Learning and Experimentation in Strategic Bandit Problems

Inaugural-Dissertation
zur Erlangung des Grades Doctor oeconomiae publicae
(Dr. oec. publ.)
an der Ludwig-Maximilians-Universität München

2010

vorgelegt von Nicolas Alexandre Klein

Referent: Prof. Sven Rady, PhD

Korreferent: Prof. Dr. Klaus Schmidt

Promotionsabschlussberatung: 17. November 2010

Mündliche Prüfung am 3. November 2010

Berichterstatter:

Prof. Sven Rady, PhD

Prof. Dr. Klaus Schmidt

Prof. Ray Rees

1 Introduction

In my dissertation, I deal with dynamic models of strategic information accumulation and transmission, meaning that I investigate situations in which strategically interacting agents seek progressively to garner intelligence, which may potentially help them make better decisions in the future. The canonical framework for the analysis of such questions is provided by the literature on strategic experimentation on (two-armed) bandits.[1] Previous related work has been done in the papers by Bolton & Harris (1999, 2000), as well as Keller, Rady, Cripps (2005), and Keller & Rady (2010). These papers investigated the case of several agents experimenting with replica bandits, with both the players' actions, as well as the outcomes of their actions, being perfectly publicly observable to their peers.

This assumption that there is perfect positive correlation across the several players' bandits is lifted in Chapter 1 of the present dissertation.[2] First investigating the case of *perfect* negative correlation in a two-player game, we fully characterize the set of equilibria. We find that there always exists an equilibrium in cutoff strategies, whereas there never exists an equilibrium in cutoff strategies in the positively correlated benchmark case. Again in marked contrast to the case of perfect positive correlation, where the probability of never finding out the true state of the world was always inefficiently large, we find that in *any* equilibrium, this probability coincides with the efficient benchmark. Furthermore, using elementary constructive methods, we characterize a symmetric equilibrium in cutoff strategies for all parameter values for the case of *arbitrary* negative correlation. These equilibria again exhibit efficient long-run patterns of learning. We moreover extend our results to three players.

In Chapter 2, I analyze a game of two players operating replica bandits with one safe arm and two risky arms, the respective types of which are perfectly negatively correlated.[3] While the general setup may seem superficially similar to that in Chapter 1, there are notable differences, which are shown to be of quite some import to the analysis: Players now endogenously decide at each "point" in time if they want to investigate a given hypothesis or its negation. While previous literature has found that there never exists an efficient equilibrium, and Chapter 1 has shown that there exists an efficient equilibrium, if, and only if, the stakes at play are *below* a certain threshold, I show in Chapter 2 that if players endogenously *choose* which risky arm to pull, there exists an efficient equilibrium if, and only if, the stakes

[1] Throughout this dissertation, I use the term *two-armed bandits* to refer to bandits with one safe arm and one risky arm, which some authors are wont to refer to as *one-armed bandits*. Indeed, the problem is isomorphic to a stopping problem on a single risky arm.

[2] Chapter 1 is joint work with my dissertation advisor Sven Rady, and is based on a joint paper entitled "Negatively Correlated Bandits" (2010).

[3] Chapter 2 is based on my paper "Strategic Learning in Teams" (2010a).

exceed a certain threshold. While the technical issues arising from the negative correlation between the risky arms can, *modulo* minor details, be solved in a manner mirroring Chapter 1, the actual construction of equilibria requires quite different technical methods: The extension of players' action sets renders a full characterization of the equilibrium set elusive; the construction rather relies on certain linearities, which guarantee the existence of a best response in the boundary of a player's action space.

Whereas Chapters 1 and 2 take the structure of agents' rewards as given, Chapter 3 endogenizes them by analyzing a game between a single principal, who enjoys full commitment power, and a single agent.[4] The principal is interested in knowing if a certain hypothesis is true or false, yet he cannot conduct the relevant investigation himself, but has to delegate this task to an agent. In continuous time, the agent in turn can either be honest and actually investigate the hypothesis, or he can shirk and be lazy, or, unbeknownst to the principal, he can manipulate the data intimating he has proved the hypothesis. In Chapter 3, I identify the parameters for which it is possible for the principal to give the agent proper incentives to investigate the hypothesis, and I construct an optimal mechanism that does so.

In Chapter 4, the agent no longer operates a bandit machine, but can rather be conceived of as the strategically acting risky arm of a two-armed bandit.[5] Specifically, in each period, a decision maker who faces a sequence of decision problems chooses whether to seek the advice of an expert. The latter is not interested in the policy decision *per se*, but rather tries to maximize the number of periods that his advice is sought. The precision of the expert's private information is initially unknown, and is gradually learnt over time. Thus, the expert may have incentives strategically to bias the cheap-talk relay of his information, in order favorably to influence the decision maker's impression of him. Should he choose to do so, he will gradually accumulate private information about his type. We show that if exogenous employment costs are low, the expert will be hired with some probability even after he has revealed himself to be of the bad type. This construction heavily relies on the assumption that the decision maker has full commitment power. In the absence of commitment power and low employment costs, we show that any fully revealing equilibrium is dominated by an equilibrium that allows the expert to accumulate some private information early in the relationship (*probationary period*).

[4] Chapter 3 is based on my paper "The Importance of Being Honest" (2010b).

[5] Chapter 4 is joint work with Tymofiy Mylovanov of Pennsylvania State University, and is based on our paper "Expert Experimentation" (2010).

Chapter 1: Negatively Correlated Bandits[*]

Nicolas Klein[†] Sven Rady[‡]

Abstract

We analyze a two-player game of strategic experimentation with two-armed bandits. Either player has to decide in continuous time whether to use a safe arm with a known payoff or a risky arm whose expected payoff per unit of time is initially unknown. This payoff can be high or low, and is negatively correlated across players. We characterize the set of all Markov perfect equilibria in the benchmark case where the risky arms are known to be of opposite type, and construct equilibria in cutoff strategies for arbitrary negative correlation. All strategies and payoffs are in closed form. In marked contrast to the case where both risky arms are of the same type, there always exists an equilibrium in cutoff strategies, and there always exists an equilibrium exhibiting efficient long-run patterns of learning. These results extend to a three-player game with common knowledge that exactly one risky arm is of the high payoff type.

Keywords: Strategic Experimentation, Two-Armed Bandit, Exponential Distribution, Poisson Process, Bayesian Learning, Markov Perfect Equilibrium.

JEL Classification Numbers: C73, D83, O32.

