Strategic Learning in Teams*

Nicolas Klein†

This version: May 25, 2011

Abstract

This paper analyzes a two-player game of strategic experimentation with three-armed exponential bandits in continuous time. Players play bandits of identical types, 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. In contrast to the previous literature, I show that in this model the long-run properties of equilibrium learning depend on the stakes at play, and that the efficient policy is incentive-compatible if, and only if, the stakes are high enough.

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

JEL Classification Numbers: C73, D83, O32.

---

*Earlier versions of this paper were circulated under the titles "Free-Riding and Delegation in Research Teams" and "Free-Riding And Delegation in Research Teams—A Three-Armed Bandit Model." I am grateful to the Department of Economics at Yale University and the Cowles Foundation for Research in Economics for their hospitality, and to Sven Rady for advice and encouragement, as well as Dirk Bergemann, Ludwig Ensthaler, Johannes Hörner, Daniel Krähmer, Jo Thori Lind, Benny Moldovanu, Tymofiy Mylovanov, Frank Riedel, Larry Samuelson, Klaus Schmidt, Lones Smith, Eilon Solan, and seminar participants at Bielefeld, Penn State, Southern Methodist University, Yale 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 GRK 801 and SFB/TR 15 is gratefully acknowledged for 2008-09; for 2010, the author thanks the National Research Fund, Luxembourg, for its support of this project.

†University of Bonn, Lennéstr. 37, D-53113 Bonn, Germany; email: kleinnic@yahoo.com.

1 Introduction

When firms explore neighboring oil fields or competing research hypotheses, they have to strike a balance between optimally using their current information on the one hand, and investing in the production of new information on the other hand. When doing so, they have to take into account the impact of their decisions not just on themselves, but on their partners and competitors also; indeed, the latter may benefit from the information a firm produces. Here, I consider a team of two researchers, who each decide whether to investigate a given hypothesis or its negation. Once one of them has found out which hypothesis is true, both benefit from the discovery. Thus, while the costs of experimentation have to be borne privately, any information a researcher produces is a public good. This makes for a situation in which a player's experimentation decisions are strategic, in that they affect the other player's payoffs.

I model this trade-off as a three-armed strategic bandit problem.[1] Specifically, I consider two players operating three-armed exponential bandits in continuous time. 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 them is good. Players are playing exact carbon copies of the same bandit machine; conditional on the state of the world, draws are iid between the players (i.e. players are playing so called *replica bandits*). 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.

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 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 the 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 use 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.

[1] For an overview of the bandit literature, see Bergemann & Välimäki (2008).

I show that the long-run properties of equilibrium learning depend on the payoff advantage of a good risky arm over the safe arm (which I shall henceforth call the *stakes*): Only if the stakes exceed a certain threshold will the overall asymptotic amount of learning be at efficient levels. This is in contrast to the previous literature: Keller, Rady, Cripps (2005) analyze multiple players playing replica exponential bandits with two arms, one safe and one risky. They find that there is always too little experimentation in the long run. Klein & Rady's (2011) setting is as in Keller, Rady, Cripps (2005), except that there are now two players, exactly one of whose risky arms is good.[2] They, by contrast, show that long-run experimentation *amounts* always reach efficient levels; however, the *speed* of learning may still be too low.

Players have to bear experimentation costs privately; the benefit, by contrast, is public. There are thus obvious incentives for players to free-ride on their partner's experimentation. On the other hand, though, the knowledge spill-overs are about options that both players can exploit, which one would think might conceivably make it easier to sustain efficient experimentation. As a matter of fact, I show that free-riding incentives can be completely overcome if, and only if, the stakes exceed a certain threshold; in this case, there exists an equilibrium in which both the amount as well as the speed of learning are at efficient levels. Keller, Rady, Cripps (2005), by contrast, have shown that in the game with positively correlated bandits, the efficient benchmark is never sustainable in equilibrium. In Klein & Rady (2011), however, full efficiency, regarding both the amount and the speed of experimentation, is incentive compatible if, and only if, stakes are *below* a certain threshold. Furthermore, on a more technical level, combinatorial approaches along the lines of those used in Keller, Rady, Cripps (2005), or in Klein & Rady (2011), in order to characterize the full equilibrium set, fail here. Instead, I rely on elementary constructive methods based on the linearity of agents' optimization problems. This enables me to construct a symmetric Markov perfect equilibrium for all parameter values while forgoing a full characterization of the equilibrium set.

My model lets agents choose themselves whether to investigate the hypothesis at hand or its negation; Klein & Rady (2011) by contrast assign one hypothesis each to either player. The comparison between my model and Klein & Rady (2011) should thus allow us to draw inferences about the effect of delegating the choice of project to the researchers themselves. It is notable that, depending on the circumstances, firms or institutions seem to pursue quite different approaches in this respect. Subsequently to marked growth in the number of its research laboratories and facing increasing competitive pressures, 3M, for instance,

[2]Klein & Rady (2011) also show that most results continue to hold in the more general case where both players' risky arms can be bad.

moved to restrict scientists' discretion over their work, which had traditionally been very vast (see 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 (see 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.[3]

Chatterjee & Evans (2004) analyze a treasure-hunting game 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. The game ends as soon as one of the players finds the treasure. Thus, as the winner takes all, their game also involves payoff externalities; in my model, by contrast, externalities are purely informational in nature. Their model is thus 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.

