Free-Riding And Delegation In Research Teams – A Three-Armed Bandit Model*

Nicolas Klein†

This version: August 24, 2009

Abstract

This paper analyzes a two-player game of strategic experimentation with three-armed exponential bandits in continuous time. Players face replica bandits, with one arm that is safe in that it generates a known payoff, whereas the likelihood of the risky arms' yielding a positive payoff is initially unknown. It is common knowledge that the types of the two risky arms are perfectly negatively correlated. I show that the efficient policy is incentive-compatible if, and only if, the stakes are high enough. Moreover, learning will be complete in *any* Markov perfect equilibrium if, and only if, the stakes exceed a certain threshold.

Keywords: Strategic Experimentation, Three-Armed Bandit, Exponential Distribution, Poisson Process, Bayesian Learning, Markov Perfect Equilibrium, R & D Teams.

JEL Classification Numbers: C73, D83, O32.

*I am grateful to Sven Rady for advice and encouragement, and to Ludwig Ensthaler, Johannes Hörner, Daniel Krähmer, Jo Thori Lind, Frank Riedel, Klaus Schmidt, Lones Smith, Eilon Solan, as well as seminar participants at Bielefeld, Penn State, Southern Methodist University, the 2009 SFB/TR 15 Young Researchers' Workshop at Humboldt University Berlin, the 2009 SFB/TR 15 Meeting in Caputh, the 2009 North American Summer Meeting of the Econometric Society, the 2009 Meeting of the Society for Economic Dynamics, the 2009 Summer School on "Limited Cognition, Strategic Thinking and Learning in Games" in Bonn, and the 2009 International Conference on Game Theory at Stony Brook, for helpful comments and suggestions. Financial support from the Deutsche Forschungsgemeinschaft through SFB/TR 15 is gratefully acknowledged.

†Seminar für Dynamische Modellierung, University of Munich, Kaulbachstr. 45, D-80539 Munich, Germany; email: kleinnic@yahoo.com.

1 Introduction

Instances abound where economic agents have to decide whether to use their current information optimally, or whether to forgo current payoffs in order to gather information which might potentially be parlayed into higher payoffs come tomorrow. Often, though, economic agents do not make these decisions in isolation; rather, the production of information is a public good. Think, for instance, of firms exploring neighboring oil fields, or a research team investigating a certain hypothesis, where it is not possible to assign credit to the individual researcher actually responsible for the decisive breakthrough. The canonical framework to analyze these questions is provided by the literature on strategic experimentation with bandits.[1] All of this previous literature has been assuming that players are exogenously *assigned* a certain type of project before the actual experimentation game starts. In this paper, I propose to let the players endogenously *choose* their project over time.

A comparison of my results to those of the previous literature will thus help us ascertain the effects of project choice delegation. Indeed, in many real-world situations, a principal has to decide whether best to delegate a decision or not. Thus, for instance, subsequent to marked growth in the number of its research laboratories and facing increasing competitive pressures, 3M moved to restrict scientists' discretion over their work, which had traditionally been very vast (cf. Bartlett & Mohammed, 1995). Conversely, Swiss pharmaceutical giant Novartis entered into a multi-million five-year agreement with the Department of Microbial and Plant Biology at Berkeley, CA, delegating project decisions to a committee being comprised of five experts, only two of whom were Novartis employees (cf. Lacetera, 2008) – a scheme that can reasonably be interpreted as a commitment device on the part of Novartis to delegate project choice to their scientific partners in academia. A somewhat similar deal had earlier been signed by Thousand Oaks, CA, based pharmaceutical company Amgen and MIT; Lawler (2003) quotes MIT biologist Nancy Hopkins: "There was no attempt by either side to change the direction of our basic research" in the aftermath of the agreement. These kinds of deals beg the question as to why a company would be willing to spend vast amounts of money sponsoring scientific research without asserting control over its direction.[2]

Manso (2007) analyzes the case of a *single* worker, who can either shirk, or take risks

[1] cf. Bolton & Harris (1999, 2000), Keller, Rady, Cripps (2005), Keller & Rady (2007), Klein & Rady (2008).

[2] The optimal allocation of research projects between academia and the commercial sector is the subject of papers by Aghion, Dewatripont, Stein (2005), as well as by Lacetera (2008), who interpret academia as a commitment device for principals not to interfere with scientists' discretion. The frictions at the heart of both of these papers rely on the assumption that scientists' preferences diverge from those of economically oriented, profit-maximizing, firms.

and innovate, or produce in an established, safe, manner. In a simple two-period model, he shows that, in order to induce risk taking, the principal will optimally be very tolerant of, or even reward, early failure and long-term success. The impact of strategic interaction among several agents on the principal's optimal incentive scheme thus far remains an open question in his setting, however. While this contract-theoretic literature focuses on possible incentive schemes that *a principal* can design to induce agents to work at first-best levels, my focus here is different in that I am investigating what organizational form *the scientists themselves* would choose *ex ante*: Would they want to commit to a particular project before starting out on their research, or, assuming they internalized all the economic consequences of their decision, would they choose to work for 3M or for the Novartis-sponsored Berkeley lab, all else being equal?

Generally, my model is applicable to teams performing scientific research, which is often characterized by the investigation of two mutually incompatible hypotheses, e.g. by a null and an alternative hypothesis.[3] In particular, two members of a research team will continually choose themselves on which, if either, of two research projects to work, rather than be assigned a given project. Furthermore, a key assumption is that any progress, or the lack thereof, as well as a team colleague's efforts, are perfectly observable to either team member. Indeed, even though the effort expounded in pursuit of a certain scientific hypothesis will hardly ever be contractible, it is often easily gauged by fellow scientists working in the same field.

Specifically, I consider two players operating replica three-armed exponential bandits in continuous time.[4] One arm is safe in that it yields a known flow payoff, whereas the other two arms are risky, i.e. they can be either good or bad. As the risky arms are meant to symbolize two mutually incompatible hypotheses, I assume that it is common knowledge that exactly one of the risky arms is good. The bad risky arm never yields a positive payoff, whereas a good risky arm yields positive payoffs after exponentially distributed times. As the expected payoff of a good risky arm exceeds that of the safe arm, players will want to know which risky arm is good. As either player's actions, as well as the outcomes of his experimentation, are perfectly publicly observable, there is an incentive for players to free-ride on the information the other player is providing; information is a public good. Moreover, observability, together with a common prior, implies that the players' beliefs agree at all times. As only a good risky arm can ever yield a positive payoff, all the uncertainty is resolved as soon as either

[3]Klein & Rady (2008) analyze a setting where players are *assigned* mutually incompatible hypotheses before the start of the actual experimentation game.

[4]The single-agent two-armed exponential model has first been analyzed by Presman (1990); Keller, Rady, Cripps (2005) have introduced strategic interaction into the model; Klein & Rady (2008) have then introduced negative correlation into the strategic model.

player has a breakthrough on a risky arm of his and beliefs become degenerate at the true state of the world. In the absence of such a breakthrough, players incrementally become more pessimistic about that risky arm that is more heavily utilized. As all the payoff-relevant strategic interaction is captured by the players' common belief process, I restrict players to using stationary Markov strategies with their common posterior belief as the state variable, thus making my results directly comparable to those in the previous strategic experimentation literature.

To the best of my knowledge, this paper is the first to analyze strategic experimentation involving bandits with more than two arms.[5] This extension of the existing literature allows me specifically to analyze the effects of project choice delegation in a setting where externalities are purely informational in nature. Chatterjee & Evans (2004), by contrast, analyze an R & D race also involving payoff externalities in discrete time, where it is common knowledge that exactly one of several projects is good. As in my model, they allow players to switch projects at any point in time. Their model is therefore better-suited e.g. to the analysis of experimentation by rival firms competing for market share; mine may be more appropriate if e.g. one wants to analyze free-riding incentives by scientists working for the same firm or in the same lab, or different jurisdictions investigating the impact of various treatment options for a particular disease, and the like.

Quite surprisingly, I find that free-riding incentives can be overcome in equilibrium if, and only if, the stakes at play, as measured by the ratio of the payoff of a good risky arm over that of a safe arm, *exceed* a certain threshold. While inefficiency on account of free-riding has been a staple result of the strategic experimentation literature, Klein & Rady's (2008) negatively correlated two-armed model found that free-riding incentives could be overcome only for *low* stakes. The reason for this is that with the stakes low enough, it is clear from the start that the initially more pessimistic player will never experiment, so that the more optimistic player will know that he will have no opportunity to free-ride, and, therefore, he will behave efficiently. In the present model, however, players are operating replica bandits, so that this effect cannot obtain here. Yet, with the stakes high enough, the safe option becomes so unattractive that it can be completely disregarded. But then, since there are no payoff rivalries or switching costs in our model, a player is willing to go for the project that looks momentarily more promising given his opponent behaves in the same fashion, which is exactly what efficiency requires. This result would suggest that with two mutually exclusive hypotheses, and high enough stakes, the problem of free-riding in teams could be overcome by delegating the choice of project to the team members. Thus, one reason why a firm such

[5]Manso (2007) embeds a three-armed bandit with two safe arms and one risky arm, operated by a *single agent*, into a simple two-period model. His focus is on the wage schemes a principal would optimally offer the agent to induce him either to choose the risky option or the principal's preferred safe option.

as Novartis or Amgen would possibly want to relinquish control to its scientists could be its desire to overcome their propensity toward free-riding. Indeed, in their case study of 3M, Bartlett & Mohammed (1995) pointedly quote one division vice-president noting the increase in what he termed "motivation and morale issues" in the aftermath of management's re-assertion of control over their research labs.

There are two distinct effects that make the three-armed setup perform better. The first effect is also apparent in the comparison of the planner's solutions, and is based on a strictly positive option value to both players' having access to the initially less auspicious project. The second effect is less obvious, and purely strategic: Indeed, while even the *lower* appertaining planner's solution is not compatible with equilibrium in the two-armed model,[6] the *higher* planner's solution can be achieved in equilibrium with three arms, if the stakes are high enough. This strategic effect is reminiscent of the "motivation and morale issues" noted by Bartlett & Mohammed (1995).

In the game with positively correlated two-armed bandits, Keller, Rady, Cripps (2005) find two dimensions of inefficiency in any equilibrium: On the one hand, the overall amount of the resource devoted over time to the risky arm conditional on there not having been a breakthrough, the so-called experimentation *amount*, is too low, as is the *intensity* of experimentation, i.e. the resources devoted to the risky arm at a given instant t. Analyzing negatively correlated two-armed bandits, Klein & Rady (2008) find that, while the experimentation intensity may be inefficient, the experimentation *amount* is always at efficient levels. In particular, learning will be complete, i.e. beliefs will almost surely eventually become degenerate at the true state of the world in any equilibrium, if, and only if, efficiency so requires. Here, I show that learning will be complete in any equilibrium for exactly the same parameter range as is the case in Klein & Rady (2008). In the present model, however, complete learning is efficient for a wider set of parameters, as *both* players can reap the benefits of a breakthrough, while in Klein & Rady (2008) one player will be stuck with the losing project.

Having characterized the single-agent and the utilitarian planner's solutions, which are both symmetric, I construct a symmetric Markov perfect equilibrium with the players' common posterior belief as the state variable for all parameter values. For those parameters where learning may be incomplete in equilibrium, I find that the experimentation amount, as well as the intensity, are inefficiently low. This obtains because, as in Keller, Rady, Cripps (2005), there is no encouragement effect in these equilibria,[7] and hence experimentation will

[6] As already mentioned, the negatively correlated case with low stakes provides a notable exception, cf. Klein & Rady (2008).

[7] The encouragement effect was first identified in the Brownian motion model of Bolton & Harris (1999).

stop at the single-agent cutoff rather than the more pessimistic efficient cutoff, which takes into account that *both* players benefit from finding out which project is good. Indeed, as is characteristic of the team production paradigm, individual players do not take into account that their efforts are also benefiting their partner.