[*] Our thanks for helpful comments and suggestions are owed to two anonymous referees, Philippe Aghion, Patrick Bolton, Kalyan Chatterjee, Martin Cripps, Matthias Dewatripont, Jan Eeckhout, Florian Englmaier, Eduardo Faingold, Chris Harris, Johannes Hörner, Philipp Kircher, George Mailath, Timofiy Mylovanov, Frank Riedel, Stephen Ryan, Klaus Schmidt, Larry Samuelson, and seminar participants at Århus, Bielefeld, Bonn, Mannheim, Munich, HEC Paris, UPenn, Yale, the 2007 SFB/TR 15 Summer School in Bronnbach, the 2007 SFB/TR 15 Workshop for Young Researchers in Bonn, the 2008 SFB/TR 15 Meeting in Gummersbach, the 2008 North American Summer Meetings of the Econometric Society, the European Summer Symposium in Economic Theory (ESSET) 2008, the 2008 Meeting of the Society for Economic Dynamics and the 2008 European Meeting of the Econometric Society. We thank the Studienzentrum Gerzensee and the Economics Departments at the University of Bonn and UPenn for their hospitality. We are both particularly grateful for an extended stay at the Cowles Foundation for Research in Economics at Yale University during which this version of the paper took shape. Financial support from the Deutsche Forschungsgemeinschaft through SFB/TR 15 and GRK 801 as well as from the National Research Fund of Luxembourg is gratefully acknowledged.

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

[‡] Department of Economics, University of Munich, Kaulbachstr. 45, D-80539 Munich, Germany; email: sven.rady@lrz.uni-muenchen.de.

1 Introduction

Starting with Rothschild (1974), two-armed bandit models have been used extensively in economics to formalize the trade-off between experimentation and exploitation in dynamic decision problems with learning; see Bergemann and Välimäki (2008) for a survey of this literature. The use of the two-armed bandit framework as a canonical model of strategic experimentation in teams is more recent: Bolton and Harris (1999, 2000) analyze the case of Brownian motion bandits, while Keller, Rady and Cripps (2005) and Keller and Rady (2010) analyze bandits where payoffs are governed by Poisson processes. These papers assume perfect *positive* correlation of the quality of the risky arm across players; all risky arms generate the same unknown expected payoff per unit of time, so what is good news to any given player is good news for everybody else.

There are many situations, however, where one man's boon is the other one's bane. Think of a suit at law, for instance: whatever is good news for one party tends to be bad news for the other. Or consider two firms pursuing research and development under different, incompatible working hypotheses. One pharmaceutical company, for example, might base its drug development strategy on the hypothesis that the cause of a particular disease is a virus, while the other might see a bacterium as the cause. An appropriate model of strategic experimentation in such a situation must assume *negative* correlation of the quality of the risky arm across players. This we propose to do in the present paper.

There are two players in our model, either one facing a continuous-time exponential bandit as in Keller, Rady and Cripps (2005). One arm is safe, generating a known payoff per unit of time. The other arm is risky, and can be good or bad. If it is good, it generates lump-sum payoffs after exponentially distributed random times; if it is bad, it never generates any payoff. A good risky arm dominates the safe one in terms of expected payoffs per unit of time, whereas the safe arm dominates a bad risky one. At the start of the game, the players hold a common belief about the types of the two risky arms. Either player's actions and payoffs are perfectly observable to the other player, so any information that a player garners via experimentation with the risky arm is a public good, and the players' posterior beliefs agree at all times.

We first analyze the case of *perfect* negative correlation, where it is common knowledge that exactly one risky arm is good. In a lawsuit, for example, this means that there exists conclusive evidence for one side which, once found, will decide the case in its favor; in the example of drug development, it means that one of the two mutually exclusive hypotheses will turn out to be true if explored long enough. The dynamics of posterior beliefs are easy to describe in this case. If both players play safe, no new information is generated and beliefs stay unchanged. If only one player plays risky and he has no success, the posterior probability that his risky arm is the good one falls gradually over time; if he obtains a lump-sum payoff, all uncertainty is resolved and beliefs become degenerate at the true state of the world. If both players play risky, finally, and there is no success on either arm, this is again uninformative about the state of the world, so beliefs are constant up to the random time when the first success occurs. It is important to note that a success on one player's risky arm is always bad news for the other player, while lack of success gradually makes the other player more optimistic.

We restrict players to stationary Markov strategies with the common posterior belief as the state variable. As is well known, this restriction is without loss of generality in the decision problem of a single agent experimenting in isolation: his optimal policy is given by a cutoff strategy, i.e. has him play risky at beliefs more optimistic than some threshold, and safe otherwise. The same structure prevails in the optimization problem of a utilitarian planner who maximizes the average of the two players' expected discounted payoffs. In the non-cooperative experimentation game, the Markov restriction rules out history-dependent behavior that is familiar from the analysis of infinitely repeated games in discrete time, yet technically quite difficult to formalize in continuous time (Simon and Stinchcombe 1989, Bergin 1992, Bergin and McLeod 1993). Imposing Markov perfection allows us to focus on the experimentation tradeoff that the players face and makes our results directly comparable to those in the previous literature on strategic experimentation in bandits. Moreover, a simple numerical evaluation of average payoffs suggests that Markov perfect equilibria are able to capture a surprisingly high fraction of the welfare gain that the planner's solution achieves relative to the safe payoff level.

The implementation of the Markov restriction needs some care in our setting because the incremental drift in beliefs can change direction as the action profile changes, which may lead to a differential equation for the state variable that possesses no, two, or a continuum of solutions. In contrast to Keller, Rady and Cripps (2005), where the drift always has the same sign, this problem cannot be remedied by the imposition of one-sided continuity requirements and arises even if both players use cutoff strategies. It is therefore impossible to define the set of a player's admissible strategies without reference to his opponent's strategy.[1] We confront this problem by calling a *pair* of strategies admissible if there exists at least one well-defined solution to the corresponding law of motion of our state variable. If there are several solutions, we select the one that can be obtained as the limit of a discrete-time approximation. We set both players' payoffs to minus infinity on any strategy profile that

[1] More generally, this problem arises whenever the types of the two risky arms are neither independent nor perfectly positively correlated. We will see this in the case of imperfect *negative* correlation below; cf. the proof of Proposition 11. For the case of imperfect *positive* correlation, see our concluding remarks.

is not admissible. A best response to the opponent's strategy thus necessarily leads to well-defined dynamics of beliefs and actions.