This paper is part of the literature on strategic experimentation with bandits. While bandit models have been analyzed as early as the 1950s (see e.g. Robbins, 1952, Bellman, 1956, Bradt, Johnson, Karlin, 1956), their use in economics harks back to the discrete-time model of Rothschild (1974). Whereas the first papers analyzing strategic interaction featured a Brownian motion model (Bolton & Harris, 1999, 2000), the exponential framework I use was first analyzed by Presman (1990) in a single-agent setting, and has proved itself to be more tractable (see Keller, Rady, Cripps, 2005, Keller & Rady, 2010, Klein & Rady, 2011). In this literature, my paper is most related to Keller, Rady, Cripps (2005) and Klein & Rady (2011). Keller, Rady, Cripps (2005) show that with positively correlated two-armed bandits there does not exist an equilibrium in cutoff strategies,[4] and that the amount, as well as the speed, of learning are inefficiently low in equilibrium. Klein & Rady (2011) show that with perfectly negatively correlated two-armed bandits there are equilibria in cutoff strategies;

[3]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.

[4]A *cutoff strategy* is a strategy of the form "play risky if, and only if, my belief exceeds a given cutoff."

the long-run amount of experimentation is always at efficient levels, though the speed of experimentation may be too low. However, there exists an efficient equilibrium if, and only if, the stakes are below a certain threshold. In my model, players play replica bandits, and both players will have access to both types of risky arm at any time.

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 (2011) analyzing strategic interaction under the assumption that only outcomes are publicly observable, while actions are private information. Rosenberg, Solan, Vieille (2007), as well as Murto & Välimäki (2011), analyze the two-armed problem of public actions and private outcomes in discrete time, assuming action choices are irreversible. 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 (2009), analyze the single-agent problem when the underlying process is a general Lévy process, while Camargo (2007) investigates 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. Manso (2010) embeds a three-armed bandit with two safe arms and one risky arm, operated by a single agent, into a principal-agent 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.

The rest of the paper is structured as follows: Section 2 introduces the model; Section 3 analyzes the utilitarian planner's problem; Section 4 analyzes some long-run properties of equilibrium learning; Section 5 analyzes the non-cooperative game, exhibiting 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, each of whom operates a three-armed bandit in continuous time. One arm is safe in that it yields a known flow payoff of $s > 0$; both other arms, A and B, are risky, and 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 $g\,dt$. Players discount payoffs at the common discount rate $r > 0$.

The constants r, λ, s and g are common knowledge; the only uncertainty is which of the two risky arms is good. The common prior is that A is good with probability p_0. This belief evolves based on the history of experimentation and payoffs. These are commonly observable and so the players continue to have a common belief (probability that arm A is good), which we denote by p_t, at time t.

At each point in time, both players receive a flow endowment of one unit of a perfectly divisible resource. Either player's objective is to maximize his own expected discounted payoffs by choosing the fraction of his endowment flow that he wants to allocate to either risky arm. Specifically, either player i chooses a stochastic process $\{(k_{i,A}, k_{i,B})(t)\}_{0\leq t}$ which is measurable with respect to the information filtration that is generated by the observations available up to time t, with $(k_{i,A}, k_{i,B})(t) \in \{(a,b) \in [0,1]^2 : a + b \leq 1\}$ for all t; $k_{i,A}(t)$ and $k_{i,B}(t)$ denote the fraction of the resource devoted by player i at time t to risky arms A and B, respectively.[5] Throughout the game, either player's actions and payoffs are perfectly observable to the other player. 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

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

[5]Here, putting a fraction of the available resources on a risky project means that the probability of getting a lump-sum reward is reduced at any moment of time, but the size of the reward does not change. It can also be viewed as an approximation for a situation in which only one arm can be pulled at any moment but a policy may change arbitrarily quickly between the arms, spending fraction $k_{i,A}$ ($k_{i,B}$) of the time on arm A (B), for instance.

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 planner's problem, which is studied in Section 3. A Markov strategy for player i is any piecewise continuous function $(k_{i,A}, k_{i,B}) : [0,1] \to \{(a,b) \in [0,1]^2 : a+b \leq 1\}$, $p_t \mapsto (k_{i,A}, k_{i,B})(p_t)$, meaning that it is continuous at all but a finite number of points. Following the approach in Klein & Rady (2011), I shall call *admissible* those strategies for which there exists a solution to the corresponding law of motion of beliefs 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. Given an admissible strategy pair $((k_{1,A}, k_{1,B})(p_t), (k_{2,A}, k_{2,B})(p_t))$, the players' belief is given by

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

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 $K_B(p_t) := k_{1,B}(p_t) + k_{2,B}(p_t)$.