The present paper is related to a fast-growing strand of literature on bandits. Whereas the introduction of strategic interaction into the model is due to Bolton & Harris (1999), the use of bandit models in economics harks back to the discrete-time model of Rothschild (1974). While the first papers analyzing strategic interaction featured a Brownian motion model (Bolton & Harris, 1999, 2000), the exponential framework I use has proved itself to be more tractable (cf. Keller, Rady, Cripps, 2005, Keller & Rady, 2007, Klein & Rady, 2008). These previous papers analyzed variants of the two-armed positively correlated model, with the exception of Klein & Rady (2008), who introduced negative correlation into the literature.

While the afore-mentioned papers, as well as the present one, assume both actions and outcomes to be public information, there has been one recent contribution by Bonatti & Hörner (2008) analyzing strategic interaction under the assumption that only outcomes are publicly observable, while actions are private information.[8] Rosenberg, Solan, Vieille (2007), as well as Murto & Välimäki (2006), analyze the two-armed problem of public actions and private outcomes in discrete time, assuming action choices are irreversible.[9]

Bergemann & Välimäki (1996, 2000) analyze strategic experimentation in buyer-seller setups. In their 1996 model, they investigate the case of a *single* buyer facing multiple firms offering a product of differing, and initially unknown, quality, and show that experimentation is efficient in any Markov perfect equilibrium in this setting. With multiple buyers and two firms, one of which offers a product of known quality, whereas the other firm's product quality is initially unknown, equilibrium results in *excessive* experimentation.[10] The reason for this is that price competition leads the "risky" firm to subsidize experimentation beyond efficient levels. If there are many different markets, though, with each having its own, separate, incumbent firm, while the same "risky" firm is active in all the markets, incumbents price

It makes players experiment at beliefs that are more pessimistic than their single-agent cutoff, because they will have a success with a non-zero probability, which will make the other players more optimistic also, thus inducing them to provide more experimentation, from which the first player can then benefit. With fully revealing breakthroughs as in this model, as well as in Keller, Rady, Cripps (2005) and Klein & Rady (2008), however, a player could not care less what others might do after a breakthrough, as there will not be anything left to learn. Therefore, there is no encouragement effect in these models.

[8] Bonatti & Hörner's (2008) is not a full-blown experimentation model, though; indeed, their game stops as soon as there has been a breakthrough, implying that there is no positive value of information. Therefore, no player will ever play risky below his myopic cutoff.

[9] In my model, by contrast, players can switch between bandit arms at any time completely free of costs.

[10] cf. Bergemann & Välimäki (2000).

more aggressively as they also benefit from the experimentation being performed in other markets. Indeed, Bergemann & Välimäki (2000) show that as the number of markets grows large, experimentation tends toward efficient levels.

Recently, there has also been an effort at generalization of existing results in the decision-theoretic bandit literature. For example, Bank & Föllmer (2003), as well as Cohen & Solan (2008), analyze the single-agent problem when the underlying process is a general Lévy process, while Camargo (2007) analyzes the effects of correlation between the arms of a two-armed bandit operated by a single decision maker.

The present paper is also somewhat related to the Moral Hazard in teams literature, to which Holmström (1982) provided the seminal contribution. He found that the introduction of a principal acting as a budget breaker was apt to achieve first best effort levels on the part of team members.

The rest of the paper is structured as follows: Section 2 introduces the model; section 3 analyzes the benchmarks provided by the single agent's and the utilitarian planner's problems; section 4 analyzes some long-run properties of equilibrium learning; section 5 analyzes the non-cooperative game, giving a symmetric Markov perfect equilibrium for all parameter values, and a necessary and sufficient condition for the existence of an efficient equilibrium; section 6 concludes. Proofs are provided in the appendix.

2 The Model

I consider a model of two players, either of whom operates a replica three-armed bandit in continuous time. Bandits are of the exponential type as studied e.g. in Keller, Rady & Cripps (2005). One arm is safe in that it yields a known flow payoff of s; both other arms, A and B, are risky, and, as in Klein & Rady (2008), it is commonly known that exactly one of these risky arms is good and one is bad. The bad risky arm never yields any payoff; the good risky arm yields a positive payoff with a probability of $\lambda\,dt$ if played over a time interval of length dt; the appertaining expected payoff increment amounts to $g\,dt$. The constants λ and g are assumed to be common knowledge between the players. In order for the problem to be interesting, we assume that a good risky arm is better than a safe arm, which is better than a bad risky arm, i.e. $g > s > 0$.

The objective of both players is to maximize their expected discounted payoffs by choosing the fraction of their flow resource they want to allocate to either risky arm. Specifically, either player i chooses a function $(k_{i,A}, k_{i,B}) : \mathbb{R}_+ \to \{(a,b) \in [0,1]^2 : a + b \leq 1\}$ such that $(k_{i,A}, k_{i,B})(t)$ be measurable with respect to the information available at time t, where $k_{i,A}(t)$

$(k_{i,B}(t))$ denotes the fraction of the resource devoted to risky arm A (B) by player i at time t. Throughout the game, either player's actions and payoffs are perfectly observable to the other player. At the outset of the game, the players share a common prior belief that risky arm A is the good one, which I denote by p_0. From these assumptions, it follows by Aumann (1976) that players share a common posterior p_t at all times t. Thus, specifically, player i seeks to maximize his total expected discounted payoff

$$\mathrm{E}\left[\int_0^\infty r\, e^{-r\,t}\, \left[(1-k_{i,A}(t)-k_{i,B}(t))s + (k_{i,A}(t)p_t + k_{i,B}(t)(1-p_t))\, g\right]\, dt\right],$$

where the expectation is taken with respect to the processes $\{p_t\}_{t\in\mathbb{R}_+}$ and $\{(k_{i,A}, k_{i,B})(t)\}_{t\in\mathbb{R}_+}$. As can immediately be seen from this objective function, there are no payoff externalities between the players; the only channel through which the presence of the other player may impact a given player is via his belief p_t, i.e. via the information that the other player is generating. Thus, ours is a game of purely informational externalities.

As only a good risky arm can ever yield a lump sum, breakthroughs are fully revealing. Thus, if there is a lump sum on risky arm A (B) at time τ, then $p_t = 1$ ($p_t = 0$) at all $t > \tau$. If there has not been a breakthrough by time τ, Bayes' Rule yields

$$p_\tau = \frac{p_0 e^{-\lambda \int_0^\tau K_{A,t}\, dt}}{p_0 e^{-\lambda \int_0^\tau K_{A,t}\, dt} + (1-p_0) e^{-\lambda \int_0^\tau K_{B,t}\, dt}},$$

where $K_{A,t} := k_{1,A}(t) + k_{2,A}(t)$ and $K_{B,t} := k_{1,B}(t) + k_{2,B}(t)$. Thus, conditional on no breakthrough having occurred, the process $\{p_t\}_{t\in\mathbb{R}_+}$ will evolve according to the law of motion

$$\dot{p}_t = -(K_{A,t} - K_{B,t})\lambda p_t(1-p_t)$$

almost everywhere.

As all payoff-relevant strategic interaction is captured by the players' common posterior beliefs $\{p_t\}_{t\in\mathbb{R}_+}$, it seems quite natural to focus on Markov perfect equilibria with the players' common posterior belief p_t as the state variable. As is well known, this restriction is without loss of generality in the single agent's and the planner's problems, which are studied in Section 3. In the non-cooperative game, the restriction rules out history-dependent play that is familiar from discrete-time models, yet technically rather intricate to formalize in continuous time.[11] However, it allows us to focus on the experimentation trade-off the players face, thus permitting a direct comparison of our results to those in the previous literature. A Markov strategy for player i is any piecewise continuous function $(k_{i,A}, k_{i,B})$:

[11] See e.g. Bergin & McLeod (1993) for appropriate continuous-time concepts.

$[0,1] \to \{(a,b) \in [0,1]^2 : a+b \leq 1\}$, $p_t \mapsto (k_{i,A}, k_{i,B})(p_t)$, implying that $k_{i,B}(p) - k_{i,A}(p)$ exhibits a finite number of jumps. However, this definition does not guarantee the existence, and even less the uniqueness, of a solution to Bayes' Rule, which now amounts to

$$p_\tau = \frac{p_0 e^{-\lambda \int_0^\tau K_A(p_t)\,dt}}{p_0 e^{-\lambda \int_0^\tau K_A(p_t)\,dt} + (1-p_0) e^{-\lambda \int_0^\tau K_B(p_t)\,dt}},$$

if there has not been a breakthrough by time τ, with $K_A(p_t) := k_{1,A}(p_t) + k_{2,A}(p_t)$ and $K_B(p_t) := k_{1,B}(p_t) + k_{2,B}(p_t)$. Further restrictions on the players' strategy spaces are hence needed to ensure that their actions and payoffs be well-defined and uniquely pinned down after *all* admissible histories. I shall call *admissible* all strategy pairs for which Bayes' rule admits of a solution that coincides with the limit of the unique discrete-time solution. This in effect boils down to ruling out those strategy pairs for which there either is no solution in continuous time, or for which the solution is different from the discrete-time limit.[12]

All that matters for the admissibility of a given strategy pair is the behavior of the function $\Delta(p) := \text{sgn}\{K_B(p) - K_A(p)\}$ at those beliefs $p^\ddagger$ where a change in sign occurs, i.e. where it is not the case that $\lim_{p \uparrow p^\ddagger} \Delta(p) = \Delta(p^\ddagger) = \lim_{p \downarrow p^\ddagger} \Delta(p)$. Given our definition of strategies, there are only finitely many such beliefs $p^\ddagger$, and hence both one-sided limits will exist. By proceeding as in Klein & Rady (2008), one can show that admissibility has to be defined for *pairs* of strategies, i.e. it is impossible to define a player's set of admissible strategies without reference to his opponent's action. Indeed, for any given strategy $(k_{j,A}, k_{j,B})$ of player j, there exists a strategy $(k_{i,A}, k_{i,B})$ of player $i \neq j$ such that the strategy pair is not admissible, as well as a strategy $(\tilde{k}_{i,A}, \tilde{k}_{i,B})$ such that it is admissible. Now, a pair of strategies is admissible if, and only if, it either exhibits no change in sign, or only changes in sign $(\lim_{p \uparrow p^\ddagger} \Delta(p), \Delta(p^\ddagger), \lim_{p \downarrow p^\ddagger} \Delta(p))$ of the following types: $(1,0,1)$, $(0,0,1)$, $(-1,0,1)$, $(-1,0,0)$, $(-1,0,-1)$, $(-1,1,1)$, $(-1,-1,1)$, $(1,0,0)$, $(0,1,1)$, $(0,0,-1)$, $(-1,-1,0)$, $(1,0,-1)$.

Each strategy pair $(k_1, k_2) = ((k_{1,A}, k_{1,B}), (k_{2,A}, k_{2,B}))$ induces a pair of payoff functions (u_1, u_2) with u_i given by

$$u_i(p|k_1,k_2) = \\ 1_{\text{adm.}} \text{E}\left[\int_0^\infty re^{-rt}\left\{(k_{i,A}(p_t)p_t + k_{i,B}(p_t)(1-p_t))g + [1-k_{i,A}(p_t)-k_{i,B}(p_t)]s\right\}dt \middle| p_0 = p\right]$$

for each $i \in \{1,2\}$, where $1_{\text{adm.}}$ is an indicator function that is 1 whenever the strategy pair is admissible. Thus, non-admissible strategy pairs lead to payoffs of $u_1 = u_2 = 0$.