Before turning to the Markov perfect equilibria of the experimentation game with perfect negative correlation, we characterize efficient behavior by solving the planner's optimization problem. When stakes (as measured by the payoff advantage of a good risky arm over a safe one) are so low that there exist beliefs at which both players are below their single-agent optimal thresholds, it is optimal for the planner to let either player apply his respective single-agent threshold, so that they both behave as if they were experimenting on their own. In particular, the planner stops all learning once the belief is in the range where both players are below their single-agent cutoffs. This is efficient because experimentation on player 1's bandit, say, can never make the belief jump into a region where experimentation on player 2's bandit became profitable. When stakes are higher, there exist beliefs at which both players are above their single-agent cutoffs, and it is optimal for the planner to have both players simultaneously use the risky arm at some beliefs. In this case, learning is complete, meaning that posterior beliefs converge to the truth almost surely.

As our first main result, we show that there always exists an equilibrium where both players use a cutoff strategy. Suppose for example that player 2 follows a cutoff strategy and player 1's best response has him play risky at a given belief. Then player 1's learning benefit from doing so must outweigh the opportunity costs. At more optimistic beliefs, the opportunity costs of playing risky are even lower while the learning benefit is at least as high because the opponent provides either the same amount of free information or less. It must therefore be optimal for player 1 to play risky at more optimistic beliefs as well, and so he must be playing a cutoff strategy himself.

If player 1's optimal cutoff lies in the region where player 2 is playing safe, it must coincide with the single-agent cutoff because the tradeoff faced by player 1 is exactly the same as that faced by an agent experimenting in isolation. If player 1's optimal cutoff lies in the region where player 2 is playing risky, it must be the same as that applied by a myopic agent who is just interested in the maximization of *current* payoffs. Indeed, when player 1 joins player 2 in playing risky, he freezes beliefs and actions until the random time when the first breakthrough resolves all uncertainty, and his total expected payoff is linear in the probabilities that he assigns to the two possible states of the world. If player 1 were now offered the possibility of observing, for a short time interval and at no cost, the payoffs generated by a replica of his own risky arm, he would be indifferent to the offer because the resulting mean-preserving spread in beliefs would leave his expected continuation payoff unchanged. Player 1 thus assigns zero value to the information he gathers when playing risky, so his decision to use the risky arm must maximize current payoffs.[2]

As the myopic cutoff belief is more optimistic than the single-agent cutoff, we obtain three cases. When stakes are so low that there exist beliefs at which both players are below their single-agent cutoffs, the unique equilibrium in cutoff strategies is for both players to behave as if they were single agents. When stakes are so high that there exist beliefs at which both players are above their myopic cutoffs, the unique equilibrium in cutoff strategies is for both players to behave as if they were myopic. When stakes are intermediate in size, finally, there exist beliefs at which either player finds himself in between his single-agent and his myopic cutoff, and thus optimally plays risky if the opponent plays safe, and safe if the opponent plays risky. Each such belief can then serve as the common threshold in an MPE in cutoff strategies.

Our second contribution is a complete characterization of all Markov perfect equilibria of the two-player game with perfect negative correlation. For low and high stakes, respectively, the cutoff equilibrium just described is the unique MPE. We prove this by characterizing the changes in the players' action profile that may occur in equilibrium, and the beliefs at which they may occur. Given that the players have dominant actions near subjective certainty (the player who is very optimistic about his risky arm uses it, the other one plays safe), the proof reduces to showing that as we vary the belief from one extreme of the state space to the other, the respective cutoff equilibrium provides the only way for the players to transition from one profile of dominant actions to the other.

For intermediate stakes, there exist equilibria that are not in cutoff strategies. Over the range of beliefs where either player's best response is to play the opposite of his opponent's action, it is possible for them to swap roles finitely often. Using the same approach as for low and high stakes, we characterize the set of all equilibria and show that in every MPE that is not in cutoff strategies, the players' payoff functions necessarily have jump discontinuities. These arise at each belief where players swap roles in a way that implies locally divergent belief dynamics. Priors arbitrarily close to each other, but on different sides of such a belief lead to very different paths of beliefs and actions, and hence to payoffs that are bounded away from each other.

The third main result of the paper concerns the asymptotics of learning. In any MPE of the two-player game with perfect negative correlation, the probability of learning the true state in the long run is the same as in the planner's solution. For low stakes, there is nothing to show because the unique equilibrium coincides with the planner's solution. For

[2] Intuitively, players cannot assign a negative value to public information when, as in the present model, the only strategic link between them is a positive informational externality. They can do so when they also exert a payoff externality on each other; see for example Harrington (1995) or Keller and Rady (2003).

intermediate and high stakes, free-riding leads to an inefficiently small set of beliefs where both players use the risky arm, yet learning is nevertheless complete in equilibrium, exactly as the planner would have it. The intuition is straightforward. If players hold common beliefs and there is perfect negative correlation between the types of the risky arms, it can never be the case that both players are simultaneously very pessimistic about their respective prospects; with stakes sufficiently high, this implies that at least one player must be using the risky arm at any time, and so learning never stops. Thus, whenever society places a lot of emphasis on uncovering the truth, as one may argue is the case with medical research or the justice system, our analysis would suggest an adversarial setup was able to achieve this goal.[3]

The existence of equilibria in cutoff strategies, the uniqueness of equilibrium for low and high stakes, and the efficiency of long-run learning outcomes stand in stark contrast to the case of perfect positive correlation analyzed in Keller, Rady and Cripps (2005). First, there is *no* equilibrium in cutoff strategies when all risky arms are of the same type. It is easy to see where the intuition given above fails. If player 2 follows a cutoff strategy and player 1's best response has him play risky at a given belief, then the learning benefit at more optimistic beliefs can be *lower* because the opponent may provide *more* free information there. So free-riding on this information may be the better choice.[4] Second, the experimentation game with identical risky arms admits a continuum of equilibria irrespective of the size of the stakes involved. As the evolution of beliefs is determined by the total number of risky arms used at a given time, one and the same equilibrium pattern of information can in fact be generated via many different assignments of the roles of experimenter and free-rider, respectively. Moreover, there is multiplicity with respect to these equilibrium patterns, yielding a continuum of average payoff functions. Third, with experimentation stopping too early, any MPE entails an inefficiently high probability of incomplete learning.

