t - \lim_{\tau \uparrow t} \mathcal{W}_\tau \right)$ for the instantaneous lump sum the principal pays the agent as a reward for his first breakthrough. By w_t I denote the expected continuation value of an agent who has achieved his first breakthrough on arm 1 at time t; formally,

$$w_t := \sup_{\{(k_{0,\tau}, k_{1,\tau})\}_{t < \tau \leq \check{T}(t)}} E\left[e^{rt} \left(\mathcal{W}_{\check{T}(t)} - \mathcal{W}_t \right) + s \int_t^{\check{T}(t)} e^{-r(\tau - t)} \left(1 - k_{0,\tau} - k_{1,\tau} \right) d\tau \, | \mathfrak{F}_t^{N^{0+1}}, N_t^1 = 1, \lim_{\tau \uparrow t} N_\tau^1 = 0, N_t^0 = 0, \{(k_{0,\tau}, k_{1,\tau})$$

The expected continuation payoff of an off-equilibrium agent, who achieves his first breakthrough on arm 0 at time t, I denote by ω_t; formally,

$$\omega_t := \sup_{\{(k_{0,\tau}, k_{1,\tau})\}_{t < \tau \leq \check{T}(t)}} E\left[e^{rt} \left(\mathcal{W}_{\check{T}(t)} - \mathcal{W}_t \right) + s \int_t^{\check{T}(t)} e^{-r(\tau - t)} \left(1 - k_{0,\tau} - k_{1,\tau} \right) d\tau \, | \mathfrak{F}_t^{N^{0+1}}, N_t^0 = 1, \lim_{\tau \uparrow t} N_\tau^0 = 0, N_t^1 = 0, \{(k_{0,\tau}, k_{1,\tau})$$

The state of the world is uncertain; clearly, whenever the agent uses arm 1, he gets new information about its quality; this *learning* is captured in the evolution of his (private) belief $\hat{p}_t$ that arm 1 is good. Formally, $\hat{p}_t \equiv E\left[\theta | \mathfrak{F}_t^{N^{0+1}} \right]$. On the equilibrium path, the principal will correctly anticipate $\hat{p}_t$; formally, $p_t = \hat{p}_t$, where p_t is defined by $p_t := E\left[\hat{p}_t | \mathfrak{F}_t^N, \{(k_{0,t}, k_{1,t})\}_t \in \mathcal{K} \right]$. In the following, I shall write p_t whenever $p_t = \hat{p}_t$, even when analyzing the agent's optimization problem.

The evolution of beliefs is easy to describe, since only a good arm 1 can ever yield a breakthrough. As the agent will always operate arm 1 until the first breakthrough, it is clear that if on the equilibrium path $N_t \geq 1$, then $p_{t+\Delta} = 1$ for all $\Delta > 0$. If $N_t = 0$, Bayes' rule implies that

$$p_t = \frac{p_0 e^{-\lambda_1 t}}{p_0 e^{-\lambda_1 t} + 1 - p_0}$$

on the equilibrium path.

Now before the first breakthrough, given an arbitrary incentive scheme $(h_t, w_t)_{0 \leq t \leq T}$, the agent seeks to choose $\{(k_{0,t}, k_{1,t})\}_{0 \leq t \leq T} \in \mathcal{U}^A$ so as to maximize

$$\int_0^T \left\{ r e^{-rt - \lambda_1 \int_0^t p_\tau k_{1,\tau} \, d\tau - \lambda_0 \int_0^t k_{0,\tau} \, d\tau} \left[(1 - k_{0,t} - k_{1,t}) s + k_{0,t} \lambda_0 (h_t + \omega_t) + k_{1,t} \lambda_1 p_t (h_t + w_t) \right] \right\} dt.$$

subject to $\dot{p}_t = -\lambda_1 k_{1,t} p_t (1 - p_t)$.

The following impossibility result is now immediate:

Proposition 3.1 *If $\lambda_0 \geq \lambda_1$, there does not exist a wage scheme $\{\mathcal{W}_t\}_{0 \leq t \leq \overline{T}}$ implementing any strategy in $\mathcal{K}$.*

PROOF: Suppose $\lambda_0 > \lambda_1$. Then, any distribution over $\{N_t\}_{0 \leq t \leq \overline{T}}$ that can be generated by a good arm 1 can be generated by a combination of arm 0 and the safe arm that puts strictly positive weight on the safe arm. As the safe arm gives the agent an instantaneous flow utility of $s > 0$, the latter option strictly dominates the former. If $\lambda_0 = \lambda_1$, arm 0 dominates arm 1 since $\hat{p}_t < 1$ before the first breakthrough. ∎

In the rest of the paper, I shall therefore assume that $\lambda_1 > \lambda_0$. When we denote the solution to the agent's problem that is implemented by an incentive scheme $(h_t, w_t)_{0 \leq t \leq T}$ as $\{(k^*_{0,t}, k^*_{1,t}) ((h_t, w_t)_{0 \leq t \leq T})\}_{0 \leq t \leq T}$, the principal's problem is to choose $(h_t, w_t)_{0 \leq t \leq T}$ so as to minimize his wage bill

$$\int_0^T r e^{-rt - \lambda_1 \int_0^t p_\tau \, d\tau} p_t \lambda_1 (h_t + w_t) \, dt$$

subject to $\{(k^*_{0,t}, k^*_{1,t}) ((h_t, w_t)_{0 \leq t \leq T})\}_{0 \leq t \leq T} \in \mathcal{K}$.

In the next two sections, I shall consider the end date T as given. In section 6, the principal will optimally choose this end date T. Thus far, we have been silent on *how* the continuation value of w_t is delivered to the agent after his first breakthrough. It will turn out, though, that the manner in which the principal gives the agent his continuation value will matter greatly, as we will see in the next section.

4 Incentives After The Breakthrough

4.1 Introduction

The purpose of this section is to analyze how the principal will deliver a promised continuation value of w_t given a first breakthrough has occurred at time t. His goal will be to find a scheme which maximally discriminates between an agent who has achieved his breakthrough on arm 1, as he was supposed to, and an agent who has been "cheating", i.e. who achieved the breakthrough on arm 0. Put differently, for any given promise w_t to the on-equilibrium agent, it is the principal's goal to push the off-equilibrium agent's continuation value ω_t down to as low a level as possible, as this will give the principal the biggest bang for his buck in terms of incentives. As an off-equilibrium agent always has the option of imitating the on-equilibrium agent's strategy, we know that $\omega_t \geq \hat{p}_t w_t$, where $\hat{p}_t \in [p_t, p_0]$ denotes his (off-equilibrium) belief at time t. Writing ω_t as a function of $\hat{p}_t$, the following proposition shows that it is possible to get arbitrarily close to this lower bound.