In the subsequent analysis, it will prove useful to make case distinctions based on the stakes at play, as measured by the ratio of the expected payoff of a good risky arm over

[12] It turns out that in this latter case, the continuous-time solution is always unique.

that of a safe arm ($\frac{g}{s}$), the players' impatience (as measured by the discount rate r), and the Poisson arrival rate of a good risky arm λ, which can be interpreted as the players' innate ability at finding out the truth: I say that the stakes are high if $\frac{g}{s} \geq \frac{4(r+\lambda)}{2r+3\lambda}$; stakes are intermediate if $\frac{2r+\lambda}{r+\lambda} < \frac{g}{s} < \frac{4(r+\lambda)}{2r+3\lambda}$; stakes are low if $\frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$; they are very low if $\frac{g}{s} < \frac{2(r+\lambda)}{r+2\lambda}$.

3 Two Benchmarks

3.1 The Single-Agent Problem

I denote by k_A and k_B the fraction of the resource that the single agent dedicates to risky arms A and B, respectively. The law of motion for the state variable is then given by the following expression:

$$\dot{p}_t = -(k_A(p_t) - k_B(p_t))\lambda p_t(1-p_t), \quad \text{for a.a. } t.$$

Straightforward computations show the Bellman equation to be given by[13]

$$u(p) = s + \max_{\{(k_A,k_B)\in[0,1]^2 : k_A+k_B\leq 1\}} \{k_A[B_A(p,u) - c_A(p)] + k_B[B_B(p,u) - c_B(p)]\},$$

where $c_A(p) := s - pg$ and $c_B(p) := s - (1-p)g$ measure the myopic opportunity costs of playing risky arm A (risky arm B) rather than the safe arm; $B_A(p,u) := \frac{\lambda}{r}p[g - u(p) - (1-p)u'(p)]$ and $B_B(p,u) := \frac{\lambda}{r}(1-p)[g - u(p) + pu'(p)]$, by contrast, measure the value of information gleaned from playing risky arm A (or risky arm B, respectively).[14]

Playing risky arm A, e.g., would yield an expected instantaneous payoff of pg rather than s. Thus, a myopic agent, i.e. one who was only interested in maximizing his *current* payoff, would prefer risky arm A over the safe arm if, and only if, $p > p^m$, where $p^m = \frac{s}{g}$ is defined by $c_A(p^m) = 0$. By the same token, he would prefer risky arm B over the safe arm, if, and only if, $p < 1 - p^m$. A far-sighted agent, however, derives a learning benefit over and above the myopic benefit from using either risky arm. Indeed, as the uncertainty is about the distribution underlying the risky arms, the only way for the agent to learn is to play a risky arm. Conceptually, while $\frac{1}{r}$ measures the discounting, $p\lambda[g - u(p)]$, measures the expected value of a potential jump, as λ is the Poisson arrival rate of a breakthrough on risky arm A given that the arm is good while p is the probability that it is good; g is the value the agent jumps to in case of a success, while $u(p)$ is the value he jumps from. The

[13] By standard arguments, if a continuously differentiable function solves the Bellman equation, it is the value function.

[14] By the standard principle of smooth pasting, the agent's payoff function from playing an optimal policy is once continuously differentiable.

second component, $-\lambda p(1-p)u'(p) = u'(p)\,dp$, captures the incremental change in value as a result of the infinitesimal movement in beliefs that is brought about by the agent's playing risky if there is no breakthrough.

As the Bellman equation is linear in the agent's choice variables, it is without loss of generality for me to restrict my attention to corner solutions, for which it is straightforward to derive closed-form solutions for the value function:

If the agent sets $(k_A, k_B)(p) = (0, 0)$, then $u(p) = s$.

If he sets $(k_A, k_B)(p) = (1, 0)$, then his value function satisfies the following ODE:

$$\lambda p(1-p)u'(p) + (r + \lambda p)u(p) = (r + \lambda)pg,$$

which is solved by

$$u(p) = pg + C(1-p)\Omega(p)^{\frac{r}{\lambda}},$$

where C is some constant of integration, and $\Omega(p) := \frac{1-p}{p}$ is the odds ratio.

If he sets $(k_A, k_B)(p) = (0, 1)$, then his value function satisfies the following ODE:

$$\lambda p(1-p)u'(p) - (r + \lambda(1-p))u(p) = -(r + \lambda)(1-p)g,$$

which is solved by

$$u(p) = (1-p)g + Cp\Omega(p)^{-\frac{r}{\lambda}}.$$

If at some belief p both $(k_A, k_B)(p) = (1, 0)$ and $(k_A, k_B)(p) = (0, 1)$ are optimal, then so is $(k_A, k_B)(p) = (\frac{1}{2}, \frac{1}{2})$, and the agent's value amounts to $u(p) = \frac{r+\lambda}{2r+\lambda}g =: \tilde{u}_{11}$.

The optimal policy for the single agent depends on whether the stakes at play, as measured by the ratio $\frac{g}{s}$, exceed the threshold of $\frac{2r+\lambda}{r+\lambda}$ or not. Note that $\frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$ if and only if $p_1^* \geq \frac{1}{2}$, where $p_1^* \equiv \frac{rs}{(r+\lambda)g-\lambda s}\,(1-p_1^*)$ denotes the optimal single-agent cutoff in the standard two-armed problem with one safe and one risky arm A (B).[15]

Proposition 3.1 (Single-Agent Solution for Low Stakes) *If $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$, the single agent will optimally play his risky arm B in $[0, 1-p_1^*[$, his safe arm in $[1-p_1^*, p_1^*]$, and his risky arm A in $]p_1^*, 1]$. His value function is given by*

$$u(p) = \begin{cases} (1-p)g + \frac{\lambda p_1^*}{\lambda p_1^* + r}\left(\Omega(p)\Omega(p_1^*)\right)^{-\frac{r}{\lambda}} pg & \text{if} \quad p \leq 1 - p_1^* \\ s & \text{if} \quad 1 - p_1^* \leq p \leq p_1^* \\ pg + \frac{\lambda p_1^*}{\lambda p_1^* + r}\left(\frac{\Omega(p)}{\Omega(p_1^*)}\right)^{\frac{r}{\lambda}}(1-p)g & \text{if} \quad p \geq p_1^*. \end{cases}$$

This solution continues to be optimal if $\frac{g}{s} = \frac{2r+\lambda}{r+\lambda}$.

[15] cf. Proposition 3.1. in Keller, Rady, Cripps (2005).

The result is illustrated in figure 1. The agent thus optimally behaves as though he was operating a two-armed bandit with one safe arm and one risky arm of that type that is initially more likely to be good. With low enough stakes, therefore, the option value of having an additional risky arm is 0.

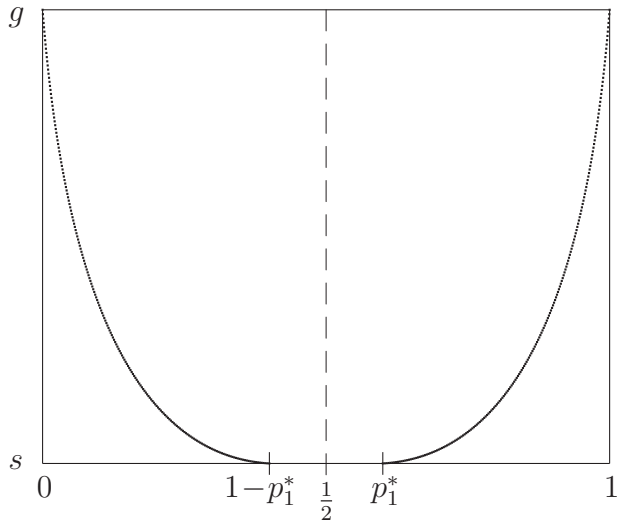

Figure 1: The single-agent value function for $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$.

As is easily verified, the optimal solution implies incomplete learning. Indeed, let us suppose that it is risky arm A that is good. Then, if the initial prior p_0 is in $[0, 1-p_1^*[$, then $\lim_{t\to\infty} p_t = 1 - p_1^*$ with probability 1. If $p_0 \in [1-p_1^*, p_1^*]$, then $p_t = p_0$ for all t, since the agent will always play safe. If $p_0 \in]p_1^*, 1]$, it is straightforward to show that the belief will converge to p_1^* with probability $\frac{\Omega(p_0)}{\Omega(p_1^*)}$, while the truth will be found out (i.e. the belief will jump to 1) with the counter-probability.

If $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$, which is the case if and only if $\tilde{u}_{11} > s$, the single agent will never avail himself of the option to play safe. Specifically, we have the following proposition:

Proposition 3.2 (Single-Agent Solution for Intermediate and High Stakes) *If $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$, the agent will play his risky arm B at all beliefs $p < \frac{1}{2}$ and his risky arm A at all beliefs $p > \frac{1}{2}$. At $p = \frac{1}{2}$, he will split his resources equally between his risky arms. His value function is given by*

$$u(p) = \begin{cases} (1-p)g + p\Omega(p)^{-\frac{r}{\lambda}} \frac{\lambda}{2r+\lambda} g & \text{if} \quad p \leq \frac{1}{2} \\ pg + (1-p)\Omega(p)^{\frac{r}{\lambda}} \frac{\lambda}{2r+\lambda} g & \text{if} \quad p \geq \frac{1}{2}. \end{cases}$$

This solution continues to be optimal if $\frac{g}{s} = \frac{2r+\lambda}{r+\lambda}$.

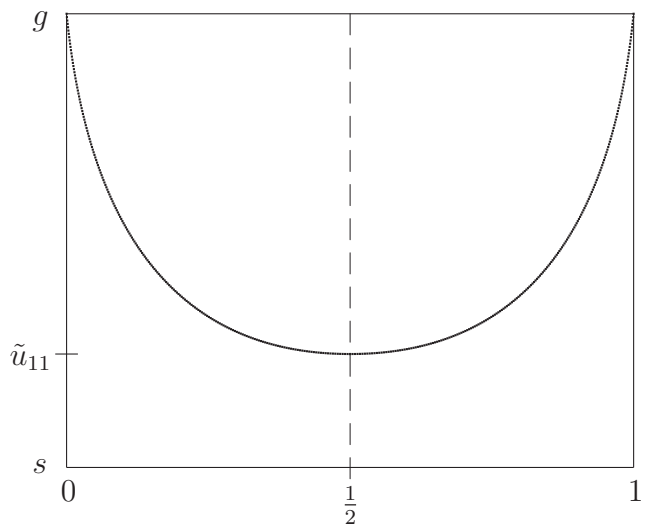

Figure 2: The single agent's value function for $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$.

The result is illustrated in figure 2.

Thus, there now is an option value to having access to the alternative risky project, as for any $p \in [0,1]$, there is now a positive probability of the agent's ending up at $p = \frac{1}{2}$, and thus using the project that initially looked less promising. The single agent's behavior at $p = \frac{1}{2}$ is dictated by the need to ensure a well-defined time path for the belief.[16] Note that whenever stakes exceed the threshold of $\frac{2r+\lambda}{r+\lambda}$, the single agent will make sure learning is complete, i.e. the truth will be found out with probability 1.

3.2 The Planner's Problem

I now turn to the investigation of a benevolent utilitarian planner's solution to the two-player problem at hand. This is the solution the players would commit to at the outset of the game if they had the means to do so. As the planner does not care about the distribution of surplus, and both players are equally apt at finding out the truth, all that matters to him is the sum of resources devoted to both risky arms of type A (B), which I denote by K_A (K_B). Straightforward computations show that the planner's Bellman equation is given by

$$u(p) = s + \max_{\{(K_A,K_B)\in[0,2]^2 : K_A+K_B\leq 2\}} \left\{ K_A \left[B_A(p,u) - \frac{c_A(p)}{2} \right] + K_B \left[B_B(p,u) - \frac{c_B(p)}{2} \right] \right\}.$$

Again, the planner's problem is linear in the choice variables, and we can therefore

[16] cf. also Presman (1990).

without loss of generality restrict our attention to corner solutions.

If $K_A = K_B = 0$ is optimal, $u(p) = s$.

If $K_A = 2$ and $K_B = 0$ is optimal, the Bellman equation is tantamount to the following ODE:

$$2\lambda p(1-p)u'(p) + (2\lambda p + r)u(p) = (2\lambda + r)pg,$$

which is solved by

$$u(p) = pg + C(1-p)\Omega(p)^{\frac{r}{2\lambda}},$$

where C is again some constant of integration.