When the quality of the risky arm is perfectly negatively correlated across players, one side's failure to produce evidence in its favor means that the other side is more likely to do so. However, in a lawsuit, for instance, there might not exist one single conclusive piece of evidence which settled the case once and for all; in the drug development example, the disease in question might be caused by a genetic defect rather than a virus or a bacterium. In a second step, therefore, we extend the model to *imperfect* negative correlation by introducing a third state of the world in which both risky arms are bad. When one side fails to produce evidence in its favor, the increase in the other side's individual optimism is now tempered

[3] Dewatripont and Tirole (1999) reach a similar conclusion in a moral hazard setting.

[4] More precisely, Keller, Rady and Cripps (2005) show that with two players, the player who is supposed to use the least optimistic cutoff in a purported MPE in cutoff strategies always has an incentive to deviate to the safe action at the other player's cutoff belief.

by an increase in collective pessimism, that is, an increase in the posterior probability that both sides will remain unsuccessful.

With three states of the world, beliefs are elements of a two-dimensional simplex, and the players' payoff functions solve linear partial differential equations. Given a fixed action profile, the trajectories of beliefs conditional on no breakthrough are straight lines in the simplex. Along each such line, we can represent the corresponding payoff function in closed form up to a constant of integration that varies with the slope of the line.

The fourth contribution of the paper is to show constructively that the game with imperfect negative correlation always admits an equilibrium in cutoff strategies, and to provide explicit representations for the players' strategies and payoff functions. As there is now a dimension of collective pessimism, the probability that learning remains incomplete in the long run is always positive. In the equilibria that we construct, this probability is the same as in the planner's solution.

These insights carry over to a game with three players and common knowledge that exactly one of them has a good risky arm. Again, there always exists an equilibrium in cutoff strategies, and the resulting asymptotics of learning are the same as in the planner's solution. Moreover, two of our findings for the two-player game with perfect negative correlation generalize to an arbitrary number of players: for sufficiently small stakes, players behave as if they were single agents experimenting in isolation, which is efficient; and learning will be complete in equilibrium if and only if efficiency requires complete learning.

The related literature on strategic experimentation with publicly observable actions and outcomes has already been addressed. Rosenberg, Solan and Vieille (2007) and Murto and Välimäki (2009) study strategic experimentation with two-armed bandits where the players' actions are publicly observable, but their payoffs are private information. These authors assume that the decision to stop playing risky is irreversible. In our model, players can freely switch back and forth between the two arms. Bonatti and Hörner (2010) study a model with private actions and publicly observable outcomes. Yet, theirs is more a model of moral hazard in teams than an experimentation model, implying, *inter alia*, that no player will ever play risky below his myopic cutoff.

There is a decision-theoretic literature on correlated bandits which analyzes correlation across different arms of a bandit operated by a single agent; see e.g. Camargo (2007) for a recent contribution to this literature, or Pastorino (2005) for economic applications. Our focus here is quite different, though, in that we are concerned with correlation between different bandits operated by two or more players who interact strategically.

Chatterjee and Evans (2004) analyze an R&D race with two firms and two projects in which it is common knowledge that exactly one of these projects will bear fruit if pursued long enough, and actions and payoffs are observable. Their discrete-time model differs from ours in several respects, chief of which is the payoff externality implied by the firms' choices. In our model, there is no payoff rivalry between players – strategic interaction arises out of purely informational concerns. Moreover, Chatterjee and Evans allow firms to change their projects at any time, so that it is possible for them to explore the same project. Our analysis, by contrast, presumes that projects of opposite type have been irrevocably assigned to players at the start of the experimentation game.[5] Finally, we allow for imperfect negative correlation between project types.

The rest of the paper is structured as follows. Section 2 introduces the game with two players and perfect negative correlation between the types of their risky arms. Section 3 solves the planner's problem. Section 4 characterizes the Markov perfect equilibria of the non-cooperative game, compares their learning outcomes and average payoffs to the planner's solution, and discusses robustness to the introduction of interior intensities of experimentation, asymmetries between the two players and news events that are not fully revealing. Section 5 constructs equilibria in the version of the game where the negative correlation between the types of the two players' risky arms is imperfect. Section 6 extends the model to three or more players. Section 7 concludes. Appendix A contains auxiliary results on payoff functions. Appendix B characterizes admissible strategy pairs in the game with perfect negative correlation. Most proofs are provided in Appendix C.

2 The Model

There are two players, 1 and 2, either one of whom faces a two-armed bandit problem in continuous time. Bandits are of the exponential type studied in Keller, Rady and Cripps (2005). One arm is safe in that it yields a known payoff flow of s; the other arm is risky in that it is either good or bad. If it is bad, it never yields any payoff; if it is good, it yields a lump-sum payoff with probability λdt when used over a length of time dt.[6] Let $g\,dt$ denote the corresponding expected payoff increment; thus, g is the product of the arrival rate λ and the average size of a lump-sum payoff. To have an interesting problem, we assume that the expected payoff of a good risky arm exceeds that of the safe arm, whereas the safe arm

[5] In the concluding remarks, we briefly report on an extension of our model in which players are given a sequential once-and-for-all choice of bandit prior to the experimentation game.

[6] The assumption of a common arrival rate of successes on a good risky arm is crucial to the analytic tractability of the model, while asymmetries in the other parameters are straightforward to accommodate; see the discussion in Section 4.6 below.

is better than a bad risky arm, i.e. $g > s > 0$. The time-invariant constants $\lambda > 0$ and $g > 0$ are common knowledge. Throughout Sections 2–4, we will further assume common knowledge that exactly one bandit's risky arm is good.

Player $i = 1, 2$ chooses actions $\{k_{i,t}\}_{t\geq 0}$ such that $k_{i,t} \in \{0, 1\}$ is measurable with respect to the information available at time t, with $k_{i,t} = 1$ indicating use of the risky arm, and $k_{i,t} = 0$ use of the safe arm. At the outset of the game, the players hold a common prior belief about which of the risky arms is good, given by the probabilities with which nature allocates the good risky arm to either player. Throughout the game, players perfectly observe each other's actions and payoffs, and so share a common posterior belief at all times. We write p_t for the players' probability assessment at time t that player 1's risky arm is good. Player 1's total expected discounted payoff, expressed in per-period units, can then be written as

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

where the expectation is taken over the stochastic processes $\{k_{1,t}\}$ and $\{p_t\}$, and $r > 0$ is the players' common discount rate. The corresponding payoff of player 2 is

$$\mathrm{E}\left[\int_0^\infty r\, e^{-r\,t}\ [k_{2,t}(1 - p_t)g + (1 - k_{2,t})s]\ dt\right].$$

The strategic link between the players stems from the impact of their actions on the evolution of beliefs.

