The Importance of Being Honest[*]

Nicolas Klein[†]

This version: April 14, 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 true, a proper investigation yields successes with a higher intensity than manipulation would; if it is false, it never yields a success. The principal is only interested in the first success *achieved through proper investigation*, yet cannot distinguish how a given success has been achieved. I show that in the optimal incentive scheme there exists some integer m such that the principal will only reward the $(m+1)$-st breakthrough, and that this reward is increasing in the time of the second breakthrough.

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, Federico Echenique, Lucas Maestri, 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 the agent 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 correct, honesty yields successes with a higher frequency than cheating; 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. My goal is to characterize the optimal incentive scheme which will make sure that the agent is always honest up to the first breakthrough at least.

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. Indeed, when designing the incentive scheme, it will be one of the principal's goals to kindle this experimentation motive, by making information valuable to the agent.

I show that even though the principal is only interested in the *first* breakthrough the agent achieves, he will only 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 minimize this option value, even to push it down to 0, the optimal incentive scheme will make sure that the off-equilibrium agent will either shirk throughout, or that he will exactly imitate the on-equilibrium agent. This is achieved by keeping the most pessimistic of all off-equilibrium agents at least indifferent between being honest and shirking, whenever the on-equilibrium agent pursues further breakthroughs. As conditional on no breakthrough arriving, the off-equilibrium agent continuously becomes even more pessimistic, rewards must be increasing in the time of the second breakthrough. Since after the second breakthrough, all off-equilibrium agents will cease learning also, rewards will be flat in the calendar times of later breakthroughs.

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 optimal mechanism through which the principal gives the agent a certain given continuation value; section 5 characterizes the optimal mechanism before the first breakthrough, and section 6 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 in the agent's private cost function, incentives are more muted, if they are complements, incentives are 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.

The paper that is probably closest 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.

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-

[1] See Holmström & Milgrom (1987).

[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*.

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

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) investigates the case where bandits have three arms, with the two risky ones being perfectly negatively correlated. While the aforementioned 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. The contribution of this paper is to introduce the question of optimal incentive provision into a fully-fledged dynamic bandit model.

3 The Model

There is one principal and one agent and an exogenously fixed end date T. 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 one that either yields breakthroughs according to $Po(\lambda_1)$ (with initial probability p_0) or never yields a breakthrough (arm 1). It is commonly known that $\lambda_1 > \lambda_0 > 0$.

Now, let $(b_{0,t}, b_{1,t}) \in \mathcal{B}_t^2$ denote the agent's information available at time t. Specifically, $\mathcal{B}_t$ is the set of all countable subsets of $[0, t]$; its generic element $(b_{0,t}, b_{1,t}) = ((\tilde{t}_1, \tilde{t}_2, ..., \tilde{t}_\xi), (t_1, t_2, ..., t_\zeta)) \in \mathcal{B}_t^2$ denotes the calendar times of breakthroughs achieved on both arms before time t. If there has been no breakthrough on arm i before time t, I set $b_{i,t} \equiv \emptyset$. A *strategy* for the agent is then simply a mapping $(k_0, k_1) : [0, T] \times \mathcal{B}_t^2 \to \{(a, b) : a + b \leq 1\}$; it describes what fraction of his flow resource[4] the agent devotes to arm 0, or arm 1, respectively, with $1 - k_0(t, b_{0,t}, b_{1,t}) - k_1(t, b_{0,t}, b_{1,t})$ being invested in the safe option.

The first breakthrough *achieved on arm 1* at time t yields the principal a payoff of $e^{-rt}\Pi$, which is only realized at date T. The principal has full commitment power but cannot condition payments on whether a breakthrough is achieved on arm 0 or 1. The idea is that he cannot observe the agent's actions, or that the latter are non-contractible, i.e. they could not be verified in a court of law. Indeed, whether a scientist's claim will turn

[4] Think of an agent who distributes his time amongst various tasks, for instance.

out to have been correct or to have been the result of data manipulation can often only be ascertained *ex post*, if at all, and will certainly hardly ever be contractible. At the outset, the principal commits to a payment schedule conditioning on the history he observes, i.e. on the history of breakthroughs. The principal's objective is to implement experimentation on arm 1 at least up until the first breakthrough, and to do so at minimal wage costs.