If $K_A = 0$ and $K_B = 2$ is optimal, the Bellman equation amounts to the following ODE:

$$-2\lambda(1-p)pu'(p) + (2\lambda(1-p) + r)u(p) = (1-p)(r+2\lambda)g,$$

which is solved by

$$u(p) = (1-p)g + Cp\Omega(p)^{-\frac{r}{2\lambda}}.$$

If $(2,0)$ and $(0,2)$, and therefore also $(1,1)$, are optimal, the planner's value satisfies

$$u(p) = \frac{r+2\lambda}{2(r+\lambda)}g =: u_{11}.$$

Which policy is optimal will again depend on the stakes at play, though this time the relevant threshold is different from the single agent's problem, namely $\frac{2(r+\lambda)}{r+2\lambda}$. Note that $\frac{g}{s} \leq \frac{2(r+\lambda)}{r+2\lambda}$ if and only if $p_2^* \geq \frac{1}{2}$, where $p_2^* \equiv \frac{rs}{(r+2\lambda)(g-s)+rs}$.

Proposition 3.3 (Planner's Solution for Very Low Stakes) *If $\frac{g}{s} < \frac{2(r+\lambda)}{r+2\lambda}$, the planner will play the same arm on both bandits at all beliefs. Specifically, he will play arm A on $]p_2^*, 1]$, arm B on $[0, 1-p_2^*[$, and safe on $[1-p_2^*, p_2^*]$. The corresponding payoff function is given by*

$$u(p) = \begin{cases} (1-p)g + \frac{2\lambda p_2^*}{2\lambda p_2^* + r} p \left(\Omega(p)\Omega(p_2^*)\right)^{-\frac{r}{2\lambda}} g & \text{if} \quad p \leq 1-p_2^*, \\ s & \text{if} \quad 1-p_2^* \leq p \leq p_2^*, \\ pg + \frac{2\lambda p_2^*}{2\lambda p_2^* + r}(1-p)\left(\frac{\Omega(p)}{\Omega(p_2^*)}\right)^{\frac{r}{2\lambda}} & \text{if} \quad p \geq p_2^*. \end{cases}$$

This solution continues to be optimal if $\frac{g}{s} = \frac{2(r+\lambda)}{r+2\lambda}$.

The planner's solution thus has pretty much the same structure as the single agent's solution for low stakes; as the latter, it implies incomplete learning. However, it is a different cutoff, namely p_2^*, that is relevant now. p_2^* is always strictly less than p_1^*, and is familiar

from the two-player two-armed bandit problem with perfect positive correlation,[17] where the utilitarian planner will apply the cutoff p_2^*. As in the low-stakes single-agent problem, the value of the risky project that is less likely to be good is so low that it does not play a role in the optimization problem. The planner is more reluctant, though, completely to forsake the less auspicious project, simply because, in case of a success, he gets twice the goodies, so information is more valuable to him than it is to the single agent. This effect is absent in the negatively correlated two-armed bandit case, which is why in Klein & Rady (2008) the relevant cutoff continues to be p_1^* for the planner.

Proposition 3.4 (Planner's Solution for Stakes that Are Not Very Low) *If $\frac{g}{s} > \frac{2(r+\lambda)}{r+2\lambda}$, the planner will play the same arm on both bandits at almost all beliefs. Specifically, he will play arm A on $]\frac{1}{2},1]$ and arm B on $[0,\frac{1}{2}[$. At $p=\frac{1}{2}$, he will split his resources equally between the risky arms. The corresponding payoff function is given by*

$$u(p) = \begin{cases} (1-p)g + \frac{\lambda}{r+\lambda} p\Omega(p)^{-\frac{r}{2\lambda}} g & \text{if} \quad p \leq \frac{1}{2}, \\ pg + \frac{\lambda}{\lambda+r}(1-p)\Omega(p)^{\frac{r}{2\lambda}} g & \text{if} \quad p \geq \frac{1}{2}. \end{cases}$$

This solution continues to be optimal if $\frac{g}{s} = \frac{2(r+\lambda)}{r+2\lambda}$.

At the knife-edge case of $\frac{g}{s} = \frac{2(r+\lambda)}{r+2\lambda}$, the planner is indifferent over all three arms at $p = \frac{1}{2}$. Yet, in order to ensure a well-defined time path of beliefs, he has to set $K_A(\frac{1}{2}) = K_B(\frac{1}{2}) \in [0,1]$.

Note that if the stakes at play are not very low, the planner's solution implies complete learning, i.e. he will make sure the truth will eventually be found out with probability 1. As a matter of fact, our solution is quite intuitive: As the planner does not care which of the risky arms is good, the solution is symmetric around $p = \frac{1}{2}$. Furthermore, it is straightforward to verify that as $\frac{g}{s} \geq \frac{2(r+\lambda)}{r+2\lambda}$, playing risky always dominates the safe arm as $u_{11} \geq s$. However, on account of the linear structure in the Bellman equation, it is always the case that either $(2,0)$ or $(0,2)$ dominates $(1,1)$. Therefore, the only candidate for a solution has the planner switch at $p = \frac{1}{2}$. At the switch point $p = \frac{1}{2}$ itself, the planner's actions are pinned down by the need to ensure a well-defined law of motion of the state variable.

4 Long-Run Equilibrium Learning

Previous literature has noted that with perfectly positively correlated two-armed bandits, learning is always incomplete, i.e. there is a positive probability that the truth will never be

[17] cf. Keller, Rady, Cripps (2005)

found out in finite time. As a matter of fact, Keller, Rady, and Cripps (2005) find that, on account of free-riding incentives, the overall amount of experimentation performed over time is inefficiently low in any equilibrium. On the other hand, Klein & Rady (2008) find that with perfectly negatively correlated bandits, the amount of experimentation is at efficient levels in any equilibrium; in particular, learning will be complete in any equilibrium if and only if efficiency so requires.

The purpose of this section is to derive conditions under which, in our framework, learning will be complete in any equilibrium. To this end, I define as u_1^* the value function of a single agent operating a bandit with only a safe arm and a risky arm A, while I denote by u_2^* the value function of a single agent operating a bandit with only a safe arm and a risky arm B. It is straightforward to verify that $u_2^*(p) = u_1^*(1-p)$ for all p and that[18]

$$u_1^*(p) = \begin{cases} s & \text{if} \quad p \leq p_1^*, \\ pg + (s - p_1^* g)\left(\frac{1-p}{1-p_1^*}\right)^{\frac{r+\lambda}{\lambda}} \left(\frac{p}{p_1^*}\right)^{-\frac{r}{\lambda}} & \text{if} \quad p \geq p_1^* \end{cases}.$$

The following lemma tells us that u_1^* and u_2^* are both lower bounds on players' value functions in *any* equilibrium.

Lemma 4.1 (Lower Bound on Equilibrium Payoffs) *Let u be a player's equilibrium value function. Then, $u(p) \geq \max\{u_1^*(p), u_2^*(p)\}$ for all $p \in [0,1]$.*

The intuition for this result is very straightforward. Indeed, there are only informational externalities, no payoff externalities, in our model. Thus, a player always has the option of ignoring the additional information that he gets for free from the other player, and to behave as though he were by himself. Moreover, a player always has the option of completely forgoing the use of one of his arms. Therefore, he can always guarantee himself the payoff of a single agent operating a two-armed bandit; lower payoffs are incompatible with his playing a best response.

Now, if $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$, then $p_1^* < \frac{1}{2} < 1 - p_1^*$, so at any belief p, we have that $u_1^*(p) > s$ or $u_2^*(p) > s$ or both. Thus, there cannot exist a p such that $(k_{1,A}, k_{1,B})(p) = (k_{2,A}, k_{2,B})(p) = (0,0)$ be mutually best responses as this would mean $u_1(p) = u_2(p) = s$. This proves the following proposition:

Proposition 4.2 (Complete learning) *If $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$, learning will be complete in any Markov perfect equilibrium.*

[18] cf. Prop.3.1 in Keller, Rady, Cripps (2005)

It is the same threshold $\frac{2r+\lambda}{r+\lambda}$ above which complete learning is efficient, and prevails in any equilibrium, in the negatively correlated two-armed bandit case.[19] In our setting, however, complete learning is efficient for a larger set of parameters, as we saw in Proposition 3.4.

Moreover, the planner's solution is an obvious upper bound on players' average equilibrium payoffs. If $\frac{g}{s} < \frac{2(r+\lambda)}{r+2\lambda}$, we know from Proposition 3.4 that the planner's value is s on the non-degenerate interval $[1-p_2^*, p_2^*]$. Since either player can always guarantee himself a payoff of s by playing safe forever, so that s is an obvious lower bound on either player's equilibrium payoffs, this means both players' value must be s on $[1-p_2^*, p_2^*]$ in any equilibrium. Therefore, in any equilibrium, both players uniquely play safe on $[1-p_2^*, p_2^*]$, implying the following proposition:

Proposition 4.3 (Incomplete Learning) *If $\frac{g}{s} < \frac{2(r+\lambda)}{r+2\lambda}$, learning will be incomplete in any equilibrium.*

5 Strategic Problem

Proceeding as before, I find that the Bellman equation for player i ($i \neq j$) is given by[20]

$$u_i(p) = s + k_{j,A}B_A(p,u_i) + k_{j,B}B_B(p,u_i) + \max_{\{(k_{i,A},k_{i,B})\in[0,1]^2 : k_{i,A}+k_{i,B}\leq 1\}} \{k_{i,A}\left[B_A(p,u_i) - c_A(p)\right] + k_{i,B}\left[B_B(p,u_i) - c_B(p)\right]\}.$$

As players are perfectly symmetric in that they are operating two replicas of the same bandit, the Bellman equation for player j looks exactly the same. It is noteworthy that a player only has to bear the opportunity costs of his own experimentation, while the benefits accrue to both, which indicates the presence of free-riding incentives.

On account of the linear structure of the optimization problem, we can restrict our attention to the nine pure strategy profiles, along with three indifference cases per player.

[19] cf. Klein & Rady (2008)

[20] By the smooth pasting principle, player i's payoff function from playing a best response is once continuously differentiable on any open interval on which $(k_{j,A}, k_{j,B})(p)$ in continuous. If $(k_{j,A}, k_{j,B})(p)$ exhibits a jump at p, $u_i'(p)$, which is contained in the definitions of B_A and B_B, is to be understood as the one-sided derivative in the direction implied by the motion of beliefs. In either instance, standard results imply that if for a certain fixed $(k_{j,A}, k_{j,B})$, the payoff function generated by the policy $(k_{i,A}, k_{i,B})$ solves the Bellman equation, then $(k_{i,A}, k_{i,B})$ is a best response to $(k_{j,A}, k_{j,B})$.

Each of these cases leads to a first-order ordinary differential equation (ODE). Details, as well as closed-form solutions, are provided in Appendix A.

5.1 Necessary Conditions for Best Responses

The linearity of the problem provides us with a powerful tool to derive necessary conditions for a certain strategy combination $((k_{1,A}, k_{1,B}), (k_{2,A}, k_{2,B}))$ to be consistent with mutually best responses on an open set of beliefs.[21] As an example, suppose player 2 is playing $(1, 0)$. If player 1's best response is given by $(1, 0)$, it follows immediately from the Bellman equation that it must be the case that $B_A(p, u_1) \geq c_A(p)$ and $B_A(p, u_1) - B_B(p, u_1) \geq c_A(p) - c_B(p)$ for all p in the open interval in question. Moreover, we know that on the open interval in question, the player's value function satisfies

$$2\lambda p(1-p)u_1'(p) + (2\lambda p + r)u_1(p) = (2\lambda + r)pg,$$

which can be plugged into the two inequalities above, yielding a necessary condition for $(k_{1,A}, k_{1,B}) = (1, 0)$ to be a best response to $(k_{2,A}, k_{2,B}) = (1, 0)$. Proceeding in this manner for the possible pure-strategy combinations gives us necessary conditions for a certain pure-strategy combination to be consistent with mutually best responses on an open interval of beliefs, which I report as an auxiliary result in Appendix A.