Proposition 4.1 *For every $\epsilon > 0$, $w_t \geq s(1 - e^{-r(T-t)})$, and $\hat{p}_t \in [p_t, p_0]$, there exists a continuation scheme such that $\omega_t(\hat{p}_t) \leq \hat{p}_t w_t + s(1 - e^{-r\epsilon})$.*

PROOF: Proof is by construction, see *infra*. ∎

The construction of this wage scheme relies on the assumption that $\lambda_1 > \lambda_0$, implying the variance in the number of successes with a good risky arm 1 is higher than with arm 0. Therefore, the principal will structure his wage scheme in such a way as to reward realizations in the number of later breakthroughs that are "extreme enough" that they are very unlikely to have been achieved on arm 0 as opposed to arm 1. Thus, even the most pessimistic of off-equilibrium agents would prefer to bet on his arm 1 being good rather than pull arm 0. Yet, now, in contrast to the off-equilibrium agents, an on-equilibrium agent will know for sure that his arm 1 is good, and therefore has a distinct advantage when facing the principal's payment scheme after a first breakthrough. The agent's anticipation of this advantage in turn gives him the right incentives to use arm 1 rather than arm 0 before the first breakthrough occurs.

4.2 Construction of An Optimal Continuation Scheme

My construction proceeds in several steps. First, the principal will only pay the agent for the m-th breakthrough, where m is chosen large enough that even the most pessimistic of

off-equilibrium agents will deem m breakthroughs more likely to occur on arm 1 than on arm 0. Then, for a given $\epsilon > 0$, I make sure that even the most pessimistic of off-equilibrium agents will not switch to playing safe with more than ϵ time left to go. This requires a certain minimum lump sum reward for the m-th breakthrough. Then, given this reward, the end date $\check{T}(t)$ is chosen appropriately so that the on-equilibrium agent exactly receive his promised continuation value of w_t in expectation.

Specifically, the agent is only paid a constant lump sum of $\overline{V}_0$ after his $(m+1)$-st breakthrough, where m is chosen sufficiently high that even for the most pessimistic of all possible off-equilibrium agents, m breakthroughs are more likely on arm 1 than on arm 0. As $\lambda_1 > \lambda_0$, such an m obviously exists; e.g. any m satisfying $p_{\overline{T}} \left(\frac{\lambda_1}{\lambda_0}\right)^m > e^{(\lambda_1-\lambda_0)\overline{T}}$ will do. Thus, for all types of off-equilibrium agents, arm 0 will be dominated by arm 1.[6]

Now, I recursively define auxiliary functions $V_i(\tilde{t}; \overline{V}_0)$ for $i = 1, ..., m-1$ according to

$$V_i(\tilde{t}; \overline{V}_0) := \max_{\{k_{i,\tau}\} \in \mathcal{M}(\tilde{t})} \int_{\tilde{t}}^{\check{T}(t)} k_{i,\tau} \left[e^{-(r+\lambda_1)(\tau-\tilde{t})} \left(\lambda_1 V_{i-1}(\tau; \overline{V}_0) - s \right) \right] d\tau$$

where $\mathcal{M}(\tilde{t})$ denotes the set of measurable functions $k_i : [\tilde{t}, \check{T}(t)] \to [0,1]$, and I set $V_0(\tau; \overline{V}_0) \equiv \overline{V}_0$.

The following lemma notes that, once the agent knows that $\theta = 1$, his best reply is given by a cutoff strategy, along with some useful properties of these functions V_i:

Lemma 4.2 *The agent's best response is given by a cutoff strategy:*

$$V_i(\tilde{t}; \overline{V}_0) := \max_{t_i^* \in [\tilde{t}, \check{T}(t)]} \int_{\tilde{t}}^{t_i^*} e^{-(r+\lambda_1)(\tau-\tilde{t})} \left(\lambda_1 V_{i-1}(\tau; \overline{V}_0) - s \right) d\tau.$$

The functions $V_i(\tilde{t}; \overline{V}_0)$ are continuous, differentiable, ad strictly decreasing in $\tilde{t}$ and strictly increasing in $\overline{V}_0$ for $\tilde{t} < t_i^*$. Moreover, for $\overline{V}_0 > \frac{s}{\lambda_1}$, it is the case that $\check{T}(t) = t_1^* > t_2^* > \cdots > t_{m-1}^*$.

PROOF: The statements obviously hold for $i = 1$, since $\overline{V}_0 = const > \frac{s}{\lambda_1}$.

[6] The formula for m explicitly only makes sure the agent prefers the strategy "always stick with arm 1, whatever befall" over the strategy "always stick with arm 0". This is sufficient for our purposes, though, because once it is optimal for the agent to play arm 0, he will no longer learn, and therefore it will always remain optimal for him to play arm 0 in the future, given he is facing a reward scheme that is constant over time. Moreover, on account of the linear structure of the agent's optimization problem, it is never strictly optimal for him to distribute his resources over several arms at the same time.

For $i > 1$, I posit the induction hypothesis that

$$V_{i-1}(\tilde{t};\overline{V}_0) := \max_{t^*_{i-1}\in[\tilde{t},\check{T}(t)]} \int_{\tilde{t}}^{t^*_{i-1}} e^{-(r+\lambda_1)(\tau-\tilde{t})}\left(\lambda_1 V_{i-2}(\tau;\overline{V}_0) - s\right)\, d\tau,$$

and that V_{i-1} is continuous, differentiable, and strictly decreasing. It now immediately follows that

$$V_i(\tilde{t};\overline{V}_0) := \max_{t^*_i\in[\tilde{t},\check{T}(t)]} \int_{\tilde{t}}^{t^*_i} e^{-(r+\lambda_1)(\tau-\tilde{t})}\left(\lambda_1 V_{i-1}(\tau;\overline{V}_0) - s\right)\, d\tau,$$

with $V_{i-1}(t^*_i) = \frac{s}{\lambda_1}$.

Now, computing the derivative $\dot{V}_i$, one finds that

$$\dot{V}_i(\tilde{t}) = -e^{-(r+\lambda_1)(t^*_i-\tilde{t})}\left(\lambda_1 V_{i-1}(t^*_i) - s\right) + \int_0^{t^*_i-\tilde{t}} e^{-(r+\lambda_1)\chi}\lambda_1\dot{V}_{i-1}(\chi+\tilde{t})\, d\chi.$$

Since V_{i-1} is strictly decreasing, we have that $V_{i-1}(t^*_i) = \frac{s}{\lambda_1}$, and that the first term is zero, while the second term is strictly negative. This establishes that, for $\tilde{t} < t^*_i$, V_i is strictly decreasing also.

Given that the V_i are strictly decreasing, we have that $V_i(t^*_{i+1}) = \frac{s}{\lambda_1} > 0$; hence $V_{i+1}(t^*_{i+1}) = 0$. As $V_i(t^*_i) = 0$, and V_i is strictly decreasing, it follows that $t^*_{i+1} < t^*_i$.