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 my definition of strategies, both one-sided limits will exist. An example of a change in sign corresponding to a non-admissible pair of strategies is given by $(\lim_{p\uparrow p^\ddagger} \Delta(p), \Delta(p^\ddagger), \lim_{p\downarrow p^\ddagger} \Delta(p)) = (0,1,0)$, as for $p_0 = p^\ddagger$, there does not exist a time path of beliefs consistent with Bayes' rule. By contrast, the change in sign $(\lim_{p\uparrow p^\ddagger} \Delta(p), \Delta(p^\ddagger), \lim_{p\downarrow p^\ddagger} \Delta(p)) = (-1,1,0)$ with $p_0 = p^\ddagger$ does admit of a unique time path $\{p_t\}_{0\leq t}$ consistent with Bayes' rule. Indeed, there exists an $\epsilon > 0$ such that this path entails $p_t < p_0$ for all $t \in (0,\epsilon)$; in the discrete-time limit, however, the belief would freeze one grid step above p_0 in the second period. Hence, my definition of admissibility of strategies also rules out changes in sign of the type $(-1,1,0)$.

By proceeding as in Klein & Rady (2011), 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. Now it can be shown that 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 admissible 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

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

for each $i\in\{1,2\}$. For strategy pairs that are not admissible, I set $u_1=u_2=-\infty$.

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 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 The Planner's Problem

First, we investigate a benevolent utilitarian planner's solution to the two-player problem at hand. 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 either type of risky arm, $K_A(p_t)$ and $K_B(p_t)$, respectively. The law of motion for the state variable is now given by

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

Straightforward computations show that the planner's Bellman equation is given by[6]

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

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. By contrast, $B_A(p,u):=$

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

$\frac{\lambda}{r}p[g-u(p)-(1-p)u'(p)]$ and $B_B(p,u) := \frac{\lambda}{r}(1-p)[g-u(p)+pu'(p)]$ measure the value of information gleaned from playing risky arm A (or risky arm B, respectively).[7]

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

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

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

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

The optimal policy depends on whether the stakes at play, as measured by the ratio $\frac{g}{s}$, exceed the threshold of $\frac{2(r+\lambda)}{r+2\lambda}$ or not. Note that $\frac{g}{s} \leq \frac{2(r+\lambda)}{r+2\lambda}$ if and only if $p_2^* \geq \frac{1}{2}$, where $p_2^* := \frac{rs}{(r+2\lambda)(g-s)+rs}$.

Proposition 3.1 (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*

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

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

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

The result is illustrated in figure 1. Note that there is no option value to the initially less promising risky arm, since the planner will never make use of it.

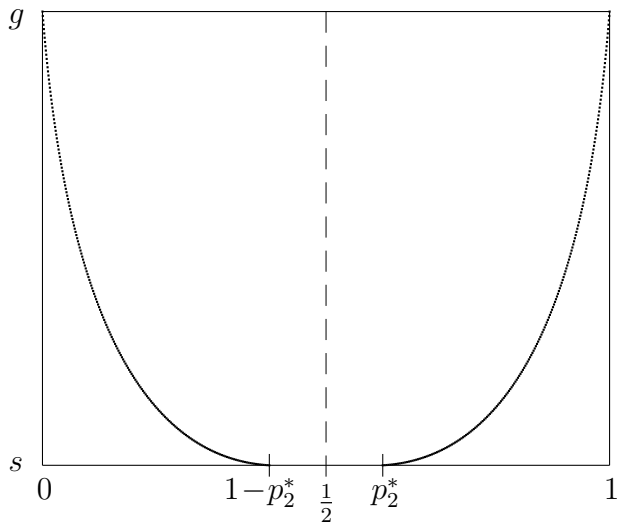

Figure 1: The planner's value function for $\frac{g}{s} < \frac{2(r+\lambda)}{r+2\lambda}$.

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 $[0, 1-p_2^*[$, we have that $\lim_{t\to\infty} p_t = 1 - p_2^*$ with probability 1. If $p_0 \in [1-p_2^*, p_2^*]$, then $p_t = p_0$ for all t, since the planner will always play safe. If $p_0 \in]p_2^*, 1]$, it is straightforward to show that the belief will converge to p_2^* with probability $\frac{\Omega(p_0)}{\Omega(p_2^*)}$, while the truth will be found out (i.e. the belief will jump to 1) with the counter-probability. Hence, there is always a positive probability that the true state of the world will not be found out, i.e. learning is incomplete.

If $\frac{g}{s} > \frac{2(r+\lambda)}{r+2\lambda}$, which is the case if and only if $\overline{u}_{11} > s$, the planner will never avail himself of the option to play safe; his solution will ensure that learning be complete, i.e. that the truth will eventually be found out with probability 1. Specifically, we have the following proposition:

Proposition 3.2 (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*

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

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

The 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 $\overline{u}_{11} \geq s$. 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. Thus, there is now an option value to the initially less promising risky project, as the planner will make use of it with strictly positive probability, no matter what his initial belief $p_0 \in]0,1[$ may be. The result is illustrated in figure 2.