The posterior belief jumps to 1 if there has been a breakthrough on player 1's bandit, and to 0 if there has been a breakthrough on player 2's bandit, where in either case it will remain ever after. If there has been no breakthrough on either bandit by time t given the players' actions $\{k_{1,\tau}\}_{0\leq\tau\leq t}$ and $\{k_{2,\tau}\}_{0\leq\tau\leq t}$, Bayes' rule yields

$$p_t = \frac{p_0 e^{-\lambda \int_0^t k_{1,\tau}\, d\tau}}{p_0 e^{-\lambda \int_0^t k_{1,\tau}\, d\tau} + (1 - p_0) e^{-\lambda \int_0^t k_{2,\tau}\, d\tau}}.$$

In particular, the posterior belief evolves continuously up to the time of the first breakthrough.

We restrict players to stationary Markov strategies with the common belief as the state variable and adopt the solution concept of Markov perfect equilibrium. As Markov strategies of player $i = 1, 2$, we allow all functions $k_i \colon [0, 1] \to \{0, 1\}$ such that both $k_i^{-1}(0)$ and $k_i^{-1}(1)$ are disjoint unions of a finite number of non-degenerate intervals, with $k_i(0) = i - 1$ and $k_i(1) = 2 - i$ (the dominant actions under subjective certainty). A Markov strategy k_1 for player 1 is called a *cutoff strategy* with cutoff $\hat{p}_1$ if $k_1^{-1}(1) = [\hat{p}_1, 1]$ or $]\hat{p}_1, 1]$. Analogously, a Markov strategy k_2 for player 2 is a cutoff strategy with cutoff $\hat{p}_2$ if $k_2^{-1}(1) = [0, \hat{p}_2]$ or $[0, \hat{p}_2[$. The action at the cutoff itself is deliberately left unspecified.[7]

A pair of Markov strategies (k_1, k_2) is called *symmetric* if $k_1(p) = k_2(1-p)$ at all p. The pair is called *admissible* if there exists at least one well-defined solution to the corresponding law of motion for posterior beliefs. This is the case if and only if for each initial belief p_0 in the unit interval, there is a function $t \mapsto p_t$ on $[0, \infty[$ that satisfies

$$p_t = \frac{p_0 e^{-\lambda \int_0^t k_1(p_\tau)\, d\tau}}{p_0 e^{-\lambda \int_0^t k_1(p_\tau)\, d\tau} + (1-p_0) e^{-\lambda \int_0^t k_2(p_\tau)\, d\tau}} \tag{1}$$

at all $t \geq 0$. This function then describes a possible time path of beliefs prior to the first breakthrough on a risky arm. If there are multiple solutions, we select the unique solution that is consistent with a discrete-time approximation; see Appendix B for details and a characterization of admissible strategy pairs.[8]

Each admissible strategy pair (k_1, k_2) induces a pair of payoff functions $u_1, u_2 \colon [0,1] \to [0, g]$ given by

$$\begin{aligned} u_1(p|k_1,k_2) &= \mathrm{E}\left[\int_0^\infty r e^{-rt} \left\{ k_1(p_t) p_t g + [1 - k_1(p_t)] s \right\} dt \,\middle|\, p_0 = p \right], \\ u_2(p|k_1,k_2) &= \mathrm{E}\left[\int_0^\infty r e^{-rt} \left\{ k_2(p_t)(1 - p_t) g + [1 - k_2(p_t)] s \right\} dt \,\middle|\, p_0 = p \right]. \end{aligned}$$

For strategy pairs that are not admissible, we set $u_1 \equiv u_2 \equiv -\infty$.

Strategy k_1 is a best response against strategy k_2 if the pair of strategies (k_1, k_2) is admissible and $u_1(p|k_1,k_2) \geq u_1(p|\tilde{k}_1,k_2)$ for all p in the unit interval and all admissible $(\tilde{k}_1, k_2)$. Analogously, strategy k_2 is a best response against strategy k_1 if (k_1, k_2) is admissible and $u_2(p|k_1,k_2) \geq u_1(p|k_1,\tilde{k}_2)$ for all p in the unit interval and all admissible $(k_1, \tilde{k}_2)$. A Markov perfect equilibrium (MPE) is a pair of strategies that are mutually best responses.

On any open interval of beliefs where an admissible pair of strategies (k_1, k_2) prescribes constant actions, the posterior belief solves the ordinary differential equation

$$\dot{p} = \lambda \left[k_2(p) - k_1(p) \right] p \,(1-p) \tag{2}$$

[7] We shall see later that there are circumstances where equilibrium requires the players to play safe at the cutoff belief, and others where equilibrium requires them to play risky.

[8] If we allowed for degenerate intervals in the interior of the unit interval, there would exist equilibria for low stakes in which one player would be forced (purely for reasons of admissibility of the strategy pair) to play risky at a belief where his resulting payoff is less than the safe payoff s. For high stakes, there would be equilibria in which an interval of beliefs where both players play risky (and achieve a payoff higher than s) is punctuated by finitely many beliefs at which both play safe. Details are available from the authors upon request. These equilibria cannot be obtained as limits of equilibria in discrete-time approximations of the continuous-time game, so we rule them out by insisting that either action must be played on a union of non-degenerate intervals.

as long as there is no breakthrough. Since the expected arrival rate of a breakthrough is $k_1(p)p\,\lambda$ on player 1's risky arm, and $k_2(p)(1-p)\lambda$ on player 2's, standard arguments imply that player 1's payoff function solves the ordinary differential equation

$$\begin{aligned} ru_1(p) \;&=\; r\left\{k_1(p)pg + [1-k_1(p)]s\right\} \\ &\quad + \lambda\left\{k_1(p)\,p\,[g-u_1(p)] + k_2(p)\,(1-p)\,[s-u_1(p)] + [k_2(p)-k_1(p)]\,p\,(1-p)\,u_1'(p)\right\} \end{aligned}$$

on any open interval where the players' actions do not change. After dividing both sides by r, we can write this ODE more succinctly as