Clearly, $V_1(\tilde{t};\overline{V}_0)$ is strictly increasing in $\overline{V}_0$ for all $\tilde{t} < \check{T}(t)$. A simple induction argument establishes that V_i is strictly increasing in $\overline{V}_0$ for all $i = 1, ..., m-1$. ∎

Next, I choose the constant $\overline{V}_0$ in such a way that $V_{m-1}(\tilde{t}) \geq \tilde{w}(\tilde{t};\hat{p}_t)$ with $\tilde{w}$ defined as

$$\tilde{w}(\tilde{t};\hat{p}_t) = \begin{cases} \frac{s}{\lambda_1} + \frac{1-\check{p}_\tau}{\check{p}_\tau}\frac{s}{r-\lambda_1}\left[\frac{r}{\lambda_1} - e^{-(r-\lambda_1)(\overline{T}-\tau)}\right] & \text{if} \quad r \neq \lambda_1 \\ \frac{s}{\lambda_1} + \frac{1-\check{p}_\tau}{\check{p}_\tau} s\left[\overline{T} - t - \frac{1}{\lambda_1}\right] & \text{if} \quad r = \lambda_1. \end{cases}$$

As I show in the appendix, the function $\tilde{w}$ denotes the reward for the next breakthrough that has to be offered an agent with belief $\check{p}_{\tilde{t}}$ for him to be exactly indifferent between choosing arm 1 and the safe arm, conditional on his always choosing arm 1 till time $\overline{T}$. Of course, in actuality, the agent will stop using the safe arm at some time in $[\check{T}(t) - \epsilon, \check{T}(t)]$, i.e. before time $\overline{T}$. Yet, as $\tilde{w}$ is an increasing function of $\overline{T}$, it provides an upper bound on the actual indifference boundary.

Now, the off-equilibrium agent definitely will not play arm 0, because m breakthroughs are more likely on arm 1 than on arm 0 and the reward for the m-th breakthrough is constant over time; our construction also makes sure that he will never switch to the safe arm before time $\check{T}(t) - \epsilon$. Hence, the option value of doing so is bounded above by $s(1 - e^{-r\epsilon})$.

As a last step, we now need to make sure that the on-equilibrium agent is indeed delivered an expected continuation value of w_t. In order to do so, I first define another auxiliary function $f(\check{T}(t), \overline{V}_0)$:

$$f(\check{T}(t), \overline{V}_0) := E_{\tau_m, t^*}\left[1_{\tau_m \leq \check{T}(t)} e^{-r(\tau_m - t)}\left(\overline{V}_0 + s(1 - e^{-r(\check{T}(t)-\tau_m)})\right) + s(1 - e^{-r(\check{T}(t)-t^*(\check{T}(t)))})|\Lambda_1\right]$$

where Λ_1 is the distribution over τ_m that is engendered by the stopping times $(t^*_{m-1}, ..., t^*_1)$ implied by the optimal behavior of the on-equilibrium agent, who knows that the state is $\theta = 1$, and t^* is the appertaining time t-expected stopping time of the on-equilibrium agent (which depends on $\check{T}(t)$).

Now, if $w_t \leq f(\overline{T}, \overline{V}_0)$, we can choose $\check{T}(t)$ so that $w_t = f(\check{T}(t), \overline{V}_0)$. Otherwise, we choose the constant $\delta > 0$ so that $w_t = f(\overline{T}, \overline{V}_0 + \delta)$.

Now, with $\check{T}(t)$ chosen as described, it may well be the case that $\epsilon \geq \check{T}(t) - t$. In this case, it might well happen that the off-equilibrium agent prefers to play safe all along on $]t, \check{T}(t)]$, in which case he collects a payoff of $s(1-e^{-r(\check{T}(t)-t)}) < s(1-e^{-r\epsilon}) < s(1-e^{-r\epsilon})+\hat{p}_t w_t$. Or otherwise, the agent might play risky for a while, and switch to safe after a period of length $\xi \leq \check{T}(t) - t \leq \epsilon$, in which case his payoff is bounded above by $s(1 - e^{-r\xi}) + \hat{p}_t w_t < s(1 - e^{-r\epsilon}) + \hat{p}_t w_t$.

Thus, in summary, the mechanism I have constructed delivers a certain given continuation value of w_t to the on-equilibrium agent; it must take care of two distinct concerns in order to harness maximal incentive power at a given cost. On the one hand, it must make sure off-equilibrium agents never continue to play arm 0; this is achieved by only rewarding the $(m + 1)$-st breakthrough. On the other hand, the mechanism must preclude the more pessimistic off-equilibrium agents from collecting an excessive option value from switching between the safe arm and arm 1, so as to make being an off-equilibrium agent none too attractive.

5 Before the Breakthrough–Optimal Incentive Scheme

Whereas in the previous section, I have investigated how a principal would optimally deliver a given *continuation* value w_t, the purpose of this section is to understand to what extent the principal would optimally give incentives via continuation values w_t, as opposed to immediate rewards h_t, which are paid out right at the moment of the first breakthrough. We recall from proposition **??** that, for any given $\hat{p}_t$, the principal can choose a continuation scheme such that $\omega_t(\hat{p}_t) = \hat{p}_t w_t$. Since the principal only cares about incentives *on path*, and, since,

for any continuation scheme, it is always the case that $\omega_t(p_t) \geq p_t w_t$, at all times t, it is clearly optimal for the principal to choose the continuation scheme that guarantees that $\omega_t(p_t) = p_t w_t$. Clearly, we have that $w_t \geq (1 - e^{-r(T-t)})s$, since otherwise the agent would prefer the safe arm over arm 1. In order to analyze this question, we first have to consider the agent's best response to a given incentive scheme $(h_t, w_t)_{0 \leq t \leq T}$, in order to derive necessary conditions for the agent to best reply by always using arm 1 until the first breakthrough. In a second step, we will then use these necessary conditions as constraints in the principal's problem as he seeks to minimize his wage bill.

While the literature on experimentation in bandits would typically use dynamic programming techniques, this would not be expedient here, as an agent's optimal strategy will depend not only on his current belief and the current incentives he is facing but also on the entire path of future incentives. To the extent we do not want to impose any *ex ante* monotonicity constraints on the incentive scheme, today's scheme need not be a perfect predictor for the future path of incentives; therefore, even a three-dimensional state variable (p_t, h_t, w_t) would be inadequate. Thus, I shall be using the Pontryagin approach of Optimal Control.

The Agent's Problem

Given an incentive scheme $(h_t, w_t)_{0 \leq t \leq T}$, the agent chooses $(k_{0,t}, k_{1,t})$ so as to maximize

$$\int_0^T \left\{ r e^{-rt - \lambda_1 \int_0^t p_\tau k_{1,\tau}\, d\tau - \lambda_0 \int_0^t k_{0,\tau}\, d\tau} \left[(1 - k_{0,t} - k_{1,t})s + k_{0,t}\lambda_0(h_t + \omega_t(p_t)) + k_{1,t}\lambda_1 p_t (h_t + w_t) \right] \right\} dt.$$

subject to $\dot{p}_t = -\lambda_1 k_{1,t} p_t (1 - p_t)$.