5.2 Efficiency

Inefficiency because of free-riding has hitherto been a staple result of the literature on strategic experimentation (cf. Bolton & Harris, 1999, 2000, Keller, Rady, Cripps, 2005, Keller & Rady, 2007). Introducing negative correlation into the strategic experimentation literature, Klein & Rady (2008) find that efficient behavior is incentive-compatible if and only if the stakes are low enough. The essential reason for this is as follows: With the stakes low enough, it is clear that the more pessimistic player will never play risky; therefore, the more optimistic player, not having an opportunity to free-ride on his opponent's efforts, will behave efficiently. As a matter of fact, Klein & Rady's (2008) efficient equilibrium disappears as soon as the players' single-agent cutoffs overlap, and free-riding incentives come into play again. Here, though, the opposite result prevails: The efficient solution is incentive-compatible if, and only if, the stakes are *high* enough, as the following proposition shows.

[21] As we keep player j's strategy $(k_{j,A}, k_{j,B})$ fixed on an open interval of beliefs, player i's value function u_i $(i \neq j)$ is of class C^1 on that open interval. Therefore, by standard arguments, u_i solves the Bellman equation on the open interval in question.

Proposition 5.1 (Efficient Equilibrium) *There exists an efficient equilibrium if and only if* $\frac{g}{s} \geq \frac{4(r+\lambda)}{2r+3\lambda}$.

Indeed, the mechanism ensuring existence of an efficient equilibrium for low stakes in Klein & Rady (2008) cannot be at work here, since both players are operating replica bandits. Therefore, if one player has an incentive to experiment given the other player abstains from experimentation, then so does the other player, and free-riding motives enter the picture, no matter how low the stakes might be. One possible intuition for why we here obtain efficiency for high stakes is as follows: For high enough stakes, players would never consider the safe option. Moreover, the efficient policy coincides with the single-agent policy, namely, either implies both players' playing risky, at full throttle, on the arm that is more likely to be good. Therefore, for a player to deviate from this policy in equilibrium, he has to be given special incentives to do so; in the absence of such incentives, e.g. when the other player sticks to the efficient policy, a player's best response calls for his doing the efficient thing also, i.e. there exists an efficient equilibrium. However, for free-riding incentives to be totally eclipsed, stakes have to exceed a threshold that is higher than the one making sure a single agent would never play safe. Indeed, as we have seen, stakes higher than this latter threshold only ensure that learning will be complete in any equilibrium, i.e. while the experimentation *amount* is at efficient levels, the *intensity* does not reach efficient levels as long as $\frac{g}{s} < \frac{4(r+\lambda)}{2r+3\lambda}$.

While it is not surprising that the utilitarian planner, who now has more options, should always be doing better than the planner in Klein & Rady (2008), who could not transfer resources between the two types of risky arm, it may seem somewhat surprising that the players should now be able to achieve even this *higher* efficient benchmark, while they could not achieve the *lower* benchmark in the negatively correlated two-armed model in Klein & Rady (2008). Indeed, with the stakes high enough, free-riding incentives can be overcome completely in non-cooperative equilibrium. While my model abstracts from possible conflicts of interest based on divergent preferences between firms and scientists, this motivational effect may constitute a possible explanation for generous hands-off corporate sponsorships of academic research, or for the morale issues noted e.g. by Bartlett & Mohammed (1995) in the aftermath of a partial revocation of scientists' autonomy at 3M.

5.3 Symmetric Equilibrium for Low And Intermediate Stakes

The purpose of this section is to construct a symmetric equilibrium for those parameter values for which there does not exist an efficient equilibrium. I define symmetry in keeping with Bolton & Harris (1999) as well as Keller, Rady, Cripps (2005):

Definition An equilibrium is said to be *symmetric* if equilibrium strategies $((k_{1,A}, k_{1,B}), (k_{2,A}, k_{2,B}))$ satisfy $(k_{1,A}, k_{1,B})(p) = (k_{2,A}, k_{2,B})(p)\ \forall p \in [0,1]$.

As a matter of course, in any symmetric equilibrium, $u_1(p) = u_2(p)$ for all $p \in [0,1]$. I shall denote the players' common value function by u.

5.3.1 Low Stakes

Recall that the stakes are low if, and only if, the single-agent cutoffs for the two risky arms do not overlap. It can be shown that in this case the symmetric equilibrium in Keller, Rady, and Cripps (Prop. 5.1, 2005) will survive in the sense that there exists an equilibrium that is essentially two copies of the Keller, Rady, and Cripps equilibrium, mirrored at the $p = \frac{1}{2}$ axis. Specifically, we have the following proposition:

Proposition 5.2 (Symmetric MPE for Low Stakes) *If $\frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$, there exists a symmetric equilibrium where both players exclusively use the safe arm on $[1-p_1^*, p_1^*]$, the risky arm A above the belief $\hat{p} > p_1^*$, and the risky arm B at beliefs below $1-\hat{p}$, where $\hat{p}$ is defined implicitly by*

$$\Omega(p^m)^{-1} - \Omega(\hat{p})^{-1} = \frac{r+\lambda}{\lambda}\left[\frac{1}{1-\hat{p}} - \frac{1}{1-p_1^*} - \Omega(p_1^*)^{-1}\ln\left(\frac{\Omega(p_1^*)}{\Omega(\hat{p})}\right)\right].$$

In $[p_1^*, \hat{p}]$, the fraction $k_A(p) = \frac{u(p)-s}{c_A(p)}$ is allocated to risky arm A, while $1-k_A(p)$ is allocated to the safe arm; in $[1-\hat{p}, 1-p_1^*]$, the fraction $k_B(p) = \frac{u(p)-s}{c_B(p)}$ is allocated to risky arm B, while $1-k_B(p)$ is allocated to the safe arm.

Let $V_h(p) := pg + C_h(1-p)\Omega(p)^{\frac{r}{2\lambda}}$, and $V_l(p) := (1-p)g + C_l p\Omega(p)^{-\frac{r}{2\lambda}}$. Then, the players' value function is given by $u(p) = W(p)$ if $1-\hat{p} \leq p \leq \hat{p}$, where $W(p)$ is defined by

$$W(p) := \begin{cases} s + \frac{r}{\lambda}s\left[\Omega(p_1^*)^{-1}\left(1-\frac{p}{p_1^*}\right) - p\ln\left(\frac{\Omega(p)}{\Omega(p_1^*)}\right)\right] & \text{if } 1-\hat{p} \leq p \leq 1-p_1^* \\ s & \text{if } 1-p_1^* \leq p \leq p_1^* \\ s + \frac{r}{\lambda}s\left[\Omega(p_1^*)\left(1-\frac{1-p}{1-p_1^*}\right) - (1-p)\ln\left(\frac{\Omega(p_1^*)}{\Omega(p)}\right)\right] & \text{if } p_1^* \leq p \leq \hat{p} \end{cases};$$

$u(p) = V_h(p)$ if $\hat{p} \leq p$, while $u(p) = V_l(p)$ if $p \leq 1-\hat{p}$, where the constants of integration C_h and C_l are determined by $V_h(\hat{p}) = W(\hat{p})$ and $V_l(1-\hat{p}) = W(1-\hat{p})$, respectively.

Thus, in this equilibrium, even though either player knows that one of his risky arms is good, whenever the uncertainty is greatest, the safe option is attractive to the point that he cannot be bothered to find out which one it is. When players are relatively certain which risky arm is good, they invest all their resources in that arm. When the uncertainty is

of medium intensity, the equilibrium has the flavor of a mixed-strategy equilibrium, with players devoting a uniquely determined fraction of their resources to the risky arm they deem more likely to be good, with the rest being invested in the safe option. As a matter of fact, the experimentation intensity decreases continuously from $k_A(\hat{p}) = 1$ to $k_A(p_1^*) = 0$ (from $k_B(1-\hat{p}) = 1$ to $k_B(1-p_1^*) = 0$). Even though players' Bellman equations are linear in the strategy variable, the equilibrium requires them to use interior levels of experimentation. Intuitively, the situation is very much reminiscent of the classical Battle of the Sexes game: If a player's partner experiments, he would like to free-ride on his efforts; if his partner plays safe, though, he would rather do the experimentation himself than give up on finding out the truth. Now, in symmetric equilibrium, the experimentation intensities are chosen in exactly such a manner as to render the other player indifferent between experimenting and playing safe, thus making him willing to mix over both his options.

Having seen that there exists an equilibrium implying incomplete learning for $\frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$, we are now in a position to strengthen our result on complete learning:

Proposition 5.3 (Complete Learning) *Learning will be complete in any Markov Perfect equilibrium if and only if* $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$.

Klein & Rady (2008) found that with the possible exception of the knife-edge case where $\frac{g}{s} = \frac{2r+\lambda}{r+\lambda}$, learning was going to be complete in any equilibrium if and only if complete learning was efficient. While complete learning obtains in any equilibrium for the exact same parameter set in both models, here, by contrast, we find that if $\frac{2(r+\lambda)}{r+2\lambda} < \frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$, efficiency uniquely calls for complete learning, yet there exists an equilibrium entailing incomplete learning. This is because with three-armed bandits information is more valuable to the planner, as in case of a success he gets the full payoff of a good risky arm. With negatively correlated two-armed bandits, however, the planner cannot shift resources between the two types of risky arm; thus, his payoff in case of a success is just $\frac{g+s}{2}$.

5.3.2 Intermediate Stakes

For intermediate stakes, the equilibrium I construct is essentially of the same structure as the previous one: It is symmetric and it requires players to mix on some interval of beliefs. However, as must be the case by Proposition 4.2, there does not exist an interval where both players play safe, so that players will always eventually find out the true state of the world, even though they do so inefficiently slowly.

Proposition 5.4 (Symmetric MPE for Intermediate Stakes) *If* $\frac{2r+\lambda}{r+\lambda} < \frac{g}{s} < \frac{4(r+\lambda)}{2r+3\lambda}$, *there exists a symmetric equilibrium. Let* $\check{p} := \frac{\lambda+r}{\lambda}(2p^m - 1)$, *and* $\mathcal{W}(p)$ *be defined by*

$$\mathcal{W}(p) := \begin{cases} s + \frac{r+\lambda}{\lambda}(g-s) - \frac{r}{\lambda}ps\,(2 + \ln(\Omega(p))) & \text{if} \quad p \leq \frac{1}{2} \\ s + \frac{r+\lambda}{\lambda}(g-s) - \frac{r}{\lambda}(1-p)s\,(2 - \ln(\Omega(p))) & \text{if} \quad p \geq \frac{1}{2} \end{cases}$$

Now, let $p_1^\dagger > \frac{1}{2}$ *and* $p_2^\dagger > \frac{1}{2}$ *be defined by* $\mathcal{W}(p_1^\dagger) = \frac{\lambda+r(1-p_1^\dagger)}{\lambda+r}g$ *and* $\mathcal{W}(p_2^\dagger) = 2s - p_2^\dagger g$, *respectively. Then, let* $p^\dagger \equiv p_1^\dagger$ *if* $p_1^\dagger \geq \check{p}$; *otherwise, let* $p^\dagger \equiv p_2^\dagger$.