Specifically, let $z_t \in \mathcal{B}_t$ denote the public information available at time t, which is coarser than the agent's private information, with the generic element $z_t \equiv (t_1, t_2, ..., t_\sigma) \in \mathcal{B}_t$ denoting the calendar times of all breakthroughs prior to time t; if no breakthrough has occurred before time t, I set $z_t \equiv \emptyset$. An *incentive scheme* is given by a function $\tilde{w} : \{(t, z_t) : t \in [0, T], z_t \in \mathcal{B}_t\} \to \mathbb{R}$, with $\int_0^t \tilde{w}(\tilde{t}, z_{\tilde{t}})\, d\tilde{t}$ denoting the monetary transfers from the principal to the agent up to time t given the public information available at the time of the transfers. The agent is protected by limited liability, i.e. $\tilde{w}(t, z_t) \geq 0$ at all t. 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: $h_t := \frac{\tilde{w}(t,(t))}{r}$ denotes the immediate lump sum reward for a breakthrough at time t; $w_t := E_t\left[\int_t^T e^{-r(\tilde{t}-t)} \tilde{w}(\tilde{t}, z_{\tilde{t}})\, d\tilde{t}\right]$ (with $z_{\tilde{t}} = (t, ..., t_\sigma)$) is the expected continuation value of the agent after his first breakthrough, with the expectation being taken with respect to his information at time t (as well as his expected future actions). As the agent always has the option of continuously pulling the safe arm, it is clear that, in equilibrium, $w_t \geq s(1 - e^{-r(T-t)})$. In this notation, the principal's objective is to minimize

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

subject to appropriate incentive constraints making sure the agent always uses arm 1.

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 that arm 1 is good. Denoting the time t belief by p_t, Bayes's rule implies that in the absence of a breakthrough on arm 1

$$p_t = \frac{p_0 e^{-\lambda_1 \int_0^t k_{1,q}\, dq}}{p_0 e^{-\lambda_1 \int_0^t k_{1,q}\, dq} + 1 - p_0};$$

if arm 1 has yielded a breakthrough at some time t, we have that $p_q = 1$ for all $q \in]t, T]$. As the principal anticipates the agent's actions in equilibrium, he will know what p_t is; however, as deviations are unobservable, they lead to the agent's holding some private belief $\hat{p}_t$, which will be different from the public belief p_t off the equilibrium path of play.

4 After the Breakthrough–The Optimal Continuation Scheme

The purpose of this section is to derive the optimal scheme by which 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 simply using his safe arm forever, we have that $\omega_t \geq s(1 - e^{-r(T-t)})$. Since he also 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 $\hat{p}_t$, we have that in the optimal wage scheme $\omega_t = \max\{s(1 - e^{-r(T-t)}), \hat{p}_t w_t\}$, a result summarized in the following proposition:

Proposition 4.1 *In the optimal mechanism delivering an expected continuation payoff of w_t to the agent if he has achieved his first breakthrough on arm 1 at time t, the off-equilibrium expected continuation value of an agent who has achieved his first breakthrough on arm 0 at time t is given by $\omega_t = \max\{s(1-e^{-r(T-t)}), \hat{p}_t w_t\}$, with $\hat{p}_t$ denoting his time t (off-equilibrium) belief.*

The principal only rewards the $(m+1)$-st success by paying a lump sum $\hat{w}(\tau)$ with m an integer satisfying $p_T \left(\frac{\lambda_1}{\lambda_0}\right)^m > e^{(\lambda_1 - \lambda_0)T}$, and $\hat{w}$ being an increasing function of τ, the time of the agent's second breakthrough.