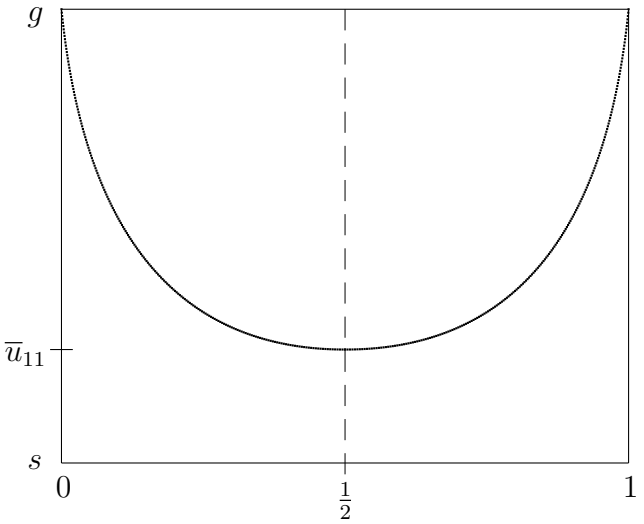

Figure 2: The planner's value function for $\frac{g}{s} > \frac{2(r+\lambda)}{r+2\lambda}$.

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 $K_A(\frac{1}{2}) = K_B(\frac{1}{2}) \in [0,1]$.

The single-agent optimum has the same structure as the planner's solution; all that changes in the relevant differential equations is that 2λ is replaced by λ. Of course, the relevant cutoffs will also change as a result: In the single-agent problem, complete learning will obtain for $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$; for $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$, the agent will switch from risky arm A to the safe arm at the cutoff belief $p_1^* := \frac{rs}{(r+\lambda)g-\lambda s} > p_2^*$, and from risky arm B to the safe arm at $1-p_1^*$. Thus, whenever the stakes are below the relevant threshold, the second risky option does

not play a role; hence, it is not surprising that the same cutoff will be applied as in the problem with two-armed exponential bandits, where p_1^* and p_2^* are the relevant cutoffs in the single-player problem and the planner's problem with two replica bandits, respectively, as Proposition 3.1 in Keller, Rady, Cripps (2005) shows.

4 Long-Run Equilibrium Learning

As already mentioned in the introduction, Keller, Rady, Cripps (2005) identified two dimensions of inefficiency in their model: On the one hand, players give up on finding out about the true state of the world too soon, i.e. the experimentation *amount* is inefficiently small. On the other hand, players also learn too slowly, i.e. the experimentation *intensity* is inefficiently low. If one were merely to focus on the long-run properties of learning, only the former effect would be of interest. Keller, Rady, Cripps (2005) show that, because of the informational externalities, all experimentation stops at the single-agent cutoff belief in any equilibrium; the efficient cutoff belief would be more pessimistic, though, as it takes into account that the information a player generates benefits the other players also.[8] Furthermore, learning is always incomplete, i.e. there is a positive probability that the truth will never be found out.[9] In Klein & Rady (2011), however, the *amount* of experimentation is always at the efficient level.[10] This is because both players cannot be exceedingly pessimistic at the same time. Indeed, as soon as players' single-agent cutoffs overlap, at any possible belief at least one of them is loath to give up completely, although players may not be experimenting with the enthusiasm required by efficiency. In particular, learning will be complete in any equilibrium if and only if efficiency so requires.

This section will show that which of these effects prevails here depends on the stakes at play: If stakes are so high that the single-agent cutoffs overlap, players would not be willing ever completely to give up on finding out the true state of the world even if they were by themselves. Yet, since all a player's partner is doing is to provide him some additional

[8] By contrast, Bolton & Harris (1999) identified an *encouragement effect* in their model. It makes players experiment at beliefs that are more pessimistic than their single-agent cutoffs. This is because they will have a success with a non-zero probability, which will make the other players more optimistic also. This then induces them to provide more experimentation, from which the first player then benefits in turn. With fully revealing breakthroughs as in Keller, Rady, Cripps (2005), Klein & Rady (2011), or this model, 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.

[9] The efficient solution in Keller, Rady, Cripps (2005) also implies incomplete learning.

[10] For perfect negative correlation, this is true in any equilibrium; for general negative correlation, there always exists an equilibrium with this property.

information for free, a player should be expected to do at least as well as if he were by himself. Hence, the Klein & Rady (2011) effect obtains if players' single-agent cutoffs overlap, and, in any equilibrium (in which at least one player's value function is smooth),[11] the true state of the world will eventually be found out with probability 1 (i.e. learning will be *complete*), as efficiency requires. In the opposite case, however, the informational externality identified by Keller, Rady, Cripps (2005) carries the day, and, as we will see in the next section, there exists an equilibrium entailing an inefficiently low amount of experimentation. For some parameters, this implies incomplete equilibrium learning while efficiency calls for complete learning.