In equilibrium, both players will exclusively use their risky arm A in $[p^\dagger, 1]$, *and their risky arm B in* $[0, 1-p^\dagger]$. *In* $]\frac{1}{2}, p^\dagger]$, *the fraction* $k_A(p) = \frac{\mathcal{W}(p)-s}{c_A(p)}$ *is allocated to risky arm A, while* $1 - k_A(p)$ *is allocated to the safe arm; in* $[p^\dagger, \frac{1}{2}[$, *the fraction* $k_B(p) = \frac{\mathcal{W}(p)-s}{c_B(p)}$ *is allocated to risky arm B, while* $1 - k_B(p)$ *is allocated to the safe arm. At* $p = \frac{1}{2}$, *a fraction of* $k_A(\frac{1}{2}) = k_B(\frac{1}{2}) = \frac{(\lambda+r)g-(2r+\lambda)s}{\lambda(2s-g)}$ *is allocated to either risky arm, with the rest being allocated to the safe arm.*

Let $V_h(p) := pg + C_h(1-p)\Omega(p)^{\frac{r}{2\lambda}}$, *and* $V_l(p) := (1-p)g + C_l p\Omega(p)^{-\frac{r}{2\lambda}}$. *Then, the players' value function is given by* $u(p) = \mathcal{W}(p)$ *in* $[1-p^\dagger, p^\dagger]$, *by* $u(p) = V_h(p)$ *in* $[p^\dagger, 1]$, *and* $u(p) = V_l(p)$ *in* $[0, 1-p^\dagger]$, *with the constants of integration* C_h *and* C_l *being determined by* $V_h(p^\dagger) = \mathcal{W}(p^\dagger)$ *and* $V_l(1-p^\dagger) = \mathcal{W}(1-p^\dagger)$.

Thus, no matter what initial prior players start out from, there is a positive probability beliefs will end up at $p = \frac{1}{2}$, and hence they will try the risky project that looked initially less auspicious. Therefore, in contrast to the equilibrium for low stakes, there is a positive value attached to the option of having access to the second risky project.

6 Conclusion

I have analyzed a game of strategic experimentation with three-armed bandits, where the two risky arms are perfectly negatively correlated. In so doing, I have constructed a symmetric equilibrium for all parameter values. Furthermore, we have seen that any equilibrium is inefficient if stakes are below a certain threshold, and that any equilibrium involves complete learning if stakes are above a certain threshold. In particular, if the stakes are high, there exists an efficient equilibrium and learning will be complete in any equilibrium. If stakes are intermediate in size, all equilibria are inefficient, though they involve complete learning, as required by efficiency. If the stakes are low but not very low, all equilibria are inefficient; there exists an equilibrium that involves incomplete learning, while efficiency requires complete learning. If the stakes are very low, the efficient solution implies incomplete learning; all equilibria involve incomplete learning and are inefficient.

The present paper is merely a first foray into the analysis of strategic experimentation on bandits with more than two arms. The underlying stochastic process was assumed to be Poisson, with a bad risky arm never yielding any payoff. Whether my results are robust to alternative distributional assumptions, such as a non-zero, yet low, arrival rate of a bad risky arm, as in Keller & Rady (2007), or to the assumption of a diffusion process, as in Bolton & Harris (1999, 2000), is an interesting question for future research. To date, even the strategic two-armed bandit problem remains unsolved for general Lévy processes, though Cohen & Solan (2008) analyze the single-agent case.

Furthermore, while Klein & Rady (2008) explore the robustness of their results to certain kinds of asymmetries between players, all the strategic experimentation papers out to date assume players are perfectly symmetric with respect to their "innate" learning abilities, as parameterized by the Poisson arrival rate of breakthroughs, or the diffusion coefficient. In the three-armed case, it could be interesting to explore the additional trade-offs that would arise, if, say, player 1 was able to learn faster on risky arm A, while player 2 was faster with risky arm B. Modeling these trade-offs might yield new insights into the dynamics of specialization in research teams. I intend to explore these questions in future research.

The assumption of perfect negative correlation between the two types of risky arm has allowed me to represent beliefs as elements of the one-dimensional unit interval. Analyzing the case of a general correlation coefficient would imply beliefs evolving in a simplex of dimension greater than 1, and the players' value functions satisfying partial, rather than ordinary, differential equations. It also remains to be investigated how the introduction of private information would affect the analysis, and the conclusions, of the model. I hope to investigate these questions in future work.

Appendix

A Closed-Form Solutions And An Auxiliary Result

If $((0,0),(0,0))$ is played, it is easy to see that $u_1(p) = u_2(p) = s$.

If $((1,0),(1,0))$ is played, both players' value functions satisfy the following ODE:

$$2\lambda p(1-p)u'(p) + (2\lambda p + r)u(p) = (2\lambda + r)pg,$$

which is solved by

$$u(p) = pg + C(1-p)\Omega(p)^{\frac{r}{2\lambda}},$$

where C is some constant of integration.

If $((0,1),(0,1))$ is played, both players' value functions satisfy the following ODE:

$$-2\lambda p(1-p)u'(p) + (2\lambda(1-p) + r)u(p) = (2\lambda + r)(1-p)g,$$

which is solved by

$$u(p) = (1-p)g + Cp\Omega(p)^{-\frac{r}{2\lambda}}.$$

If $((0,1),(1,0))$ is played, player 1's value function is linear:

$$u_1(p) = \frac{\lambda + r(1-p)}{\lambda + r}g.$$

By the same token, player 2's value is also linear,

$$u_2(p) = \frac{\lambda + rp}{\lambda + r}g.$$

Symmetrically, if $((1,0),(0,1))$ is played we have:

$$u_1(p) = \frac{\lambda + rp}{\lambda + r}g,$$

and

$$u_2(p) = \frac{\lambda + r(1-p)}{\lambda + r}g.$$

If $((0,0),(1,0))$ is played, player 1's value satisfies the following ODE:

$$\lambda p(1-p)u'(p) + (\lambda p + r)u(p) = rs + \lambda pg,$$

which is solved by

$$u_1(p) = s + \frac{\lambda}{\lambda + r}p(g-s) + C(1-p)\Omega(p)^{\frac{r}{\lambda}},$$

while player 2's value satisfies

$$\lambda p(1-p)u'(p) + (\lambda p + r)u(p) = (\lambda + r)pg,$$

which is solved by

$$u_2(p) = pg + C(1-p)\Omega(p)^{\frac{r}{\lambda}}.$$

Symmetrically, if $((1,0),(0,0))$ is played, player 1's value satisfies the following ODE:

$$\lambda p(1-p)u'(p) + (\lambda p + r)u(p) = (\lambda + r)pg,$$

which is solved by

$$u_1(p) = pg + C(1-p)\Omega(p)^{\frac{r}{\lambda}}.$$

Meanwhile, player 2's value satisfies:

$$\lambda p(1-p)u'(p) + (\lambda p + r)u(p) = rs + \lambda pg,$$

which is solved by

$$u_2(p) = s + \frac{\lambda}{\lambda + r}p(g-s) + C(1-p)\Omega(p)^{\frac{r}{\lambda}}.$$

If $((0,0),(0,1))$ is played, player 1's value satisfies the following ODE:

$$\lambda p(1-p)u'(p) - (r + \lambda(1-p))u(p) = -rs - \lambda(1-p)g,$$

which admits of the solution

$$u_1(p) = s + \frac{\lambda}{r}g + \frac{\lambda}{r}p[\frac{\lambda}{r}g - (g-s)] + Cp\Omega(p)^{-\frac{r}{\lambda}}.$$

As for player 2, his value evolves according to:

$$\lambda p(1-p)u'(p) - (r + \lambda(1-p))u(p) = -(1-p)(r+\lambda)g,$$

which is solved by

$$u_2(p) = (1-p)g + Cp\Omega(p)^{-\frac{r}{\lambda}}.$$

Symmetrically, if $((0,1),(0,0))$ is played, player 1's value satisfies the following ODE:

$$\lambda p(1-p)u'(p) - (r + \lambda(1-p))u(p) = -(1-p)(r+\lambda)g,$$

which is solved by

$$u_1(p) = (1-p)g + Cp\Omega(p)^{-\frac{r}{\lambda}}.$$

Player 2's value, by contrast, satisfies

$$\lambda p(1-p)u'(p) - (r + \lambda(1-p))u(p) = -rs - \lambda(1-p)g,$$

which admits of the solution

$$u_2(p) = s + \frac{\lambda}{r}g + \frac{\lambda}{r}p[\frac{\lambda}{r}g - (g-s)] + Cp\Omega(p)^{-\frac{r}{\lambda}}.$$

Moreover, there are three indifference cases for player i: He might be indifferent between his risky arm A and his safe arm, between his risky arm B and his safe arm, or between his two risky arms of opposite types.

If player i is indifferent between his safe arm and his risky arm A, his value function satisfies the following ODE:

$$\lambda p(1-p)u'(p) + \lambda p u(p) = (\lambda + r)pg - rs,$$

which is solved by

$$u_i(p) = s + \frac{r+\lambda}{\lambda}(g-s) + \frac{r}{\lambda}s(1-p)\ln\left[\Omega(p)\right] + C(1-p).$$

If player i is indifferent between his safe arm and his risky arm A, his value function satisfies the following ODE:

$$\lambda p(1-p)u'(p) - \lambda(1-p)u(p) = rs - (r+\lambda)(1-p)g,$$

which is solved by

$$u_i(p) = s + \frac{r+\lambda}{\lambda}(g-s) - \frac{r}{\lambda}sp\ln\left[\Omega(p)\right] + Cp.$$

If player i is indifferent between both his risky arms, his value function satisfies the following ODE:

$$2\lambda p(1-p)u'(p) + \lambda(2p-1)u(p) = (\lambda + r)(2p-1)g,$$

which is solved by

$$u_i(p) = \frac{r+\lambda}{\lambda}g + C\sqrt{p(1-p)}.$$

An Auxiliary Result

The logic we discussed in section 5.1 of the main text gives us the following auxiliary result, which will be useful in the proofs of Propositions 4.3, 5.1, and 5.4.

Lemma A.1 *Let $\mathcal{P} \subset]0,1[$ be an open interval of beliefs in which the action profile remains constant, and let $p \in \mathcal{P}$.*

Let $k_j(p) = (0,0)$. Then the following statements hold:

- *If player i's best response is given by $k_i(p) = (0,0)$, then $u_i(p) = s$.*
- *If player i's best response is given by $k_i(p) = (1,0)$ or $k_i(p) = (0,1)$, then $u_i(p) \geq \max\{s, \frac{r+\lambda}{2r+\lambda}g\}$.*

Let $k_j(p) = (1,0)$. Then the following statements hold:

- *If player i's best response is given by $k_i(p) = (0,0)$, then $\frac{\lambda+r(1-p)}{\lambda+r}g \leq u_i(p) \leq 2s - pg$.*
- *If player i's best response is given by $k_i(p) = (1,0)$, then $u_i(p) \geq \max\{\frac{\lambda+r(1-p)}{\lambda+r}g, 2s - pg\}$.*
- *If player i's best response is given by $k_i(p) = (0,1)$, then $u_i(p) = \frac{\lambda+r(1-p)}{\lambda+r}g$ and $p \leq \min\{1 - p^m, \frac{r+\lambda}{2r+3\lambda}\}$.*

Let $k_j(p) = (0,1)$. Then the following statements hold:

- *If player i's best response is given by $k_i(p) = (0,0)$, then $\frac{\lambda+rp}{\lambda+r}g \leq u_i(p) \leq 2s - (1-p)g$.*

- *If player i's best response is given by $k_i(p) = (1,0)$, then $u_i(p) = \frac{\lambda+rp}{\lambda+r}g$ and $p \geq \max\{p^m, \frac{r+2\lambda}{2r+3\lambda}\}$.*
- *If player i's best response is given by $k_i(p) = (0,1)$, then $u_i(p) \geq \max\{\frac{\lambda+rp}{\lambda+r}g, 2s-(1-p)g\}$.*

As $\frac{r+\lambda}{2r+3\lambda} < \frac{1}{2} < \frac{r+2\lambda}{2r+3\lambda}$, the lemma immediately implies that in no equilibrium $((1,0),(0,1))$ or $((0,1),(1,0))$ can arise on an open interval. If furthermore $\frac{g}{s} \geq 2$, and hence $2s - pg \leq \frac{\lambda+r(1-p)}{\lambda+r}g$ for all $p \in [0,1]$, then $((1,0),(0,0))$, $((0,0),(1,0))$, $((0,1),(0,0))$ and $((0,0),(0,1))$ cannot arise on an open interval either.