Proof: Proof is by construction, see *infra.* ▮

In order to force off-equilibrium agents down to their lower bound, the principal will endeavor to ensure that any off-equilibrium agent will always either exactly imitate the on-equilibrium agent or play safe forever. Clearly, arm 0 is dominated for an on-equilibrium agent, who knows that arm 1 is good. In order to make arm 0 dominated for all types of off-equilibrium agents,[5] the principal will only pay for the m-th breakthrough after time t, 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 $\lambda_1 > \lambda_0$,

[5] The *type* of an off-equilibrium agent is defined by his belief $\hat{p}_t$ at the moment of his first breakthrough (which occurs on arm 0).

such an m obviously exists. Now, in order to make sure that whenever any off-equilibrium agent uses arm 1 at all, he does so if, and only if, the on-equilibrium agent also uses arm 1, the principal will reward the m-th breakthrough in a way that depends on the time of the first breakthrough after t. Indeed, before achieving their first breakthroughs on arm 1, off-equilibrium agents continue to learn about the quality of arm 1 whenever they use it; therefore, their incentives for playing arm 1 versus using the safe arm vary with the new information they are accumulating. The trick is now either to reward the on-equilibrium agent in such a manner that even the most *optimistic* off-equilibrium agent would rather play safe throughout, or, alternatively, to reward him so much that even the most *pessimistic* off-equilibrium agent is at least indifferent between using arm 1 and the safe arm at all times. This latter option implies that the reward for the m-th breakthrough will be increasing in the time of the second breakthrough.

One additional problem to be taken care of is that for intermediate values of w_t, some off-equilibrium agents might be willing to experiment for a while, and then switch to safe, once they will have become too pessimistic to carry on at a time an on-equilibrium agent would still be experimenting. Whenever agents have a strict incentive to employ this course of action, the option value of doing so would give them a payoff above the lower bound we have identified. To prevent this, the principal will not give out any rewards if the second breakthrough occurs before some time $\hat{\tau}$, thus implementing the safe action on $[t, \hat{\tau}]$ for both the on-equilibrium agent as well as all types of off-equilibrium agents. After $\hat{\tau}$, though, rewards will be so high that even the most pessimistic off-equilibrium agents will play risky, with $\hat{\tau}$ chosen appropriately to give the on-equilibrium agent an expected continuation value of w_t. In the following, I shall make this construction precise.

For the rest of this section, let t and w_t be fixed, and let m be an integer, e.g. the smallest, satisfying $p_T \left(\frac{\lambda_1}{\lambda_0}\right)^m > e^{(\lambda_1 - \lambda_0)T}$. As $\lambda_1 > \lambda_0$, such an m exists. The principal will now pay a lump sum of $\hat{w}(\tau)$ $(\tau > t)$ at the time of the m-th breakthrough after time t, the first of which occurs at time τ, and nothing otherwise. Clearly, this scheme makes arm 0 dominated for any off-equilibrium agent with any plausible belief $\hat{p}_t \in [p_t, p_0]$.[6]

We now consider the family of functions $\overline{w}(.; t, \hat{p}_t) : [t, T] \longrightarrow \mathbb{R}$, where $\overline{w}(\tau; t, \hat{p}_t)$ is the lump sum that would have to be paid after the m-th breakthrough, with the first of these m breakthroughs occurring at time τ, in order to keep the off-equilibrium agent indifferent

[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. 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.

between using arm 1 or using the safe arm at time τ given his belief that arm 1 is good is given by $\frac{\hat{p}_t e^{-\lambda_1(\tau-t)}}{\hat{p}_t e^{-\lambda_1(\tau-t)}+1-\hat{p}_t}$, i.e. under the assumption that he has all the time experimented at full throttle on arm 1. To state the next lemma, I define $\overline{w}_1 := \overline{w}(.;t,1)$.

Lemma 4.2 *$\overline{w}(.;t,\hat{p}_t)$ is strictly increasing for any $\hat{p}_t < 1$; it is strictly decreasing in $\hat{p}_t < 1$ for any given τ. Moreover, it is continuous in both arguments. $\overline{w}_1$ is constant.*

Proof: Using the indifference conditions, it is easy to derive a recursive representation for $\overline{w}$ for any given m. Indeed, fix t and $\hat{p}_t$, and let $V^i_\tau(\tilde{t})$ be the player's value at time $\tilde{t}$ if he is i breakthroughs removed from collecting the lump sum reward and his first breakthrough after t has previously occurred at time τ. (If $i = m$, I set $\tau \equiv \tilde{t}$ and suppress the argument $\tilde{t}$.) Define $\hat{p}_\tau := \frac{\hat{p}_t e^{-\lambda_1(\tau-t)}}{\hat{p}_t e^{-\lambda_1(\tau-t)}+1-\hat{p}_t}$.