It will turn out to be useful to work with the log-likelihood ratio $x_t := \ln\left(\frac{1-p_t}{p_t}\right)$, and the probability of no success on arm 0, $y_t := e^{-\lambda_0 \int_0^t k_{0,\tau}\, d\tau}$, as the state variables in our variational problem. These evolve according to $\dot{x}_t = \lambda_1 k_{1,t}$ (to which law of motion I assign the co-state μ_t) and $\dot{y}_t = -\lambda_0 k_{0,t} y_t$ (co-state γ_t), respectively. The initial values $x_0 \equiv \ln\left(\frac{1-p_0}{p_0}\right)$ and $y_0 = 1$ are given, and x_T and y_T are free. The agent's controls are $(k_{0,t}, k_{1,t}) \in \{(a, b) \in \mathbb{R}_+ : a + b \leq 1\}$.

Neglecting a constant factor, the Hamiltonian $\mathfrak{H}_t$ is now given by

$$\begin{aligned} \mathfrak{H}_t = e^{-rt} y_t [(1 - k_{0,t} - k_{1,t})s + k_{0,t}\lambda_0(h_t + \omega_t(x_t))] \\ + y_t e^{-rt - x_t} \left[(1 - k_{0,t} - k_{1,t})s + k_{0,t}\lambda_0(h_t + \omega_t(x_t)) + k_{1,t}\lambda_1(h_t + w_t) \right] \} \\ + \mu_t \lambda_1 k_{1,t} - \gamma_t \lambda_0 k_{0,t} y_t. \end{aligned}$$

By the Maximum Principle, the equations (1), (2), (3), together with the transversality conditions $\gamma_T = \mu_T = 0$, are necessary for the agent's behaving optimally by setting $k_{1,t} = 1$ for all t:

$$\dot{\mu}_t = e^{-rt} y_t \left\{ e^{-x_t} \left[(1 - k_{0,t} - k_{1,t})s + k_{0,t}\lambda_0(h_t + \omega_t(x_t)) + k_{1,t}\lambda_1(h_t + w_t) \right] - k_{0,t}\lambda_0(1 + e^{-x_t})\omega'(x_t) \right\}, \tag{1}$$

and

$$\begin{aligned} \dot{\gamma}_t = -e^{-rt}\{ & [(1 - k_{0,t} - k_{1,t})s + k_{0,t}\lambda_0(h_t + \omega_t(x_t))] \\ & + e^{-x_t} \left[(1 - k_{0,t} - k_{1,t})s + k_{0,t}\lambda_0(h_t + \omega_t(x_t)) + k_{1,t}\lambda_1(h_t + w_t) \right] \} + \gamma_t \lambda_0 k_{0,t}. \end{aligned} \tag{2}$$

$$\begin{aligned} & e^{-rt} y_t \left[e^{-x_t}\lambda_1(h_t + w_t) - (1 + e^{-x_t})s \right] + \mu_t \lambda_1 \\ & \qquad \geq \max \left\{ 0, e^{-rt} y_t (1 + e^{-x_t})[\lambda_0(h_t + \omega_t(x_t)) - s] - \gamma_t \lambda_0 y_t \right\}. \end{aligned} \tag{3}$$

In the appendix, it is shown that these conditions are also sufficient for optimality of the agent's behavior, thus validating my first-order approach.

The Principal's Problem

Now, we turn to the principal's problem, who will take the agent's incentive constraint into account when designing his incentive scheme with a view toward implementing $k_{1,t} = 1$ for almost all $t \in [0, T]$; we note that $k_{1,t} = 1$ for all t implies $y_t = 1$ for all t. Thus, the principal's objective is to choose $(h_t, w_t)_{0 \leq t \leq T} \in [0, L] \times [s(1 - e^{-r(T-t)}), L]$ (for some L which I choose large enough) so as to minimize

$$\int_0^T r e^{-rt - \lambda_1 \int_0^t p_\tau \, d\tau} p_t \lambda_1 (h_t + w_t) \, dt$$

subject to the constraints (1), (2), (3), and the transversality conditions $\mu_T = \gamma_T = 0$.

Neglecting constant factors and using the fact that $x_t = x_0 + \lambda_1 t$, which we derived when we analyzed the agent's problem, one can re-write the principal's objective in terms of the log-likelihood ratio as

$$\int_0^T e^{-(r+\lambda_1)t}(h_t + w_t) \, dt.$$

This can be viewed as a variational problem with the co-state variables (μ_t, γ_t) from the agent's problem as the state variables. I denote the co-state associated with μ_t as ξ_t, and

the one associated with γ_t as η_t. I define ν_t as the Lagrangian parameters associated with the agent's incentive constraints (3).

As we have seen, μ_t and γ_t evolve according to

$$\dot{\mu}_t = e^{-rt-x_t}\lambda_1(h_t + w_t),$$

and

$$\dot{\gamma}_t = -e^{-rt-x_t}\lambda_1(h_t + w_t) = -\dot{\mu}_t.$$

The generalized Hamiltonian is given by

$$\begin{aligned}\mathcal{H}_t &= -re^{-(r+\lambda_1)t}(h_t + w_t) + (\xi_t - \eta_t)e^{-rt-x_t}\lambda_1(h_t + w_t) \\ &+\nu_t\left\{e^{-rt}\left[e^{-x_t}\lambda_1(h_t + w_t) - (1 + e^{-x_t})s\right] + \mu_t\lambda_1 - \max\left\{0, e^{-rt}(1 + e^{-x_t})\left[\lambda_0(h_t + \omega_t(x_t)) - s\right] - \gamma_t\lambda_0\right\}\right\}.\end{aligned}$$

In the appendix, I prove the existence of an optimal plan, a result I state formally in the following lemma.

Lemma 5.1 *There exists an optimal wage scheme.*

PROOF: See the discussion of the principal's optimization problem in the appendix. ∎

By Pontryagin's Principle, any optimal plan must maximize $\mathcal{H}_t$; in the appendix, I show that the optimal plans are actually characterized by these first-order conditions. In the following lemma, I make precise the intuition that if a plan is optimal, the agent's incentive constraint will bind for almost all t.

Lemma 5.2 *In any optimal plan, the agent's incentive constraint binds a.s.*

PROOF: Suppose $(h_t, w_t)_{0\leq t\leq T}$ is an optimal plan. As the plan is optimal, it must be incentive compatible for a.a. t. This means that either $h_t > 0$ or $w_t > s(1 - e^{-r(T-t)})$ for a.a. t, for otherwise playing safe is a strictly dominant action for the agent. Now, suppose that, under $(h_t, w_t)_{0\leq t\leq T}$, the incentive constraint was slack on a set of positive measure. This means that there exists an interval $[t_1, t_2]$, with $t_1 < t_2$, such that the incentive constraint is slack a.e. on $[t_1, t_2]$. Then there exists a collection $(\epsilon_t)_{t_1\leq t\leq t_2}$ with $\epsilon_t > 0$ and an incentive compatible plan $(\tilde{h}_t, \tilde{w}_t)_{0\leq t\leq T}$ satisfying $(\tilde{h}_t, \tilde{w}_t) = (h_t, w_t)$ if $t \in [0, t_1] \cup [t_2, T]$, and $\tilde{h}_t + \tilde{w}_t = h_t + w_t - \epsilon_t$ if $t \in [t_1, t_2]$. As $[t_1, t_2]$ has positive measure, the principal is strictly better off under $(\tilde{h}_t, \tilde{w}_t)_{0\leq t\leq T}$, contradicting the optimality of $(h_t, w_t)_{0\leq t\leq T}$. ∎