$$u_1(p) = s + k_2(p)\beta_1(p,u_1) + k_1(p)[b_1(p,u_1) - c_1(p)],$$

where $c_1(p) = s - pg$ is the opportunity cost player 1 has to bear when he plays risky, $b_1(p,u_1) = \frac{\lambda}{r}p[g - u_1(p) - (1-p)u_1'(p)]$ is the learning benefit accruing to player 1 when he plays risky, and $\beta_1(p,u_1) = \frac{\lambda}{r}(1-p)[s - u_1(p) + pu_1'(p)]$ is his learning benefit from player 2's playing risky. The corresponding equation for player 2's payoff function is

$$u_2(p) = s + k_1(p)\beta_2(p,u_2) + k_2(p)[b_2(p,u_2) - c_2(p)],$$

where $c_2(p) = s - (1-p)g$ is the opportunity cost player 2 has to bear when he plays risky, $b_2(p,u_2) = \frac{\lambda}{r}(1-p)[g - u_2(p) + pu_2'(p)]$ is the learning benefit accruing to player 2 when he plays risky, and $\beta_2(p,u_2) = \frac{\lambda}{r}p[s - u_2(p) - (1-p)u_2'(p)]$ is his learning benefit from player 1's playing risky. It is straightforward to obtain closed-form solutions for these differential equations; see Appendix A for details.

Given a Markov strategy k_j of player j, standard arguments imply that on any open interval where player j's action is constant, player i's payoff function from playing a best response is once continuously differentiable[9] and solves the Bellman equation

$$u_i(p) = s + k_j(p)\beta_i(p,u_i) + \max_{k_i\in\{0,1\}} k_i[b_i(p,u_i) - c_i(p)].$$

Conversely, a standard verification argument yields the following sufficiency result. Given the Markov strategy k_j, consider the set $S(k_j)$ of all Markov strategies of player i that form an admissible strategy pair with k_j. For any belief p, let $K_i(p,k_j) = \{k_i(p)\colon k_i \in S(k_j)\}$; this is the set of all actions player i can choose at the belief p under the constraint that his Markov strategy be admissible together with k_j. At all those beliefs where player j's action does not change, $K_i(p,k_j) = \{0,1\}$. At a belief where player j's action does change, by contrast, $K_i(p,k_j)$ may be a singleton, in which case player i's action is already pinned

[9] At a belief where the opponent's action changes while the best response does not, the payoff function from this best response typically has a kink. At a belief where both the opponent's action and the best response change, the payoff function may possess a jump discontinuity; see Proposition 7 below.

down by admissibility.[10] Now, strategy $k_i \in S(k_j)$ is a best response if the resulting payoff function u_i satisfies the modified Bellman equation

$$u_i(p) = s + k_j(p)\beta_i(p, u_i) + \max_{k_i \in K_i(p,\, k_j)} k_i[b_i(p, u_i) - c_i(p)]$$

everywhere on the unit interval. It is understood here that whenever the players' actions differ, the right-hand side is evaluated at the one-sided derivative in the direction of the infinitesimal changes in beliefs implied by the respective strategy pair. When the players' actions coincide, the terms involving derivatives cancel.

If players were myopic, i.e. merely maximizing current payoffs, player 1 would use the cutoff $p^m = \frac{s}{g}$ and player 2 the cutoff $1-p^m$. If they were forward-looking but experimenting in isolation, player 1 would optimally use the single-agent cutoff computed in Keller, Rady and Cripps (2005), $p^* = \frac{rs}{(r+\lambda)g-\lambda s} < p^m$, and player 2 the cutoff $1 - p^*$.

We will find it useful below to distinguish three cases depending on the size of the stakes involved, i.e. on the value of information as measured by the ratio $\frac{g}{s}$, and on the parameters λ and r that govern the speed of resolution of uncertainty and the player's impatience, respectively. We speak of *low stakes* if $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$, *intermediate stakes* if $\frac{2r+\lambda}{r+\lambda} < \frac{g}{s} < 2$, and *high stakes* if $\frac{g}{s} > 2$. These cases are easily distinguished by the positions of the cutoffs p^m and p^*: stakes are low if and only if $p^* > \frac{1}{2}$; intermediate if and only if $p^* < \frac{1}{2} < p^m$; and high if and only if $p^m < \frac{1}{2}$. The boundary cases $\frac{g}{s} = \frac{2r+\lambda}{r+\lambda}$ and $\frac{g}{s} = 2$ will be treated separately when needed.

3 The Planner's Problem

In this section, we examine a utilitarian social planner's behavior in our setup. The Bellman equation for the maximization of the average payoff from the two bandits is

$$u(p) = s + \max_{(k_1,k_2)\in\{0,1\}^2} \left\{ k_1 \left[B_1(p,u) - \frac{c_1(p)}{2} \right] + k_2 \left[B_2(p,u) - \frac{c_2(p)}{2} \right] \right\},$$

where $B_1(p, u) = \frac{\lambda}{r} p[\frac{g+s}{2} - u(p) - (1 - p)u'(p)]$ measures the expected learning benefit of playing risky arm 1, and $B_2(p, u) = \frac{\lambda}{r}(1-p)[\frac{g+s}{2} - u(p) + pu'(p)]$ the expected learning benefit of playing risky arm 2. The planner's problem is clearly symmetric with respect to $p = \frac{1}{2}$. By standard arguments, the corresponding value function is convex; by symmetry, it admits its global minimum at $p = \frac{1}{2}$.

[10] We will first encounter this phenomenon when determining best responses to cutoff strategies in Proposition 3 below.

If it is optimal to set $k_1 = k_2 = 0$, the value function works out as $u(p) = s$. If it is optimal to set $k_1 = k_2 = 1$, the Bellman equation reduces to $u(p) = \frac{\lambda}{r}\left[\frac{g+s}{2} - u(p)\right] + \frac{g}{2}$, and so $u(p) = u_{11} = \frac{g}{2} + \frac{\lambda}{r+\lambda}\,\frac{s}{2}$. As one risky arm is good for sure, playing both of them is certain to generate an expected average payoff of $\frac{g}{2}$. At some random time τ, the first success on the good risky arm causes the planner to switch to the safe arm on the other bandit; his expected total payoff from that bandit is therefore $\frac{s}{2}$ times the expectation of $e^{-r\tau}$. As τ is exponentially distributed with rate parameter λ, this expectation is $\frac{\lambda}{r+\lambda}$. In the remaining cases where it is optimal to set $k_1 = 0$ and $k_2 = 1$, or $k_1 = 1$ and $k_2 = 0$, explicit solutions of the Bellman equation are obtained as the average of the individual payoff functions stated in Appendix A.