Then, we have that $V^m_\tau = (\hat{p}_\tau \lambda_1 V^{m-1}_\tau - s)dt + (1-rdt)(1-\hat{p}_\tau\lambda_1 dt)(V^m_\tau + \dot{V}^m_\tau dt)$, i.e. $(r+\hat{p}_\tau\lambda_1)V^m_\tau = \hat{p}_\tau\lambda_1 V^{m-1}_\tau - s + \dot{V}^m_\tau$. Now, by indifference, $V^m_\tau = \dot{V}^m_\tau = 0$. Thus, $V^{m-1}_\tau = \frac{s}{\hat{p}_\tau\lambda_1}$. As given that the first breakthrough has occurred at time τ, the lump sum $\overline{w}_\tau$ is constant, it follows that $\dot{V}^{m-n}_\tau = 0$ for all $n = 1, ..., m-1$. Noting that after the first breakthrough, the agent will know that the arm is good, we have, by the same logic as before, that $(r+\lambda_1)V^{m-n}_\tau = \lambda_1 V^{m-n-1}_\tau - s$, yielding $V^{m-n-1}_\tau = \frac{s}{\lambda_1} + \frac{r+\lambda_1}{\lambda_1}V^{m-n}_\tau$, with $V^0_\tau \equiv \overline{w}(\tau;t,\hat{p}_t)$. Thus, $\overline{w}(\tau;t,\hat{p}_t)$ is increasing in τ and decreasing in $\hat{p}_t$.

For $\hat{p}_t = 1$, the same steps show that $\overline{w}_1$ is constant. ∎

Now, given m and t, it will be useful to define a mapping $f : \mathcal{L}[t,T] \to \mathbb{R}$ by the following equation:[7]

$$f(h) := r\frac{\lambda_1^m}{(m-1)!}$$
$$\times \int_t^T \int_\tau^T (1-\lambda_1(\tau-t))e^{-(\lambda_1+r)(\check{\tau}-t)}(\check{\tau}-\tau)^{m-2}(\chi-1-\lambda_1(\check{\tau}-\tau))\left(h(\tau)+\frac{s}{r}(1-e^{-r(T-\check{\tau})})\right) d\check{\tau}\, d\tau.$$

Thus, $f(\overline{w}(.;t,\hat{p}_t))$ is the time-t-expected payoff of the on-equilibrium agent given the incentive scheme $\overline{w}(.;t,\hat{p}_t)$ conditional on his always using arm 1 after his first breakthrough. Clearly, $f(\overline{w}(.;t,\hat{p}_t))$ is decreasing in $\hat{p}_t$.

If $w_t \leq f(\overline{w}_1)$, I set

$$\hat{w}(\tau) = \begin{cases} 0 & \text{if } \tau < \hat{\tau} \\ \overline{w}_1 & \text{if } \tau \geq \hat{\tau}, \end{cases}$$

[7] $\mathcal{L}[t,T]$ denotes the space of Lebesgue-integrable real-valued functions defined on the interval $[t,T]$.

with $\hat{\tau}$ chosen such that $w_t = f(\hat{w}) + s(1 - e^{-r(\hat{\tau}-t)})$. In this case, the on-equilibrium agent will play safe on $[t, \hat{\tau}[$, and arm 1 forever thereafter; all off-equilibrium agents will always play safe and thus collect $s(1 - e^{-r(T-t)})$.

Now, suppose $w_t > f(\overline{w}_1)$.

If $w_t > f(\overline{w}(.;t,p_t))$, I set $\hat{w}(\tau) = \overline{w}(\tau;t,p_t) + \delta$ for all $\tau \in [t,T]$, where $\delta > 0$ is the (unique) constant chosen so that $w_t = f(\hat{w})$. In this case, all off-equilibrium agents will strictly prefer to use arm 1 all the way, and their respective payoffs are $\hat{p}_t w_t$.