In the next lemma, we shall see that it can never be strictly optimal for the principal to pay for the first breakthrough:

Lemma 5.3 *The principal can without loss restrict himself to plans $(h_t, w_t)_{0\leq t\leq T}$ with $h_t = 0$ for all t.*

PROOF: Consider an incentive compatible plan $(\hat{h}_t, \hat{w}_t)_{0\leq t\leq T}$ with $\hat{h}_t > 0$ for some t. Consider the alternative plan $(h_t, w_t)_{0\leq t\leq T}$ with $h_t \equiv 0$ and $w_t = \hat{w}_t + \hat{h}_t$. Applying proposition 4.1 for an ϵ satisfying $1 - e^{-r\epsilon} \leq \frac{1-p_t}{2}\frac{\hat{h}_t}{s}$ (which exists, because $\hat{h}_t > 0$), shows that $(h_t, w_t)_{0\leq t\leq T}$ is incentive compatible. Moreover, it gives the principal exactly the same payoff as the original plan $(\hat{h}_t, \hat{w}_t)_{0\leq t\leq T}$. ∎

Now, we are ready to characterize the optimal incentive scheme, which is essentially unique in the class of optimal schemes with $h_t = 0$ for a.a. t, as the following proposition shows. The characterization relies on the fact, which we have formalized in lemma 5.2, that it never pays for the principal to give strict rather than weak incentives for the agent to do the right thing, because if he did, he could lower his expected wage bill while still providing adequate incentives. This means that the agent is indifferent between doing the right thing and using arm 1, on the one hand, and his next best outside option on the other hand. Yet, the wage scheme we have constructed in section 4 makes sure that the agent's best outside option can never be arm 0. Indeed, playing arm 0 yields the agent approximately $p_t w_t$ after a breakthrough, which occurs with an instantaneous probability of $\lambda_0 dt$ if arm 0 is pulled over a time interval of infinitesimal length dt. Arm 1, by contrast, yields w_t in case of a breakthrough, which occurs with an instantaneous probability of $p_t\lambda_1 dt$; thus, as $\lambda_1 > \lambda_0$, arm 1 dominates arm 0. Any optimal incentive scheme now has the property that the agent is exactly indifferent between the safe arm and arm 1. It is therefore no surprise that our optimal scheme mirrors the function $\tilde{w}$, which we have introduced in section 5. These insights are summarized in the following two propositions:

Proposition 5.4 *An optimal plan is given by $h_t = 0$ and*

$$w_t = \frac{s}{\lambda_1} e^{\lambda_1^2 (T-t)} + \frac{s}{1+\lambda_1}\frac{1-p_t}{p_t}\left[\frac{1}{\lambda_1} + e^{\lambda_1(1+\lambda_1)(T-t)}\right]$$

for all $t \in [0, T]$.

PROOF: By lemmas 5.1 and 5.3, we know that there exists an optimal wage scheme $(h_t, w_t)_{0\leq t\leq T}$ with $h_t = 0$ for all t. In this scheme, lemma 5.2 implies that one of the following two constraints will bind almost surely:

$$e^{-rt}\left[e^{-x_t}\lambda_1 w_t - (1+e^{-x_t})s\right] + \mu_t\lambda_1 \geq 0, \tag{4}$$

$$e^{-rt}\left[e^{-x_t}\lambda_1 w_t - (1+e^{-x_t})s\right] + \mu_t\lambda_1 \geq e^{-rt}(1+e^{-x_t})\left[\lambda_0\omega_t(x_t) - s\right] - \gamma_t\lambda_0. \tag{5}$$

Moreover, by the maximum principle, we know that, for any t, $\mu_t = -\gamma_t$. Now, suppose that the constraint (4) is slack on a set of positive measure. This means that there exist times $t_1 < t_2$ such that (4) is slack a.s. on $[t_1, t_2]$. Lemma 5.2 implies that constraint (5) will bind a.s. on $[t_1, t_2]$. Simple algebra now shows that for (4) to hold given that (5) binds, it has to be the case that $\omega_t \geq p_t w_t + \left(\frac{1}{\lambda_0} - \frac{1}{\lambda_1}\right) s$ a.e. on $[t_1, t_2]$. Yet, by proposition 4.1, there exists an alternative scheme $(\tilde{h}_t, \tilde{w}_t, \tilde{\omega}_t)$ with $\tilde{h}_t \equiv h_t \equiv 0$ and $\tilde{w}_t \equiv w_t$ for all t, yet $\tilde{\omega}_t < p_t w_t + \left(\frac{1}{\lambda_0} - \frac{1}{\lambda_1}\right) s$. Clearly, $(\tilde{h}_t, \tilde{w}_t, \tilde{\omega}_t) \equiv (h_t, w_t, \tilde{\omega}_t)$ still satisfies (4) with slackness a.e. on $[t_1, t_2]$, since (4) is independent of $\tilde{\omega}_t$. Since $\tilde{\omega}_t < \omega_t$ a.e. on $[t_1, t_2]$, it follows that (5) is now also slack a.s. on $[t_1, t_2]$. Hence, there exists a sequence of $(\delta_t)_{t_1 \leq t \leq t_2}$ with $\delta_t > 0$ such that, for $\hat{w}_t := \tilde{w}_t - \delta_t$, $(h_t, \hat{w}_t, \tilde{\omega}_t)$ satisfy both constraints and imply lower wage costs for the principal on $[t_1, t_2]$, a set of positive measure, contradicting the optimality of (h_t, w_t, ω_t).

Thus, we have shown that if $h_t = 0$ for all t, and (h_t, w_t) is optimal, then (4) binds a.s. Furthermore, by the maximum principle, $\gamma_t = \lambda_1 \int_t^T e^{-r\tau - x_\tau} w_\tau$, which completes the proof. ∎

That the agent will be kept indifferent between arm 1 and the safe arm is a feature of *any* optimal wage scheme, as the following proposition shows:

Proposition 5.5 *Any optimal wage scheme $(h_t, w_t)_{0 \leq t \leq T}$ has the property that, prior to the first breakthrough, it keeps the agent indifferent between arm 1 and the safe arm almost surely.*

PROOF: Proof is by contradiction. Suppose to the contrary that $(h_t, w_t)_{0 \leq t \leq T}$ is an optimal wage scheme with the property that the agent strictly prefers arm 1 over the safe arm a.s. on some interval $[t_1, t_2]$ with $0 \leq t_1 < t_2 \leq T$. In a first step, I shall show that this implies that $h_t > 0$ a.s. on $[t_1, t_2]$, which, as I show in a second step, contradicts the optimality of $(h_t, w_t)_{0 \leq t \leq T}$.

Indeed, suppose it is not the case that $h_t > 0$ a.s. on $[t_1, t_2]$. Then, there exists a time interval $[t_1', t_2'] \subseteq [t_1, t_2]$ such that $t_1' < t_2'$ and $h_t = 0$ a.s. on $[t_1', t_2']$. Since the agent a.s. strictly prefers arm 1 over the safe arm on this interval, it follows by lemma 5.2 that the

constraint (5) will bind a.s. on $[t_1', t_2']$, which, as we have seen in the proof of Proposition 5.4, contradicts optimality.