It is clear that $(k_1, k_2) = (1, 0)$ will be optimal in a neighborhood of $p = 1$, and $(k_1, k_2) = (0, 1)$ in a neighborhood of $p = 0$. What is optimal at beliefs around $p = \frac{1}{2}$ depends on which of the two possible plateaus s and u_{11} is higher. This in turn depends on the size of the stakes involved. In fact, $s > u_{11}$ if and only if stakes are low, i.e. $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$. This is the case we consider first.

Proposition 1 (Planner's solution for low stakes) *If $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$, and hence $p^* > \frac{1}{2}$, the planner's optimum is to apply the single-agent cutoffs p^* and $1 - p^*$, respectively, that is, to set $(k_1, k_2) = (0, 1)$ on $[0, 1 - p^*[$, $k_1 = k_2 = 0$ on $[1 - p^*, p^*]$, and $(k_1, k_2) = (1, 0)$ on $]p^*, 1]$. This solution remains optimal in the limiting case where $\frac{g}{s} = \frac{2r+\lambda}{r+\lambda}$ and $p^* = \frac{1}{2}$.*

Proof: See Appendix C. ∎

Thus, when the value of information, as measured by $\frac{g}{s}$, is so low that the single-agent cutoff p^* exceeds $\frac{1}{2}$, it is optimal for the planner to let the players behave as though they were solving two separate, completely unconnected, problems.[11] The left panel of Figure 1 illustrates the corresponding value function.

Next, we turn to the case where $u_{11} > s$, which is obtained for intermediate and high stakes.

Proposition 2 (Planner's solution for intermediate and high stakes) *If $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$, and hence $p^* < \frac{1}{2}$, the planner's optimum is to apply the cutoffs $\bar{p} = \frac{(r+\lambda)s}{(r+\lambda)g+\lambda s} \in \,]p^*, \frac{1}{2}[$ and $1 - \bar{p}$, respectively, that is, to set $(k_1, k_2) = (0, 1)$ on $[0, \bar{p}[$, $k_1 = k_2 = 1$ on $[\bar{p}, 1 - \bar{p}]$, and*

[11] This would be different if playing the risky arm could also lead to "bad news events" that triggered downward jumps in beliefs. If, starting from p^*, such a jump were large enough to take the belief below $1 - p^*$, then letting player 1 play risky at beliefs somewhat below p^* would raise average payoffs.

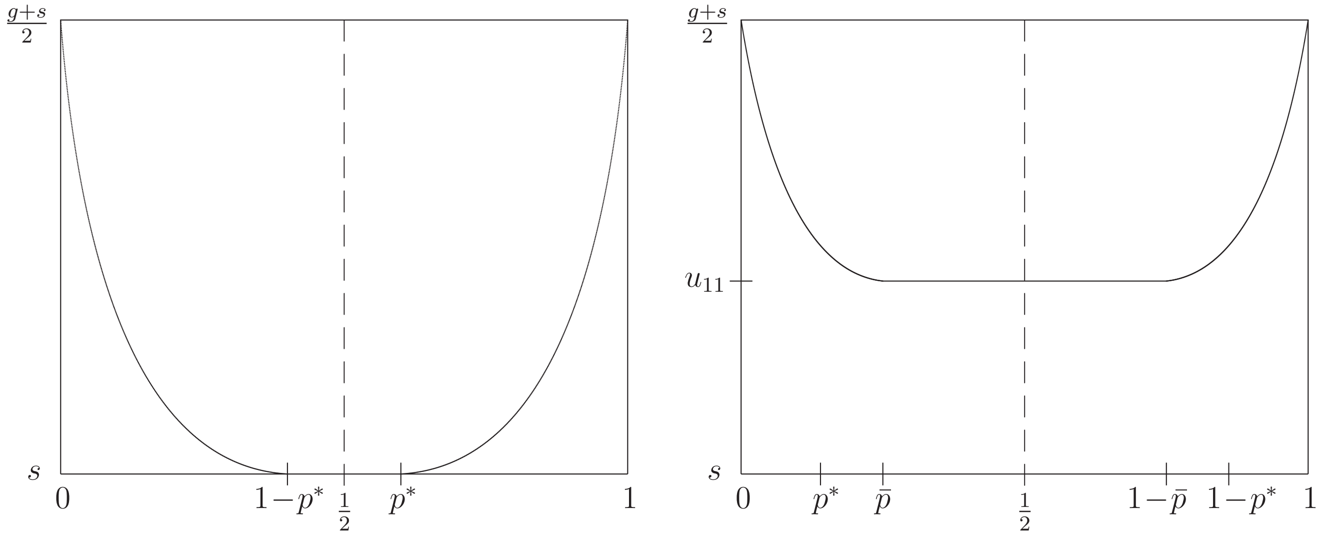

Figure 1: The planner's value function for $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$ (left panel) and $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$ (right panel).

$(k_1, k_2) = (1, 0)$ *on* $]1-\bar{p}, 1]$. *This solution, with* $\bar{p} = p^* = \frac{1}{2}$, *remains optimal in the limiting case where* $\frac{g}{s} = \frac{2r+\lambda}{r+\lambda}$.

PROOF: It is straightforward to check that $p^* \leq \bar{p} \leq \frac{1}{2}$ if $\frac{g}{s} \geq \frac{2r+\lambda}{r+\lambda}$. The rest of the proof proceeds along the same lines as that of Proposition 1 and is therefore omitted. ∎

The right panel of Figure 1 illustrates this result. To understand why the planner has either player use the risky arm on a smaller interval of beliefs than in the respective single-agent optimum, consider the effect of player 1's action on the aggregate payoff when player 2 is playing risky. If the planner is indifferent between player 1's actions at the belief $\bar{p}$, it must be the case that $\frac{\lambda}{r}\bar{p}[g + s - 2u_{11}] = c_1(\bar{p})$, with the possibility of a jump in the sum of the two players' payoffs from $2u_{11}$ to $g + s$ exactly compensating for the opportunity cost of player 1 using the risky arm. For a player 1 experimenting in isolation, the corresponding equation reads $\frac{\lambda}{r}p^*[g - s] = c_1(p^*)$. When $u_{11} > s$, the jump from s to g is larger than the one from $2u_{11}$ to $g + s$, so we cannot have $\bar{p} = p^*$. That $\bar{p}$ must be greater than p^* follows from the fact that the opportunity cost of using player 1's risky arm is decreasing in p.