B Proofs

Proof of Proposition 3.1

The policy (k_A, k_B) implies a well-defined law of motion for the posterior belief. The function u satisfies value matching and smooth pasting at p_1^* and $1-p_1^*$, hence is of class C^1. It is strictly decreasing on $]0, 1-p^*[$ and strictly increasing on $]p^*, 1[$. Moreover, $u = s + B_B - c_B$ on $[0, 1-p^*]$, $u = s$ on $[1-p^*, p^*]$, and $u = s + B_A - c_A$ on $[p^*, 1]$ (I drop the arguments for simplicity), which shows that u is indeed the planner's payoff function from (k_1, k_2).

To show that u and this policy (k_A, k_B) solve the agent's Bellman equation, and hence that (k_1, k_2) is optimal, it is enough to establish that $B_A < c_A$ and $B_B > c_B$ on $]0, 1-p^*[$, $B_A < c_A$ and $B_B < c_B$ on $]1-p^*, p^*[$, and $B_A > c_A$ and $B_B < c_B$ on $]p^*, 1[$. Consider this last interval. There, $u = s + B_A - c_A$ and $u > s$ (by monotonicity of u) immediately imply $B_A > c_A$. It remains to be shown that $B_A - c_A > B_B - c_B$. Using the appertaining differential equation, we have that $B_A - B_B = 2(u - pg) - \frac{\lambda}{r}(g-u)$. It is now straightforward to show that $B_A - B_B > c_A - c_B$ if and only if $u > \frac{r+\lambda}{2r+\lambda}g$. By the afore-mentioned monotonicity properties, we know that $u > s$; yet, $\frac{r+\lambda}{2r+\lambda}g \leq s$ if and only if $\frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$, i.e. if and only if the stakes are low. The other intervals are dealt with in similar fashion. ∎

Proof of Proposition 3.2

The policy (k_A, k_B) implies a well-defined law of motion for the posterior belief. The function u satisfies value matching and smooth pasting at $p = \frac{1}{2}$, hence is of class C^1. It is strictly decreasing on $]0, \frac{1}{2}[$ and strictly increasing on $]\frac{1}{2}, 1[$. Moreover, $u = s + B_B - c_B$ on $[0, \frac{1}{2}]$ and $u = s + B_A - c_A$ on $[\frac{1}{2}, 1]$, which shows that u is indeed the agent's payoff function from (k_1, k_2).

Note that on account of $\tilde{u}_{11} \geq s$, it can never be the case that $0 > \max\{B_A - c_A, B_B - c_B\}$. Thus, all that remains to be shown is that $B_B - c_B > B_A - c_A$ on $]0, \frac{1}{2}[$ and $B_A - c_A > B_B - c_B$ on $]\frac{1}{2}, 1[$. Consider this last interval. Plugging in the relevant ODE, we have that $B_A - c_A = u - s$, and $B_B - c_B = (1 + \frac{\lambda}{r})(g-u) - s$; hence $B_A - c_A > B_B - c_B$ is equivalent to $u > \frac{r+\lambda}{2r+\lambda}g = \tilde{u}_{11} = u(\frac{1}{2})$, which is satisfied on account of the afore-mentioned monotonicity properties. The other interval is dealt with in a similar way. ∎

Proof of Proposition 3.3

The policy (K_A, K_B) implies a well-defined law of motion for the posterior belief. The function u satisfies value matching and smooth pasting at p_2^* and $1-p_2^*$, hence is of class C^1. It is strictly decreasing on $[0, 1-p_2^*]$ and strictly increasing on $[p_2^*, 1]$. Moreover, $u = s+2B_B-c_B$ on $[0, 1-p_2^*]$, $u = s$ on $[1-p_2^*, p_2^*]$, and $u = s+2B_A-c_A$ on $[p_2^*, 1]$, which shows that u is indeed the planner's payoff function from (k_1, k_2).

To show that u and this policy (K_A, K_B) solve the planner's Bellman equation, it is enough to establish that $B_B - \frac{c_B}{2} > \max\{0, B_A - \frac{c_A}{2}\}$ on $]0, 1-p_2^*[$, $0 > \max\{B_A - \frac{c_A}{2}, B_B - \frac{c_B}{2}\}$ on $]1-p_2^*, p_2^*[$, $B_A - \frac{c_A}{2} > \max\{0, B_B - \frac{c_B}{2}\}$ on $]p_2^*, 1[$. Consider this last interval. There, $u = s+2B_A-c_A$ and $u > s$ (by monotonicity of u) immediately imply $2B_A - c_A > 0$. It remains to be shown that $2B_A - c_A > 2B_B - c_B$. Using the appertaining differential equation, we have that $B_A - B_B = u - pg - \frac{\lambda}{r}(g-u)$. It is now straightforward to show that $B_A - B_B > \frac{c_A - c_B}{2}$ if and only if $u > \frac{2\lambda+r}{2(r+\lambda)}g$. By the afore-mentioned monotonicity properties, we know that $u > s$; yet, $s \geq \frac{2\lambda+r}{2(r+\lambda)}g$ if and only if $\frac{g}{s} \leq \frac{2(r+\lambda)}{2\lambda+r}$, i.e. if and only if the stakes are very low. The other intervals are dealt with in similar fashion. ∎

Proof of Proposition 3.4

The policy (K_A, K_B) implies a well-defined law of motion for the posterior belief. The function u satisfies value matching and smooth pasting at $p = \frac{1}{2}$, hence is of class C^1. It is strictly decreasing on $]0, \frac{1}{2}[$ and strictly increasing on $]\frac{1}{2}, 1[$. Moreover, $u = s+2B_B-c_B$ on $[0, \frac{1}{2}]$ and $u = s+2B_A-c_A$ on $[\frac{1}{2}, 1]$, which shows that u is indeed the planner's payoff function from (K_A, K_B).

To show that u and this policy (K_A, K_B) solve the planner's Bellman equation, it is enough to establish that $B_B - \frac{c_B}{2} > \max\{0, B_A - \frac{c_A}{2}\}$ on $]0, \frac{1}{2}[$, and $B_A - \frac{c_A}{2} > \max\{0, B_B - \frac{c_B}{2}\}$ on $]\frac{1}{2}, 1[$. To start out, note that on account of $u_{11} \geq s$, it can never be the case that $0 > \max\{B_A - \frac{c_A}{2}, B_B - \frac{c_B}{2}\}$. Thus, all that remains to be shown is that $B_B - \frac{c_B}{2} > B_A - \frac{c_A}{2}$ on $]0, \frac{1}{2}[$ and $B_A - \frac{c_A}{2} > B_B - \frac{c_B}{2}$ on $]\frac{1}{2}, 1[$. Consider this last interval. Using the appertaining differential equation, we have that $B_A - B_B = u - pg - \frac{\lambda}{r}(g-u)$. It is now straightforward to show that $B_A - B_B > \frac{c_A - c_B}{2}$ if and only if $u > \frac{2\lambda+r}{2(r+\lambda)}g = u_{11}$, which is satisfied on account of the afore-mentioned monotonicity properties and the fact that $u(\frac{1}{2}) = u_{11}$. The other interval is treated in a similar fashion. ∎

Proof of Lemma 4.1

I shall first prove that u_1^* is a lower bound on player i's value function u, writing $B_A^*(p) = B_A(p, u_1^*)$, and $B_B^*(p) = B_B(p, u_1^*)$. Henceforth, I shall suppress arguments whenever this is convenient. Since p_1^* is the single-agent cutoff belief for player 1, we have $u_1^* = s$ for $p \leq p_1^*$ and $u_1^* = s + b_1^* - c_1 = pg + b_1^*$ for $p > p_1^*$. Thus, if $p < p_1^*$, the claim obviously holds as s is a lower bound on u.

Now, let $p \geq p_1^*$. Then, noting that $B_A^* = u_1^* - pg$, we have $B_B^* = \frac{\lambda}{r}[g - u_1^*] - (u_1^* - gp)$. Thus, $B_B^* \geq 0$ if and only if $u_1^* \leq \frac{\lambda+rp}{\lambda+r}g =: w_1(p)$. Let $\tilde{p}$ be defined by $w_1(\tilde{p}) = s$; it is straightforward to show that $\tilde{p} < p_1^*$. Noting furthermore that $u_1^*(p_1^*) = s$, $w_1(1) = u_1^*(1) = g$, and that w_1 is linear whereas u_1^* is strictly convex in p, we conclude that $u_1^* < w_1$ and hence $B_B^* > 0$ on $[p_1^*, 1[$. As a

consequence, we have $u_1^* = pg + B_A^* \leq pg + k_{2,B}B_B^* + B_A^*$ on $[p^*, 1]$.

Now, suppose $u_1 < u_1^*$ at some belief. Since s is a lower bound on u_1, this implies existence of a belief strictly greater than p_1^* where $u_1 < u_1^*$ and $u_1' \leq (u_1^*)'$. This immediately yields $B_A > B_A^* > c_A$, as well as

$$k_{j,A}B_A + k_{j,B}B_B + \max\{B_A - c_A, B_B - c_B, 0\} < \max\{B_A^* - c_A, 0\},$$

which, as $B_A^* \geq 0$ (cf. Keller, Rady, Cripps, 2005), in turn implies $B_B < 0$ and $k_{j,B} = 1$. If $k_{i,B} = 1$, then u would amount to $(1-p)g + 2B_B < (1-p)g$, a contradiction. Therefore, we have $k_{i,A} = 1$, and $u = pg + B_B + B_A$ at the belief in question. But now,

$$u_1 - u_1^* \geq pg + B_B + B_A - (pg + B_B^* + B_A^*) = \frac{\lambda}{r}(u_1^* - u_1) > 0,$$

a contradiction.

An analogous argument applies for u_2^*. ∎

Proof of Proposition 4.3

First, I note that $2s - p_2^* g = 2s - \frac{rsg}{(r+2\lambda)(g-s)+rs}$ and $\frac{\lambda+r(1-p_2^*)}{\lambda+r}g = g - \frac{r}{r+\lambda}\frac{rsg}{(r+2\lambda)(g-s)+rs}$ are strictly bigger than s. As $p \mapsto 2s - pg$ and $p \mapsto \frac{\lambda+r(1-p)}{\lambda+r}g$ are both strictly decreasing in p, this implies that either player i's payoff function satisfies $u_i < \min\{2s - pg, \frac{\lambda+r(1-p)}{\lambda+r}g\}$ on the entire interval $]1-p_2^*, p_2^*[$. By Lemma A.1, this rules out $((1,0),(1,0))$, $((0,1),(0,1))$, $((0,0),(1,0))$ and $((1,0),(0,0))$ on any open subinterval. Noting that $p \mapsto 2s - (1-p)g$ and $p \mapsto \frac{\lambda+rp}{\lambda+r}g$ are both strictly increasing in p, the same calculations rule out $((0,1),(0,0))$ and $((0,0),(0,1))$. Therefore, $((0,0),(0,0))$ uniquely prevails almost everywhere on $]1-p_2^*, p_2^*[$.