To state our next lemma, I write u_1^* for 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[12]

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

The following lemma tells us that u_1^* and u_2^* are both lower bounds on a player's value in *any* equilibrium, provided his value is smooth.

Lemma 4.1 (Lower Bound on Equilibrium Payoffs) *Let $u \in C^1$ 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, intuitively, a player can only benefit from any information his opponent provides him for free; therefore, he should be expected to do at least as well as if he were by himself, forgoing the use of one of his risky arms to boot.

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 $(k_{1,A}, k_{1,B})(p) = (k_{2,A}, k_{2,B})(p) =$

[11] The technical requirement that at least one player's value function be C^1 is needed on account of complications pertaining to the admissibility of strategies. I use it in the proof of Lemma 4.1 to establish that the safe payoff s constitutes a lower bound on the player's equilibrium value. However, by e.g. insisting on playing $(1,0)$ at a single belief $\hat{p}$ while playing $(0,0)$ everywhere else in a neighborhood of $\hat{p}$, a player could e.g. force the other player to play $(0,1)$ at $\hat{p}$ for mere admissibility reasons. Thus, both players' *equilibrium* value functions might be pushed below s at certain beliefs $\hat{p}$. For the purposes of this section, I rule out such implausible behavior by restricting attention to equilibria in which at least one player's value function is smooth.

[12] See Prop.3.1 in Keller, Rady, Cripps (2005).

$(0,0)$ 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 in which at least one player's value function is of class* C^1.

It is the same threshold $\frac{2r+\lambda}{r+\lambda}$ above which complete learning is efficient, and prevails in any equilibrium, in the perfectly negatively correlated two-armed bandit case.[13] In our setting, however, complete learning is efficient for a larger set of parameters, as we saw in Proposition 3.2. In the following section, I shall proceed to a more thorough analysis of the strategic problem.

5 Equilibria of the Non-Cooperative Game

5.1 The Bellman Equation

Proceeding as before, I find that the Bellman equation for player i $(i \neq j)$ is given by[14]

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

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. For future reference, I define the myopic cutoff belief $p^m := \frac{s}{g}$ by $c_A(p^m) = 0$. A player who was only interested in maximizing his current payoff would switch from risky arm A (B) to the safe arm at p^m $(1 - p^m)$.

[13] See Proposition 8 in Klein & Rady (2011).

[14] 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 $(k_{j,A}, k_{j,B})$.

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. 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.

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.[15] 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 in 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. I report these necessary conditions as an auxiliary result in Appendix A.

5.2 Efficiency

As already mentioned, the planner's solution is compatible with equilibrium if and only if stakes exceed a certain threshold. This may at first glance seem surprising given that a player provides a positive informational externality through his experimentation; indeed, the information he generates helps his partner make better decisions in turn. This is the reason why efficiency is not sustainable in equilibrium in Keller, Rady, Cripps (2005). In Klein & Rady (2011), this calculation changes, though, when the stakes are so low that the players' respective single-agent cutoffs do not overlap: In this case, the more pessimistic player will never play risky under any circumstances, which the more optimistic player will anticipate, and hence behave efficiently. However, the efficient equilibrium disappears as soon as the relevant single-agent cutoffs overlap and free-riding incentives kick in again.

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 (2011), who could not transfer

[15] 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.

resources between the two types of risky arm, it may seem somewhat surprising that, for high stakes, the players should now be able to achieve even this *higher* efficient benchmark, while they could not achieve the *lower* benchmark in the perfectly negatively correlated two-armed model in Klein & Rady (2011). Indeed, with the stakes high enough, free-riding incentives can be overcome completely in non-cooperative equilibrium, as the following proposition shows.

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

Since players are playing replica bandits, there will never arise a situation in which one player is optimistic while the other one is pessimistic; as soon as one player finds it optimal to experiment in isolation then so will the other player, and free-riding incentives enter the picture again. Therefore, the Klein & Rady (2011) channel effecting efficiency cannot be at work here, no matter what the stakes might be. For high stakes, a different channel will kick in, though: On account of perfect negative correlation between the risky arms, players will never simultaneously be very pessimistic about both prospects. Hence, for stakes above a certain threshold, they would never consider the safe option. Moreover, since there are no switching costs in my model, players would use the risky arm that looks momentarily more promising if they were left to their own devices. Thus, in the absence of specific incentives to deviate from this policy, they would do what efficiency requires. In particular, if the other player behaves efficiently, a player's best response calls for behaving efficiently also; i.e. there exists an efficient equilibrium.[16]

Note that the relevant threshold above which free-riding incentives are totally eclipsed is lower than 2 (above which experimentation becomes costless). This is because players are not myopic and take the learning benefit of experimentation into account, at least to the extent it benefits the player himself. Thus, it is no surprise that the relevant threshold should be increasing in the players' impatience r, and decreasing in the informativeness of experimentation, as measured by λ. However, note that 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 in which at least one

[16]In his canonical Moral Hazard in Teams paper, Holmström (1982) shows that a team cannot produce efficiently in the absence of a budget-breaking principal, on account of payoff externalities between team members. By contrast, my analysis shows that, in a model with purely informational externalities in which players can choose whether to investigate a given hypothesis or its negation, the efficient solution becomes incentive compatible if the stakes at play exceed a certain threshold. In the treasure-hunting game, Chatterjee & Evans (2004) also show efficiency can be sustained in equilibrium, assuming the treasure is big enough.

player's value function is smooth; i.e. while the experimentation *amount* is at efficient levels, the *intensity* does not reach efficient levels as long as $\frac{g}{s} < \frac{4(r+\lambda)}{2r+3\lambda}$.