Therefore, $h_t > 0$ a.s. on $[t_1, t_2]$. As the agent strictly prefers arm 1 over the safe arm, lemma 5.2 implies that the incentive constraint

$$e^{-rt}\left[e^{-x_t}\lambda_1(h_t + w_t) - (1 + e^{-x_t})s\right] + \mu_t\lambda_1 \geq e^{-rt}(1 + e^{-x_t})\left[\lambda_0(h_t + \omega_t(x_t)) - s\right] - \gamma_t\lambda_0$$

will bind for a.a. $t \in [t_1, t_2]$. Now, consider the alternative plan $(\hat{h}_t, \hat{w}_t)_{0\leq t\leq T}$ with $\hat{h}_t = 0$ and $\hat{w}_t = w_t + h_t$ for all $t \in [t_1, t_2]$, and $\hat{h}_t = h_t$ and $\hat{w}_t = w_t$ for all $t \in [0, t_1[\cup]t_2, T]$. Arguing as in the proof of lemma 5.3, one shows that $(\hat{h}_t, \hat{w}_t)_{0\leq t\leq T}$ satisfies the incentive constraint (3) with slackness, and gives the principal exactly the same payoff as the original plan $(h_t, w_t)_{0\leq t\leq T}$. Therefore, by lemma 5.2, the principal can strictly improve over $(\hat{h}_t, \hat{w}_t)_{0\leq t\leq T}$, and hence over $(h_t, w_t)_{0\leq t\leq T}$. ∎

Thus, an immediate implication of the preceding proposition is that the optimal incentive scheme is essentially unique in that $w_t + h_t$ is a.s. uniquely pinned down in any optimal incentive scheme:

Corollary 5.6 *If $(h_t, w_t)_{0\leq t\leq T}$ is optimal, then*

$$h_t + w_t = \frac{s}{\lambda_1}e^{\lambda_1^2(T-t)} + \frac{s}{1+\lambda_1}\frac{1-p_t}{p_t}\left[\frac{1}{\lambda_1} + e^{\lambda_1(1+\lambda_1)(T-t)}\right]$$

t-a.s.

6 The Optimal Stopping Time

In this section, the principal can optimally choose the end date T, which we have taken to be given thus far. As the main result of this section, I shall show that agency costs do *not* imply that the principal will stop the project inefficiently early. As the first-best benchmark, I use the solution which is given by the hypothetical situation in which the principal operates the bandit himself. He would obviously never use arm 0. Therefore, as is well known, his problem is equivalent to a one-armed bandit problem where he has to decide at what time to stop using the risky arm, which he pulls at a flow cost of s, conditional on not having obtained a success thus far, i.e. he chooses a stopping time T so as to maximize

$$\int_0^T \left\{re^{-rt-\lambda_1\int_0^t p_\tau\, d\tau}(p_t\lambda_1\Pi - s)\right\} dt. \tag{6}$$

Clearly, the integrand is positive if, and only if, $p_t\lambda_1\Pi \geq s$, i.e. as long as $p_t \geq \frac{s}{\lambda_1\Pi} =: p^m$. As the principal is only interested in the first breakthrough, information has no value for him, meaning that, very much in contrast to the classical bandit literature, he is not willing to forgo current payoffs in order to *learn* something about the state of the world. In other words, he will behave myopically, i.e. as though the future was of no consequence to him, and stops playing risky at his myopic cutoff belief p^m.

Now, as the principal delegates the investigation to an agent, he will choose T so as to maximize

$$\int_0^T \left\{ re^{-rt-\lambda_1 \int_0^t p_\tau \, d\tau} p_t\lambda_1 \left(\Pi - (h_t + w_t)\right) \right\} dt. \tag{7}$$

Thus, all that changes with respect to the first best problem (6) is that the opportunity cost flow s is now replaced by the wage costs $h_t + w_t$, which only have to be paid out in case of a success, which happens with an instantaneous probability of $p_t\lambda_1 dt$. Yet, recall from the preceding sections that given the optimal incentive scheme we have computed there, the principal only needs to compensate the agent for his outside option of using the safe arm; yet, this is exactly what risky arm 1 has to compensate the principal for in the first-best problem. Thus, the following result should come as little surprise:

Proposition 6.1 *The principal stops the delegated project at the time T^* when $p_{T^*} = p^m$.*

PROOF: After we plug the $h_t + w_t$ we have determined in corollary 5.6 into the principal's objective (7), we find that his problem is equivalent to maximizing

$$\int_0^T e^{-(r+\lambda_1)t} \left[\lambda_1\Pi - se^{\lambda_1^2(T-t)} - \frac{s}{1+\lambda_1} e^{x_0+\lambda_1 t} \left(1 + \lambda_1 e^{\lambda_1(1+\lambda_1)(T-t)}\right) \right] dt.$$

If the principal were to stop at time t, his payoff after time t would be zero. If instead he decided to stop "an instant later", i.e. at time $t+dt$, his payoff from doing so would amount to $e^{-(r+\lambda_1)t} \left[\lambda_1\Pi - se^{\lambda_1^2 dt} - \frac{s}{1+\lambda_1} e^{x_0+\lambda_1 t} \left(1 + \lambda_1 e^{\lambda_1(1+\lambda_1)dt}\right) \right] dt$, which exceeds the payoff from stopping at time t if, and only if, $\lambda_1\Pi - \frac{s}{p_t} \geq 0$. As p_t monotonely decreases over t, the left-hand side of this expression is monotonely decreasing in t, so that there is a unique optimal stopping time T^*, which is given by $p_{T^*} = p^m$. ∎

Thus, while delegating the project to an agent forces the principal to devise quite a complicated incentive scheme, it does not force him to stop the exploration inefficiently early; hence, in the second-best solution, there is no efficiency loss stemming from agency costs. In summary, if $\lambda_0 \geq \lambda_1$, it is impossible to have the agent use arm 1; if $\lambda_0 < \lambda_1$, it is possible to give incentives in a manner that even achieves efficiency.

7 Conclusion

The present paper introduces the question of optimal incentive design into a dynamic single-agent model of experimentation on bandits. I have shown that even though the principal only cares about the first breakthrough, he only rewards later ones. Structuring incentives appropriately allows the principal to achieve first-best levels of exploration.

The present paper only investigates the case of a single agent. It would be interesting to explore how the structure of the optimal incentive scheme would change if several agents were simultaneously working for the same principal. Intuition would suggest that the rationale for only rewarding later breakthroughs should carry over to that case. Previous literature on strategic experimentation on bandits with exogenously given rewards has found that in most cases the efficient amount of experimentation cannot be achieved in any Markov perfect equilibrium.[7] It would be quite compelling to investigate under what conditions efficiency could be sustained with several players. I intend to explore these questions in future work.