Proof of Proposition 5.1

Suppose $\frac{g}{s} \geq \frac{4(r+\lambda)}{2r+3\lambda}$. What is to be shown is that the action profiles $((1,0),(1,0))$ and $((0,1),(0,1))$ are mutually best responses on $]\frac{1}{2}, 1]$, and $[0, \frac{1}{2}[$, respectively. By the characterization of efficiency (cf. Proposition 3.4), both players' respective value function if efficiency prevails is given by:

$$u(p) = \begin{cases} (1-p)g + p\Omega(p)^{-\frac{r}{2\lambda}}\frac{\lambda}{r+\lambda}g & \text{if} \quad p \leq \frac{1}{2} \\ pg + (1-p)\Omega(p)^{\frac{r}{2\lambda}}\frac{\lambda}{r+\lambda}g & \text{if} \quad p \geq \frac{1}{2}. \end{cases}$$

Now, by Lemma A.1, it is sufficient to show that $u(p) > \max\{\frac{\lambda+r(1-p)}{\lambda+r}g, 2s - pg\}$ on $]\frac{1}{2}, 1]$, and $u(p) > \max\{\frac{\lambda+rp}{\lambda+r}g, 2s-(1-p)g\}$ on $[0, \frac{1}{2}[$. I shall only consider the former interval, as the argument pertaining to the latter is perfectly symmetric.

Simple algebra shows that if $\frac{g}{s} \geq \frac{4(r+\lambda)}{2r+3\lambda}$, $w(p) := \frac{\lambda+r(1-p)}{\lambda+r}g \geq 2s - pg$ everywhere in $[\frac{1}{2}, 1]$. Since $u(\frac{1}{2}) = w(\frac{1}{2})$, and u is strictly increasing while w is strictly decreasing in $]\frac{1}{2}, 1[$, the claim follows.

Suppose $\frac{2(r+\lambda)}{r+2\lambda} \leq \frac{g}{s} < \frac{4(r+\lambda)}{2r+3\lambda}$, and define $\tilde{w}(p) := 2s - pg$. It is now straightforward to show that $\tilde{w}(\frac{1}{2}) > w(\frac{1}{2}) = u(\frac{1}{2})$, and, therefore, by Lemma A.1, there exists a neighborhood to the right of $p = \frac{1}{2}$ in which $(1,0)$ is not a best response to $(1,0)$.

Suppose that the stakes are very low, i.e. $\frac{g}{s} < \frac{2(r+\lambda)}{r+2\lambda}$. From our characterization of the efficient solution (cf. Proposition 3.3), we know that $B_A(p_2^*, u) = \frac{c_A(p_2^*)}{2}$, and that the players' value function is given by

$$u(p) = \begin{cases} (1-p)g + \frac{2\lambda p_2^*}{2\lambda p_2^* + r} p \left(\Omega(p)\Omega(p_2^*)\right)^{-\frac{r}{2\lambda}} g & \text{if} \quad p \leq 1 - p_2^*, \\ s & \text{if} \quad 1 - p_2^* \leq p \leq p_2^*, \\ pg + \frac{2\lambda p_2^*}{2\lambda p_2^* + r}(1-p)\left(\frac{\Omega(p)}{\Omega(p_2^*)}\right)^{\frac{r}{2\lambda}} & \text{if} \quad p \geq p_2^*. \end{cases}$$

For the efficient actions to be incentive-compatible, it is necessary that $B_A \geq c_A$ on $]p_2^*, 1]$. Yet, since u is of class C^1, we have that $B_A \to_{p \downarrow p_2^*} \frac{c_A(p_2^*)}{2} < c_A(p_2^*)$, as $p_2^* < p^m$. ∎

Proof of Proposition 5.2

First, I show that $\hat{p}$ as defined in the proposition indeed exists and is unique in $]p_1^*, 1[$. It is immediate to verify that the left-hand side of the defining equation is decreasing, while the right-hand side is increasing in $\hat{p}$. Moreover, for $\hat{p} = p_1^*$, the left-hand side is strictly positive, while the right-hand side is zero. Now, for $\hat{p} \uparrow 1$, the left-hand side tends to $-\infty$, while the right-hand side is positive. The claim thus follows by continuity.

The proposed policies imply a well-defined law of motion for the posterior belief. The function u satisfies value matching and smooth pasting at p_1^* and $1 - p_1^*$, hence is of class C^1. It is strictly decreasing on $]0, 1 - p_1^*]$ and strictly increasing on $]p_1^*, 1[$. Moreover, $u = s + 2B_B - c_B$ on $[0, 1 - \hat{p}]$, $u = s + k_B B_B$ on $[1 - \hat{p}, 1 - p_1^*]$, $u = s$ on $[1 - p_1^*, p_1^*]$, $u = s + k_A B_A$ on $[p_1^*, \hat{p}]$ and $u = s + 2B_A - c_A$ on $[\hat{p}, 1]$, which shows that u is indeed the players' payoff function from $((k_A, k_B), (k_A, k_B))$.

Consider first the interval $]1 - p_1^*, p_1^*[$. It has to be shown that $B_A - c_A < 0$ and $B_B - c_B < 0$. On $]1 - p_1^*, p_1^*[$, we have that $u = s$ and $u' = 0$, and therefore $B_A - c_A = \frac{\lambda + r}{r}(pg - s)$. This is strictly negative if and only if $p < p^m$, which is verified as $p_1^* < p^m$. By the same token, $B_B - c_B = \frac{\lambda + r}{r}\left((1-p)g - s\right)$. This is strictly negative if and only if $p > 1 - p^m$, which is verified as $1 - p^m < 1 - p_1^*$.

Now, consider the interval $]p_1^*, \hat{p}[$. Here, $B_A = c_A$ by construction, as k_A is determined by the indifference condition and symmetry. It remains to be shown that $B_B \leq c_B$ here. Using the relevant differential equation, I find that $B_B = \frac{\lambda}{r}(g - u) + pg - s$. This is less than $c_B = s - (1-p)g$ if and only if $u \geq \frac{\lambda + r}{\lambda} g - \frac{2r}{\lambda} s$. Yet, $\frac{\lambda + r}{\lambda} g - \frac{2r}{\lambda} s \leq s$ if and only if $\frac{g}{s} \leq \frac{2r + \lambda}{r + \lambda}$, so that the relevant inequality is satisfied. The interval $]1 - \hat{p}, 1 - p_1^*[$ is treated in an analogous way.

Finally, consider the interval $]\hat{p}, 1[$. Plugging in the relevant differential equation yields $B_A - B_B = u - pg - \frac{\lambda}{r}(g - u)$. This exceeds $c_A - c_B = (1 - 2p)g$ if and only if $u \geq \frac{\lambda + r(1-p)}{\lambda + r} g$, which is satisfied as $p \mapsto \frac{\lambda + r(1-p)}{\lambda + r} g$ is decreasing and $\frac{\lambda + r(1 - p_1^*)}{\lambda + r} g < s$ whenever $1 - p_1^* < p^m$. The interval $]0, 1 - \hat{p}[$ is dealt with in similar fashion. ∎

Proof of Proposition 5.4

The proposed policies imply a well-defined law of motion for the posterior belief. u is strictly decreasing on $]0, \frac{1}{2}[$ and strictly increasing on $]\frac{1}{2}, 1[$. Furthermore, as $\lim_{p \uparrow \frac{1}{2}} u'(p) = \lim_{p \downarrow \frac{1}{2}} u'(p) = 0$,

the function u is of class C^1. Moreover, $u = s + 2B_B - c_B$ on $[0, 1-p^\dagger]$, $u = s + k_B B_B$ on $[1-p^\dagger, \frac{1}{2}]$, $u = s + k_A B_A$ on $[\frac{1}{2}, p^\dagger]$ and $u = s + 2B_A - c_A$ on $[p^\dagger, 1]$, which shows that u is indeed the players' payoff function from $((k_A, k_B), (k_A, k_B))$.

To establish existence and uniqueness of $p^\dagger$, note that $p \mapsto \frac{\lambda+r(1-p)}{\lambda+r}g$ and $p \mapsto 2s - pg$ are strictly decreasing in p, whereas $\mathcal{W}$ is strictly increasing in p on $]\frac{1}{2}, 1[$. Now, $\mathcal{W}(\frac{1}{2}) = \frac{r+\lambda}{\lambda}g - \frac{2r}{\lambda}s$. This is strictly less than $\frac{\lambda+\frac{r}{2}}{\lambda+r}g$ and $2s - \frac{g}{2}$ whenever $\frac{g}{s} < \frac{4(r+\lambda)}{2r+3\lambda}$. Moreover, $\mathcal{W}(\frac{1}{2})$ strictly exceeds $\frac{\lambda+r(1-p^m)}{\lambda+r}g = g - \frac{r}{r+\lambda}s$ and $2s - p^m g = s$ whenever $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$. Thus, I have established uniqueness and existence of $p^\dagger$ and that $p^\dagger \in]\frac{1}{2}, p^m[$.

By construction, $u > \max\{\frac{\lambda+r(1-p)}{\lambda+r}g, 2s - pg\}$ in $]p^\dagger, 1]$, which, by Lemma A.1, implies that $((1,0),(1,0))$ are mutually best responses in this region; by the same token, $u > \max\{\frac{\lambda+rp}{\lambda+r}g, 2s - (1-p)g\}$ in $[0, 1-p^\dagger[$, which, by Lemma A.1, implies that $((0,1),(0,1))$ are mutually best responses in that region.

Now, consider the interval $]\frac{1}{2}, p^\dagger[$. Here, $B_A = c_A$ by construction, so all that remains to be shown is $B_B \leq c_B$. By plugging in the indifference condition on u', I get $B_B = \frac{\lambda}{r}(g-u) + pg - s$. This is less than $c_B = s - (1-p)g$ if and only if $u \geq \frac{\lambda+r}{\lambda}g - \frac{2r}{\lambda}s = \mathcal{W}(\frac{1}{2}) = u(\frac{1}{2})$, which is satisfied by the monotonicity properties of u. An analogous argument establishes $B_A \leq c_A$ on $]1-p^\dagger, \frac{1}{2}[$. ∎

References

Aghion, P., Dewatripont, M. and Stein, J. (2005): "Academic Freedom, Private-Sector Focus, and the Process of Innovation," Harvard Institute of Economic Research Discussion paper No. 2089.

Aumann, R. (1976): "Agreeing to Disagree," *Annals of Statistics*, Vol. 4, No. 6, 1236–1239.

Bank, P. and H. Föllmer (2003): "American Options, Multi-armed Bandits, and Optimal Consumption Plans: A Unifying View," in: *Paris-Princeton Lectures on Mathematical Finance 2002*, ed. by R. A. Carmona et al.. Springer-Verlag, Berlin and Heidelberg.

Bartlett, C. and Mohammed, A. (1995): "3M: Profile of an Innovating Company," Harvard Business School Case Study 9-395-016.

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.

Bergin, J. and W.B. MacLeod (1993): "Continuous Time Repeated Games," *International Economic Review*, 34, 21–37.

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 (2008): "Collaborating," working paper, Yale University.

Camargo, B. (2007): "Good News and Bad News in Two-Armed Bandits," *Journal of Economic Theory*, 135, 558–566.

Chatterjee, K. and R. Evans (2004): "Rivals' Search for Buried Treasure: Competition and Duplication in R&D," *RAND Journal of Economics*, 35, 160–183.

Cohen, A. and E. Solan (2008): "One-Arm Lévy Bandits," working paper, University of Tel Aviv.

Holmström, B. (1982): "Moral Hazard in Teams," *Bell Journal of Economics*, 13, 324–40.

Keller, G. and S. Rady (2007): "Strategic Experimentation with Poisson Bandits," working paper, University of Oxford and University of Munich.

Keller, G., S. Rady and M. Cripps (2005): "Strategic Experimentation with Exponential Bandits," *Econometrica*, 73, 39–68.

Klein, N. and S. Rady (2008): "Negatively Correlated Bandits," working paper, University of Munich.

Lacetera, N. (2008): "Different Missions and Commitment Power in R & D Organization: Theory and Evidence on Industry-University Alliances," *Organization Science*, published online before print, September, 17, 2008.

Lawler, A. (2003): "Last of the big-time spenders?," *Science*, 299, 330-333.

Manso, G. (2007): "Motivating Innovation," Hudson Institute Research Paper No. 08-01.

Murto, P. and J. Välimäki (2006): "Learning in a Model of Exit," Helsinki Center of Economic Research Working Paper No. 110.

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.