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 $(k_{1,A}, k_{1,B})(p) = (k_{2,A}, k_{2,B})(p)$ for all $p \in [0,1]$.

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 there exists an equilibrium that is essentially two copies of the Keller, Rady, Cripps (2005) symmetric equilibrium (see their Proposition 5.1), 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

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

$u(p) = V_h(p)$ 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 $k_B(1 - p_1^*) = 0$). Intuitively, the situation is very much reminiscent of the classical Battle of the Sexes game: If one's partner experiments, one would like to free-ride on his efforts; if one's partner plays safe, though, one would rather do the experimentation oneself than give up on finding out the truth. On the relevant range of beliefs it is the case that as players become more optimistic, they have to raise their experimentation intensities in order to increase free-riding incentives for their partner. This is necessary to keep their partner indifferent, and hence willing to mix, over both options.

Having seen that for $\frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$, there exists an equilibrium with smooth value functions that implies incomplete learning, we are now in a position to strengthen our result on the long-run properties of equilibrium learning:

Corollary 5.3 (Complete Learning) *Learning will be complete in any Markov Perfect equilibrium in which at least one player's value function is smooth, if and only if $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$.*

For perfect negative correlation, Klein & Rady (2011) 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 the proposition pertains to the exact same parameter set on which complete learning prevails in Klein & Rady (2011), we here find by contrast 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 $\frac{g+s}{2}$.

Thus, while our analysis would unambiguously suggest that, if stakes were high, delegating project choice to the agents was a good idea since it increases experimentation intensities,

this conclusion need not hold for $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$. For this case, Klein & Rady (2011) have shown that if agents are assigned one of the projects each, the unique equilibrium features an experimentation intensity of 1 for the more promising project throughout $[0, 1-p_1^*[\cup]p_1^*, 1]$. By contrast, in the equilibrium we discussed in Proposition 5.2, the overall experimentation intensity increases continuously from 0 at p_1^* to 2 at $\hat{p}$ (decreases continuously from 2 at $1-\hat{p}$ to 0 at $1-p_1^*$). Thus, for initial beliefs just above p_1^*, for instance, the rate of experimentation may be higher if scientists do not have the freedom to choose the hypothesis they are working on. Hence, if the stakes are low, as arguably they might be at a company like 3M, which makes products such as adhesives and abrasives, it might make sense to restrict scientists' discretion.

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, 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*

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

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* $p^\dagger := p_1^\dagger$ *if* $p_1^\dagger \geq \check{p}$; *otherwise, let* $p^\dagger := p_2^\dagger$.

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 that 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 in which at least one player's value is smooth 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 in which at least one player's value function is smooth. If stakes are intermediate in size, all equilibria are inefficient, though they involve complete learning (provided both players' value functions are not kinked), as required by efficiency. If the stakes are low, all equilibria are inefficient, and there exists an equilibrium implying an inefficiently low amount of experimentation. In particular, if the stakes are low but not very low, there exists an equilibrium that involves incomplete learning while efficiency requires complete learning; if the stakes are very low, the efficient solution also implies incomplete learning.

While I have only investigated the case of *perfect* negative correlation, the impact of general pessimism à la Klein & Rady (2011) on the existence of an efficient equilibrium might constitute an interesting object of further investigation. It seems clear that, in this problem, the planner's solution would feature $(0,0)$ on $[0,p_2^*]^2$, $(2,0)$ for $p_A > \max\{p_B, p_2^*\}$, $(0,2)$ for $p_B > \max\{p_A, p_2^*\}$, and $(1,1)$ for $p_A = p_B > p_2^*$. One would have to expect that $((1,0),(1,0))$ could not be sustained in equilibrium for $p_A > \max\{p_B, p_2^*\}$ and p_A close to p_2^* for the same reasons as in this paper. However, $((1,0),(1,0))$ would clearly prevail in a neighborhood of $(p_A, p_B) = (1,0)$. As is easy to see from the appertaining laws of motion, $\frac{p_{B,t}}{1-p_{A,t}}$ would remain constant in this neighborhood. One might now expect that if the ratio of initial beliefs $\frac{p_{B,0}}{1-p_{A,0}}$ was close enough to 1, $((1,0),(1,0))$ could be sustained along the ray $p_{B,t} = (1-p_{A,t})\frac{p_{B,0}}{1-p_{A,0}}$ all the way to the hyperplane $p_{A,t} = p_{B,t}$. Once this hyperplane is reached, admissibility considerations should guarantee the implementability of $((\frac{1}{2},\frac{1}{2}),(\frac{1}{2},\frac{1}{2}))$ for all $p_{A,t} = p_{B,t} \in [p_2^*, \frac{1}{2}]$. This would imply that the efficient solution was incentive compatible if, and only if, both hypotheses' being wrong was *initially* very unlikely. I leave a full exploration of these conjectures for future work.