[7]See Bolton & Harris (1999) or Keller, Rady, Cripps (2005), for positively correlated bandits, for instance. Klein & Rady (2010), by contrast, find that for perfectly negatively correlated bandits the overall amount of experimentation is efficient in any equilibrium, while there exist equilibria for imperfectly negatively correlated bandits with the same efficiency properties; the speed of learning will typically be inefficiently slow, unless the exogenously given stakes are low. Klein (2010), by contrast, shows that when players are able to choose if they want to investigate a given hypothesis or its negation, the efficient solution can be implemented in a Markov perfect equilibrium if, and only if, the stakes exceed a certain threshold.

Appendix

Derivation of the Function $\tilde{w}$

Let t and $\hat{p}_t$ be given. As in section 5, it is again convenient to work with the log-likelihood ratio $x_\tau := \ln\left(\frac{1-\check{p}_\tau}{\check{p}_\tau}\right)$. We now consider the Hamiltonian for the hypothetical problem where the off-equilibrium agent with initial belief $\hat{p}_t$ has to allocate his flow endowment between the safe arm and arm 1 over the time interval $[t, \overline{T}]$, subject to $\dot{x}_\tau = \lambda_1 k_{1,\tau}$ (co-state variable μ_τ), x_t given and x_T free, and his first breakthrough on arm 1 is rewarded according to the function $\tilde{w}(\tau; \hat{p}_t)$, which is to be determined:[8]

$$\tilde{\mathfrak{H}}_\tau = e^{-r(\tau-t)} k_{1,\tau} \left[-(1+e^{-x_\tau})s + e^{-x_\tau}\lambda_1 \tilde{w}(\tau; \hat{p}_t)\right] + \mu_\tau \lambda_1 k_{1,\tau}.$$

Clearly, the agent's choice set is closed and bounded, the set of admissible policies is non-empty, and the state variable is bounded. Moreover, the objective is linear in the choice variable; thus, existence of an optimal plan follows from the Existence Theorem of Filippov-Cesari (Thm. 8 in Seierstad & Sydsæter, 1987, p. 132).

To show sufficiency of the first-order Pontryagin conditions, I use the same variable transformation as Bonatti & Hörner (2010), $q_\tau := e^{-x_\tau}$. The maximized Hamiltonian is then clearly concave in q_τ, so that sufficiency follows from Arrow's Sufficiency Theorem (Thm. 5 in Seierstad & Sydsæter, 1987, p. 107).

Now, Pontryagin's conditions are given by $\mu_{\overline{T}} = 0$,

$$\dot{\mu}_\tau = e^{-r(\tau-t)-x_\tau} k_{1,\tau} \left[\lambda_1 \tilde{w}(\tau; \hat{p}_t) - s\right];$$

and the agent is indifferent between choosing arm 1 and the safe arm at time τ if, and only if,

$$e^{-r(\tau-t)} \left[e^{-x_\tau}\lambda_1 \tilde{w}(\tau; \hat{p}_t) - (1+e^{-x_\tau})s\right] + \mu_\tau \lambda_1 = 0.$$

Now, plugging in $k_{1,\tau} = 1$ for all $\tau \in [t, \overline{T}]$ gives us $\tilde{w}$ as defined in the text. ∎

The Agent's Optimization Problem

As derived in the text, the agent's Hamiltonian is given by

$$\begin{aligned}\mathfrak{H}_t = e^{-rt} y_t [(1-k_{0,t}-k_{1,t})s + k_{0,t}\lambda_0(h_t + \omega_t(x_t))] \\ + y_t e^{-rt-x_t}\left[(1-k_{0,t}-k_{1,t})s + k_{0,t}\lambda_0(h_t+\omega_t(x_t)) + k_{1,t}\lambda_1(h_t+w_t)\right]\} \\ + \mu_t \lambda_1 k_{1,t} - \gamma_t \lambda_0 k_{0,t} y_t.\end{aligned}$$

[8] Here, it turns out to be convenient to normalize payoffs by subtracting a flow of s. This way, all the dynamics stop as soon as the agent switches to safe.

with the state variables evolving according to $\dot{x}_t = \lambda_1 k_{1,t}$ (co-state μ_t) and $\dot{y}_t = -\lambda_0 k_{0,t} y_t$ (co-state γ_t), x_0 given, $y_0 = 1$, and x_T and y_T free, and $(k_{0,t}, k_{1,t}) \in \{(a,b) \in \mathbb{R}_t : a + b \leq 1\} \equiv \mathcal{U}_T^A$. Clearly, $\mathcal{U}_T^A$ is closed and bounded, the set of admissible policies is non-empty, and the state variables are bounded. Moreover, the objective is linear in the choice variables; thus, existence of an optimal plan follows from the Existence Theorem of Filippov-Cesari (Thm. 8 in Seierstad & Sydsæter, 1987, p. 132).

To show sufficiency of Pontryagin's conditions, I invoke Arrow's Sufficiency Theorem (Thm. 5 in Seierstad & Sydsæter, 1987, p. 107). To do so, I define the new state variables $\tilde{y}_t := -\ln y_t$ (co-state $\tilde{\gamma}_t$), and $z_t := -e^{\tilde{y}_t - x_t}$ (co-state ζ_t). Thus, $\dot{\tilde{y}}_t = \lambda_0 k_{0,t}$ and $\dot{z}_t = z_t(\lambda_0 k_{0,t} - \lambda_1 k_{1,t})$. Then,

$$\begin{aligned}\mathfrak{H}_t = -e^{-rt+\tilde{y}_t}&[(1 - k_{0,t} - k_{1,t})s + k_{0,t}\lambda_0(h_t + \omega_t(z_t))] \\ &+ e^{-rt} z_t \left[(1 - k_{0,t} - k_{1,t})s + k_{0,t}\lambda_0(h_t + \omega_t(z_t)) + k_{1,t}\lambda_1(h_t + w_t)\right]\} \\ &+ \tilde{\gamma}_t \lambda_0 k_{0,t} + \zeta_t z_t (k_{0,t}\lambda_0 - k_{1,t}\lambda_1).\end{aligned}$$

As $\mathfrak{H}_t$ is linear in the control variables, the maximized Hamiltonian is given by plugging in either $k_{0,t} = k_{1,t} = 0$, $(k_{0,t}, k_{1,t}) = (0,1)$ or $(k_{0,t}, k_{1,t}) = (1,0)$ in the above expression. We shall now show that $\mathfrak{H}_t$ is concave in $(\tilde{y}_t, z_t)$ for each of these three cases, which in turn implies sufficiency of the first-order Pontryagin conditions by Arrow's Theorem.

If $k_{0,t} = k_{1,t} = 0$,

$$\mathfrak{H}_t = -e^{-rt+\tilde{y}_t} s + e^{-rt} z_t s,$$

and hence clearly concave.

If $(k_{0,t}, k_{1,t}) = (0,1)$,

$$\mathfrak{H}_t = e^{-rt} z_t \lambda_1 (h_t + w_t) - \zeta_t z_t \lambda_1,$$

which is linear, and hence concave.