If $w_t \leq f(\overline{w}(.;t,p_t))$, we set

$$\hat{w}(\tau) = \begin{cases} 0 & \text{if } \tau < \hat{\tau} \\ \overline{w}(\tau;t,p_t) & \text{if } \tau \geq \hat{\tau}, \end{cases}$$

with $\hat{\tau}$ again chosen such that $w_t = f(\hat{w}) + s(1 - e^{-r(\hat{\tau}-t)})$. In this case, the on-equilibrium agent will play safe on $[t, \hat{\tau}[$, and arm 1 forever thereafter; all off-equilibrium agents will behave in the exact same fashion, and thus collect $\hat{p}_t w_t$.

Clearly, no player will switch back to playing safe at any time after their first breakthrough, because whenever players play risky, it is the case that $\hat{w}(\tau) \geq \overline{w}_1$, and $\overline{w}_1$ is what is needed to make a player with a belief of 1 willing to play arm 1 at any time when there are still m breakthroughs to go; thus, they are definitely willing to play risky when there are fewer breakthroughs to go (and a weakly higher lump sum to be gotten).

Thus, in summary, the optimal mechanism delivering a certain given continuation value of w_t to the on-equilibrium agent 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 switching between the safe arm and arm 1 according to a timing schedule that is different from that of the on-equilibrium agent. Indeed, if they did, they could collect an additional option value as opposed to a situation where they were forced to behave as an on-equilibrium agent would. This latter concern can be remedied by having rewards that are increasing in the time of the second breakthrough in precisely such a way as to neutralize the effect of off-equilibrium agents' learning.

5 Before the Breakthrough–The 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 incentivize agents 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 the previous section that $\frac{\partial \omega_t}{\partial w_t} < 1$. By feasibility, we have that $w_t \geq (1-e^{-r(T-t)})s$. In order to analyze this question, we first have to consider the agent's best response to a given incentive scheme $(h_t, w_t)_{0\leq t \leq T}$.

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

Agent's controls are $k_{0,t}$, $k_{1,t}$; the state variable is p_t, which evolves according to $\dot{p}_t = -\lambda_1 k_{1,t} p_t(1-p_t)$ (co-state μ_t).

The agent maximizes

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

The appertaining Hamiltonian is given by

$$\begin{aligned}\mathfrak{H}_t = re^{-rt-\lambda_1\int_0^t p_\tau k_{1,\tau}\,d\tau-\lambda_0\int_0^t k_{0,\tau}\,d\tau}\left[(1-k_{0,t}-k_{1,t})s+k_{0,t}\lambda_0\left(h_t+\omega_t(p_t)\right)+k_{1,t}\lambda_1 p_t(h_t+w_t)\right]\\ -\mu_t\lambda_1 p_t(1-p_t)k_{1,t}\end{aligned}$$

s.t. $k_{0,t}+k_{1,t}\leq 1$. From this we get that

$$\dot{\mu}_t = \lambda_1 k_{1,t}\mu_t(1-2p_t) - re^{-rt-\lambda_1\int_0^t p_\tau k_{1,\tau}\,d\tau-\lambda_0\int_0^t k_{0,\tau}\,d\tau}\left[k_{1,t}\lambda_1(h_t+w_t)+k_{0,t}\lambda_0\omega_t'(p_t)\right]$$

and that $k_{1,t}=1$ is a best response if, and only if,

$$\begin{aligned} re^{-rt-\lambda_1\int_0^t p_\tau k_{1,\tau}\,d\tau-\lambda_0\int_0^t k_{0,\tau}\,d\tau}[\lambda_1 p_t(h_t+w_t)-s]-\mu_t\lambda_1 p_t(1-p_t)\\ \geq \max\left\{0, re^{-rt-\lambda_1\int_0^t p_\tau k_{1,\tau}\,d\tau-\lambda_0\int_0^t k_{0,\tau}\,d\tau}[-s+\lambda_0(h_t+\omega_t)]\right\}.\end{aligned}$$

From the transversality condition, we get $\mu_T \geq 0$. $\{\mu_t\}_{0\leq t\leq T}$ captures the evolution of the opportunity cost of forgone future continuation values in case of a breakthrough today. Clearly, this opportunity cost cannot be negative at the very end of the interaction.