Furthermore, it could be interesting to explore the additional trade-offs arising when players differed with respect to their innate learning abilities, as parameterized by the Poisson arrival rate of breakthroughs. Analyzing these additional trade-offs that would appear, if, say, player 1 was able to learn faster on risky arm A, while player 2 was faster with risky arm B might yield insights into conditions under which there is excessive, or insufficient, specialization in equilibrium. I intend to explore these questions in future research.

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

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

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

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

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

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

If player i is indifferent between his safe arm and his risky arm B, 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

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

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 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_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 3.2

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, $u = s+2B_B-c_B$ 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 $\overline{u}_{11} \geq s$, 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 $u > \frac{2\lambda+r}{2(r+\lambda)}g = \overline{u}_{11}$, which is satisfied on account of the afore-mentioned monotonicity properties and the fact that $u(\frac{1}{2}) = \overline{u}_{11}$. The other interval is treated in a similar fashion. ∎

Proof of Lemma 4.1

In a first step, I show that s is a lower bound on u. Assume to the contrary that there exists a belief $p^\dagger \in]0,1[$ such that $u(p^\dagger) < s$. Then, since u is C^1 and $u(0) = u(1) = g > s$, there exists a belief $\tilde{p} \in]0,1[$ such that $u(\tilde{p}) < s$ and $u'(\tilde{p}) = 0$. I write B_A and B_B for $B_A(p,u)$ and $B_B(p,u)$, respectively, suppressing arguments whenever this is convenient. Moreover, I define $\hat{B}_A(p) := \frac{\lambda}{r}p(g-s) > 0$ and $\hat{B}_B(p) := \frac{\lambda}{r}(1-p)(g-s) > 0$, while denoting by $(k_{j,A}, k_{j,B})$ the other player's action at $\tilde{p}$ in the equilibrium underlying the value function u. Now, at $\tilde{p}$, $u < s$ immediately implies $B_A = \frac{\lambda}{r}\tilde{p}(g-u) > \hat{B}_A$ and $B_B = \frac{\lambda}{r}(1-\tilde{p})(g-u) > \hat{B}_B$, and we have that

$$u - s \geq k_{j,A}(B_A - \hat{B}_A) + k_{j,B}(B_B - \hat{B}_B) = (k_{j,A}\tilde{p} + k_{j,B}(1-\tilde{p}))(s-u) \geq 0,$$

a contradiction to $u < s$.[17] Thus, we have already shown that u_1^* bounds u from below at all beliefs $p \leq p_1^*$.

Now, suppose there exists a belief $p > p_1^*$ at which $u < u_1^*$. I now write $B_A^* := \frac{\lambda}{r}p[g - u_1^* - (1-p)(u_1^*)'(p)] = u_1^* - pg$ and $B_B^* := \frac{\lambda}{r}(1-p)[g - u_1^* + p(u_1^*)'(p)]$. Since $B_A^* + B_B^* = \frac{\lambda}{r}(g - u_1^*)$, and hence $B_B^* = \frac{\lambda}{r}(g-u_1^*) - (u_1^* - pg)$, we have that $B_B^* \geq 0$ 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 $[p_1^*, 1[$. Moreover, since $B_A^* \geq 0$ (see Keller, Rady, Cripps, 2005), we have $u_1^* = pg + B_A^* \leq pg + k_{j,B}B_B^* + (1 + k_{j,A})B_A^*$ on $[p^*, 1]$, for any $(k_{j,A}, k_{j,B})$.

Since s is a lower bound on u, by continuity, $u(p) < u_1^*(p)$ implies the existence of a belief strictly greater than p_1^* where $u < u_1^*$ and $u_1' \leq (u_1^*)'$. This immediately yields $B_A > B_A^* > c_A$, as well as

$$\begin{aligned} u - u_1^* &\geq pg + k_{j,B}B_B + (1 + k_{j,A})B_A - [pg + (1 + k_{j,A})B_A^* + k_{j,B}B_B^*] \\ &= k_{j,B}(B_A + B_B - B_A^* - B_B^*) + (1 + k_{j,A} - k_{j,B})(B_A - B_A^*) \\ &= k_{j,B}\frac{\lambda}{r}(u_1^* - u_1) + (1 + k_{j,A} - k_{j,B})(B_A - B_A^*) > 0, \end{aligned}$$

a contradiction.[18]

An analogous argument applies for u_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. At $p = \frac{1}{2}$, admissibility uniquely pins

[17] Strictly speaking, the first inequality relies on the admissibility of the action $(0,0)$ at $\tilde{p}$. However, even if $(0,0)$ should not be admissible at $\tilde{p}$, my definition of strategies still guarantees the existence of a neighborhood of $\tilde{p}$ in which $(0,0)$ is admissible everywhere except at $\tilde{p}$. Hence, by continuous differentiability of u, there exists a belief $\tilde{\tilde{p}}$ in this neighborhood at which the same contradiction can be derived.