If $(k_{0,t}, k_{1,t}) = (1,0)$ we note that $z_t = p_t(z_t - e^{\tilde{y}_t})$, and thus

$$\mathfrak{H}_t = e^{-rt}\lambda_0 z_t (h_t + \hat{\delta}_t + w_t) - e^{-rt+\tilde{y}_t}\lambda_0(h_t + \hat{\delta}_t) + \lambda_0(\zeta_t z_t + \tilde{\gamma}_t),$$

hence concave also.[9] ∎

The Principal's Optimization Problem

As derived in the text, the principal's Hamiltonian is given by

$$\begin{aligned}&\mathcal{H}_t = -re^{-(r+\lambda_1)t}(h_t + w_t) + (\xi_t - \eta_t)e^{-rt-x_t}\lambda_1(h_t + w_t) \\ &+\nu_t \left\{e^{-rt}\left[e^{-x_t}\lambda_1(h_t + w_t) - (1 + e^{-x_t})s\right] + \mu_t\lambda_1 - \max\left\{0, e^{-rt}(1 + e^{-x_t})\left[\lambda_0(h_t + \omega_t(x_t)) - s\right] - \gamma_t\lambda_0\right\}\right\}.\end{aligned}$$

[9] Here, I set $\omega_t \equiv p_t w_t + \hat{\delta}_t$. In proposition **??**, we have seen that for any $\epsilon > 0$, there exists a continuation scheme such that $p_t w_t \leq \omega_t \leq p_t w_t + s(1 - e^{-r\epsilon})$. Equivalently, we can think of the principal choosing $\hat{\delta}_t \in [\underline{\epsilon}, L]$ with $\underline{\epsilon} > 0$ sufficiently small.

with the state variables evolving according to $\dot{\mu}_t = e^{-rt-x_t}\lambda_1(h_t + w_t)$ (co-state ξ_t), and $\dot{\gamma}_t = -e^{-rt-x_t}\lambda_1(h_t + w_t)$ (co-state η_t), and $\mu_T = \gamma_T = 0$.

Clearly, the set of admissible controls and states is non-empty for $L < \infty$ large enough; e.g. the plan expounded in proposition 5.4 is admissible. Moreover, μ_t and γ_t are bounded for all admissible $(\mu_t, \gamma_t, h_t, w_t)$ because $(h_t, w_t)_{0\leq t\leq T} \in [0, L] \times [s(1 - e^{-r(T-t)}), L]$ and $\gamma_T = \mu_T = 0$. As the set of feasible controls is bounded, the set of incentive compatible controls is also bounded. Since the objective is linear in the controls, all that remains to be shown is that the constraint set is convex for any given t and (μ_t, γ_t). Existence of an optimal plan then follows by Fillipov-Cesari's Theorem (Thm. 2, Seierstad & Sydsæter, 1987, p. 285).

The incentive compatibility constraint can be written as the following two conditions:

$$e^{-rt}\left[e^{-x_t}\lambda_1(h_t + w_t) - (1 + e^{-x_t})s\right] + \mu_t\lambda_1 \geq 0,$$

and

$$e^{-rt}\left[e^{-x_t}\lambda_1(h_t + w_t) - (1 + e^{-x_t})s\right] + \mu_t\lambda_1 \geq e^{-rt}(1 + e^{-x_t})\left[\lambda_0(h_t + \hat{\delta}_t + \frac{w_t}{1 + e^{x_t}}) - s\right] - \gamma_t\lambda_0$$

which, in $(h_t, w_t, \hat{\delta}_t)$-space, is the intersection of two half-spaces, and therefore convex.

As both the objective as well as the above constraints are linear in both control and state variables, sufficiency of first-order conditions immediately follows from Mangasarian's Theorem (see Seierstad & Sydsæter, 1987, Thm. 5, p. 287). ∎

References

Bergemann, D. and U. Hege (2005): "The Financing of Innovation: Learning and Stopping," *RAND Journal of Economics*, 36, 719–752.

Bergemann, D. and U. Hege (1998): "Dynamic Venture Capital Financing, Learning and Moral Hazard," *Journal of Banking and Finance*, 22, 703–735.

Bergemann, D. and J. Välimäki (2008): "Bandit Problems," in: *The New Palgrave Dictionary of Economics*, 2nd edition ed. by S. Durlauf and L. Blume. Basingstoke and New York, Palgrave Macmillan Ltd.

Bergemann, D. and J. Välimäki (2000): "Experimentation in Markets," *Review of Economic Studies*, 67, 213–234.

Bergemann, D. and J. Välimäki (1996): "Learning And Strategic Pricing," *Econometrica*, 64, 1125–1149.

Bolton, P. and C. Harris (1999): "Strategic Experimentation," *Econometrica*, 67, 349–374.

Bolton, P. and C. Harris (2000): "Strategic Experimentation: the Undiscounted Case," in: *Incentives, Organizations and Public Economics – Papers in Honour of Sir James Mirrlees*, ed. by P.J. Hammond and G.D. Myles. Oxford: Oxford University Press, 53–68.

Bonatti, A. and J. Hörner (2010): "Collaborating," *American Economic Review*, forthcoming.

De Marzo, P. and Y. Sannikov (2008): "Learning in Dynamic Incentive Contracts," mimeo, Stanford University.

Gerardi, D. and L. Maestri (2008): "A Principal-Agent Model of Sequential Testing," Cowles Foundation Discussion Paper No. 1680.

Grossman, S. and O. Hart (1983): "An Analysis of the Principal-Agent Problem," *Econometrica*, 51, 7–45.

Hörner, J. and L. Samuelson (2009): "Incentives for Experimenting Agents," Cowles Foundation Discussion Paper No. 1726.

Keller, G. and S. Rady (2010): "Strategic Experimentation with Poisson Bandits," *Theoretical Economics*, forthcoming.

Keller, G., S. Rady and M. Cripps (2005): "Strategic Experimentation with Exponen-

tial Bandits," *Econometrica*, 73, 39–68.

Klein, N. and S. Rady (2010): "Negatively Correlated Bandits," working paper, University of Munich.

Klein, N. (2010): "Strategic Learning in Teams," working paper, University of Munich.

Manso, G. (2010): "Motivating Innovation," working paper, MIT Sloan School of Management.

Murto, P. and J. Välimäki (2009): "Learning and Information Aggregation in an Exit Game," working paper, Helsinki School of Economics.

Presman, E.L. (1990): "Poisson Version of the Two-Armed Bandit Problem with Discounting," *Theory of Probability and its Applications*, 35, 307–317.

Rahman, D. (2010): "Detecting Profitable Deviations," mimeo, University of Minnesota.

Rahman, D. (2009): "Dynamic Implementation," mimeo, University of Minnesota.

Rosenberg, D., E. Solan and N. Vieille (2007): "Social Learning in One-Armed Bandit Problems," *Econometrica*, 75, 1591–1611.

Rothschild, M. (1974): "A Two-Armed Bandit Theory of Market Pricing," *Journal of Economic Theory*, 9, 185–202.

Seierstad, A. and K. Sydsæter (1987): *Optimal Control Theory With Economic Applications*. Elsevier Science.