Moreover, the principal will clearly never give strict incentives to use arm 1 in equilibrium, because if he was he could always lower his wage costs and still be providing adequate incentives. Thus, the incentive constraint will bind in equilibrium, and the agent will be just indifferent between doing the right thing and his next best option.

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 to with a view toward implementing $k_{1,t} = 1$ for all $t \in [0, T]$. The principal's controls are w_t and h_t; the state variable is μ_t , i.e. the co-state variable from the agent's problem, which evolves according to

$$\dot{\mu}_t = -\lambda_1 \left[re^{-rt-\lambda_1 \int_0^t p_\tau \, d\tau}(h_t + w_t) - \mu_t(1 - 2p_t) \right].$$

To μ_t I assign the co-state variable ν_t. Moreover, for each t, there is a Lagrangian constraint (co-state ζ_t)

$$\begin{aligned} re^{-rt-\lambda_1 \int_0^t p_\tau \, d\tau}[\lambda_1 p_t(h_t + w_t) - s] - \mu_t\lambda_1 p_t(1 - p_t)& \\ \geq \max\left\{0, re^{-rt-\lambda_1 \int_0^t p_\tau \, d\tau}[-s + \lambda_0(h_t + \omega_t)]\right\}&, \end{aligned}$$

which, as we have argued, will bind in equilibrium, as well as the constraints $h_t \geq 0$ and $w_t \geq s(1 - e^{-r(T-t)})$. The Hamiltonian is

$$\begin{aligned} \mathfrak{H}_t = &-re^{-rt-\lambda_1 \int_0^t p_\tau \, d\tau} p_t\lambda_1(h_t + w_t) - \nu_t\lambda_1[re^{-rt-\lambda_1 \int_0^t p_\tau \, d\tau}(h_t + w_t) - \mu_t(1 - 2p_t)] \\ &- \zeta_t[-re^{-rt-\lambda_1 \int_0^t p_\tau \, d\tau} p_t\lambda_1(h_t + w_t) + re^{-rt-\lambda_1 \int_0^t p_\tau \, d\tau} s + \mu_t\lambda_1 p_t(1 - p_t) \\ &+ \max\{0, re^{-rt-\lambda_1 \int_0^t p_\tau \, d\tau}[-s + \lambda_0(h_t + \omega_t)]\}]. \end{aligned}$$

By mere inspection, we see that it is less costly to give incentives through w_t than through h_t. Therefore, the principal will set $h_t = 0$, and w_t such that the incentive constraint will just bind. This proves the following result:

Proposition 5.1 *Even though the principal only cares about the first breakthrough, he optimally gives incentives exclusively by committing to pay for later breakthroughs (and does not reward the first breakthrough at all), i.e.* $h_t = 0$ *and* $w_t = \frac{\mu_t(1-p_t)}{re^{-rt-\lambda_1 \int_0^t p_\tau \, d\tau}} + \frac{\max\{s, \lambda_0\omega_t(p_t)\}}{\lambda_1 p_t}$.

The intuition for the result is that by paying for later breakthroughs, the principal is in a position to discriminate between honest agents, who achieved their breakthrough on arm 1, and those who cheated and achieved their breakthrough on arm 0. Agents' anticipation of eventually being found out in turn provides them with adequate incentives to ensure that they will use arm 1 throughout. At each point in time, the agent's promised continuation value has to compensate him for his next best outside option, as well as for the forgone option value of having a later first breakthrough in case of a breakthrough today.

6 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. In doing so, he makes sure that off-equilibrium agents either always play safe or exactly imitate the on-equilibrium agent's actions. This can be achieved by making the rewards increasing in the time of second breakthrough.

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 working for the same principal. Furthermore, while my investigation assumes an exogenous end date T, endogenizing the optimal stopping date would make for an interesting extension. I intend to explore these questions in future work.

References

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.

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 Exponential 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.

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.