[18] Again, strictly speaking, the first inequality relies on the admissibility of the action $(1,0)$ at the belief in question, and my previous remark applies.

down a player's response to the other player's action. By the characterization of efficiency (see Proposition 3.2), both players' respective value function if efficiency prevails is given by:

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

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 (see Proposition 3.1), we know that $B_A(p_2^*, u) = \frac{c_A(p_2^*)}{2}$, and that the players' value function is given by

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

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 $\lim_{p \downarrow p_2^*} B_A(p, u) = \frac{c_A(p_2^*)}{2} < c_A(p_2^*)$, 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. It is immediate to verify that the function u satisfies value matching and smooth pasting at p_1^* and $1 - p_1^*$. To show that it is of class C^1, it remains to be shown that smooth pasting is satisfied at $\hat{p}$ and $1 - \hat{p}$. From the appertaining ODEs, we have that

$$\lambda\hat{p}(1-\hat{p})u'(\hat{p}-) + \lambda\hat{p}u(\hat{p}) = (\lambda + r)\hat{p}g - rs$$

and

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

where I write $u'(\hat{p}-) := \lim p \uparrow \hat{p} u'(p)$ and $u'(\hat{p}+) := \lim p \downarrow \hat{p} u'(p)$. Now, $u'(\hat{p}-) = u'(\hat{p}+)$ if and only if $u(\hat{p}) = 2s - \hat{p}g$. Now, algebra shows that indeed $W(\hat{p}) = 2s - \hat{p}g$. By symmetry, we can thus conclude that $W(1-\hat{p}) = 2s - (1-\hat{p})g$ and that u is of class C^1. Furthermore, 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 $B_A - c_A = \frac{\lambda+r}{r} pg - \frac{\lambda p + r}{r} s$. This is strictly negative if and only if $p < p_1^*$. By the same token, $B_B - c_B = \frac{\lambda+r}{r}(1-p)g - \frac{\lambda(1-p)+r}{r} s$. This is strictly negative if and only if $p > 1 - p_1^*$.

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 $u \geq \frac{\lambda + r(1-p)}{\lambda+r} g$. At $\hat{p}$, the indifference condition gives us $k_A(\hat{p}) = 1$, which implies $u(\hat{p}) = 2s - \hat{p}g$. Since $p \mapsto \frac{\lambda + r(1-p)}{\lambda+r} g$ is decreasing and u is increasing, it is sufficient for us to show that $u(\hat{p}) \geq \frac{\lambda + r(1-\hat{p})}{\lambda+r} g$, which is equivalent to $\hat{p} \leq \frac{\lambda+r}{\lambda}(2p^m - 1)$. From the indifference condition for the experimentation intensity $\tilde{k}_A(p) := \frac{u(p)-s}{c_A(p)}$, we see that $\tilde{k}_A$ is strictly increasing on $]p_1^*, p^m[$, and that $\lim_{p \uparrow p^m} \tilde{k}_A(p) = +\infty$; hence $\hat{p} < p^m$. Therefore, it is sufficient to show that $p^m \leq \frac{\lambda+r}{\lambda}(2p^m - 1)$, which is equivalent to $\frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$. ∎

Proof of Proposition 5.4

The proposed policies imply a well-defined law of motion for the posterior belief. The function 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 for 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.

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.

Bellman, R. (1956): "A Problem in the Sequential Design of Experiments," *Sankhya: The Indian Journal of Statistics (1933–1960)*, Vol. 16, No. 3/4, 221–229.

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.

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 (2011): "Collaborating," *American Economic Review*, 101(2), 632–663.

Bradt, R., S. Johnson and S. Karlin (1956): "On Sequential Designs for Maximizing the Sum of n Observations," *The Annals of Mathematical Statistics*, 27, 1060–1074.

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 (2009): "Bandit Problems with Lévy Payoff Processes," working paper, University of Tel Aviv, archived at http://arxiv.org/abs/0906.0835v1.

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

Keller, G. and S. Rady (2010): "Strategic Experimentation with Poisson Bandits," Theoretical Economics, 5, 275–311.

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

Klein, N. and S. Rady (2011): "Negatively Correlated Bandits," *Review of Economic Studies*, 78(2), 693–732.

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. (2010): "Motivating Innovation," *Journal of Finance*, forthcoming.

Murto, P. and J. Välimäki (2011): "Learning in a Model of Exit," *Review of Economic Studies*, forthcoming.

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

Robbins, H. (1952): "Some Aspects of the Sequential Design of Experiments," *Bulletin of the American Mathematical Society*, 58, 527–535.

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.