4 Markov Perfect Equilibria

Our next aim is to characterize the Markov perfect equilibria of the experimentation game. To start out, we shall establish that the best response to certain cutoff strategies is in turn a cutoff strategy.

To get a first intuition for the results to come, suppose player 2 follows a cutoff strategy and player 1 plays a best response. If this response involves playing risky at some belief p, then the expected benefit of player 1's experimentation must outweigh its opportunity cost at p. At a belief $p' > p$, the opportunity cost is lower than at p and, since player 2 does not provide more free information to player 1 at p' than he does at p, the expected benefit of player 1's own experimentation should be at least as high as at p. So player 1 should also play risky at the belief p'. Thus, $k_1^{-1}(1)$ should be an interval with right boundary 1, implying a cutoff strategy for player 1.

The following proposition confirms this intuition and characterizes best-response cutoffs.

Proposition 3 (Best responses to cutoff strategies) *For player 1, a best response to* $k_2^{-1}(1) = [0, \hat{p}_2[$ *with* $\hat{p}_2 \leq p^*$ *is* $k_1^{-1}(1) =]p^*, 1]$*; to* $k_2^{-1}(1) = [0, \hat{p}_2]$ *with* $\hat{p}_2 \geq p^m$*, it is* $k_1^{-1}(1) = [p^m, 1]$*; and to* $k_2^{-1}(1) = [0, \hat{p}_2]$ *with* $p^* \leq \hat{p}_2 < p^m$*, it is* $k_1^{-1}(1) = [\hat{p}_2, 1]$*.*

For player 2, a best response to $k_1^{-1}(1) =]\hat{p}_1, 1]$ *with* $\hat{p}_1 \geq 1 - p^*$ *is* $k_2^{-1}(1) = [0, 1 - p^*[$*; to* $k_1^{-1}(1) = [\hat{p}_1, 1]$ *with* $\hat{p}_1 \leq 1 - p^m$*, it is* $k_2^{-1}(1) = [0, 1 - p^m]$*; and to* $k_1^{-1}(1) = [\hat{p}_1, 1]$ *with* $1 - p^m < \hat{p}_1 \leq 1 - p^*$*, it is* $k_2^{-1}(1) = [0, \hat{p}_1]$*.*

Proof: See Appendix C. ▮

While it is intuitive that player 1 should apply the single-agent cutoff p^* against an opponent who plays safe, and thus provides no information, at beliefs $p \geq p^*$, it is surprising that the myopic cutoff p^m determines player 1's best response against an opponent who plays risky. Technically, this result is due to the fact that along player 1's payoff function for $k_1 = k_2 = 1$, $u_1(p) = pg + (1-p)\frac{\lambda}{r+\lambda}s$, his learning benefit from playing risky vanishes:

$$b_1(p, u_1) = \frac{\lambda}{r} p \left[g - \left(pg + (1-p)\frac{\lambda}{r+\lambda}s \right) - (1-p)\left(g - \frac{\lambda}{r+\lambda}s \right) \right] = 0,$$

and so $k_1 = 1$ is optimal against $k_2 = 1$ if and only if $c_1(p) \leq 0$, that is, $p \geq p^m$.

Intuitively, this is best understood by recalling the law of motion of beliefs in the absence of a success on either arm, $\dot{p} = -(k_1 - k_2)\lambda p(1-p)$, which tells us that for $k_1 = k_2 = 1$, the state variable, and hence the players' actions, will not budge until the first success occurs and all uncertainty is resolved. Conditional on having the good risky arm, player 1 can thus look forward to a total expected discounted payoff equal to g. Conditional on having the bad risky arm, his total payoff equals s times the expectation of $e^{-r\tau}$ where τ is the exponentially distributed random time at which player 2 experiences his first success, causing player 1 to switch to the safe arm irrevocably. Weighting each state with its subjective probability, we obtain a payoff function that is linear in p. This means that player 1 is risk neutral with respect to lotteries over beliefs, so if he were offered the possibility of observing, for a short time interval and at no cost, the payoffs generated by a replica of his own risky arm, he would be indifferent to the offer, assigning zero value to this information because the resulting mean-preserving spread in beliefs would leave his continuation payoff unchanged on average. But if the value of information is zero, the decision to use the risky arm must be myopically optimal.

This insight also explains the third part of the proposition. If player 2 uses a cutoff $\hat{p_2}$ in between player 1's single-agent cutoff p^* and myopic cutoff p^m, player 1 does not want to play risky to the left of $\hat{p_2}$ because doing so is not myopically optimal there. Just to the right of $\hat{p_2}$, by contrast, he faces an opponent playing safe, so he views the situation exactly as a single agent experimenting in isolation would, and plays risky accordingly. Thus, player 1 uses the same cutoff as player 2. At $\hat{p_2}$ itself, player 1's behavior is pinned down by the requirement that his action be part of an admissible strategy pair. If he played safe at $\hat{p_2}$, the incremental drift of the state variable p would be positive for $p \leq \hat{p_2}$, and negative for $p > \hat{p_2}$. As we show in Appendix B, there would then be no solution to the law of motion of beliefs starting from the prior $p_0 = \hat{p_2}$. So player 1 can only use the risky arm at $\hat{p_2}$, and this action is indeed compatible with admissibility.

Using Proposition 3, it is straightforward to draw best-response correspondences in the space of cutoff pairs $(\hat{p}_1, \hat{p}_2)$ and characterize the resulting MPE in cutoff strategies. The nature of these equilibria depends on the relative position of the cutoffs p^*, p^m, $1 - p^*$ and $1 - p^m$, which, as previously noted, gives us a distinction between low, intermediate, and high stakes. We defer details to Propositions 4–6 below, each of which covers one of these three cases. For the moment, we just take note of the following stark contrast to Keller, Rady, Cripps (2005).

Corollary 1 (Equilibria in cutoff strategies) *For any combination of the parameters g, s, r, and λ, there exists an equilibrium in cutoff strategies.*

When investigating whether there exist Markov perfect equilibria beyond those in cutoff strategies, we shall make use of combinatoric arguments, exploiting the fact that for any admissible pair of Markov strategies, there can be but finitely many beliefs at which a change in action profile occurs. Appendix B characterizes the types and possible loci of these changes, allowing us to determine all manners in which equilibrium play can transition from the action profile $(0, 1)$ at $p = 0$ to the profile $(1, 0)$ at $p = 1$.

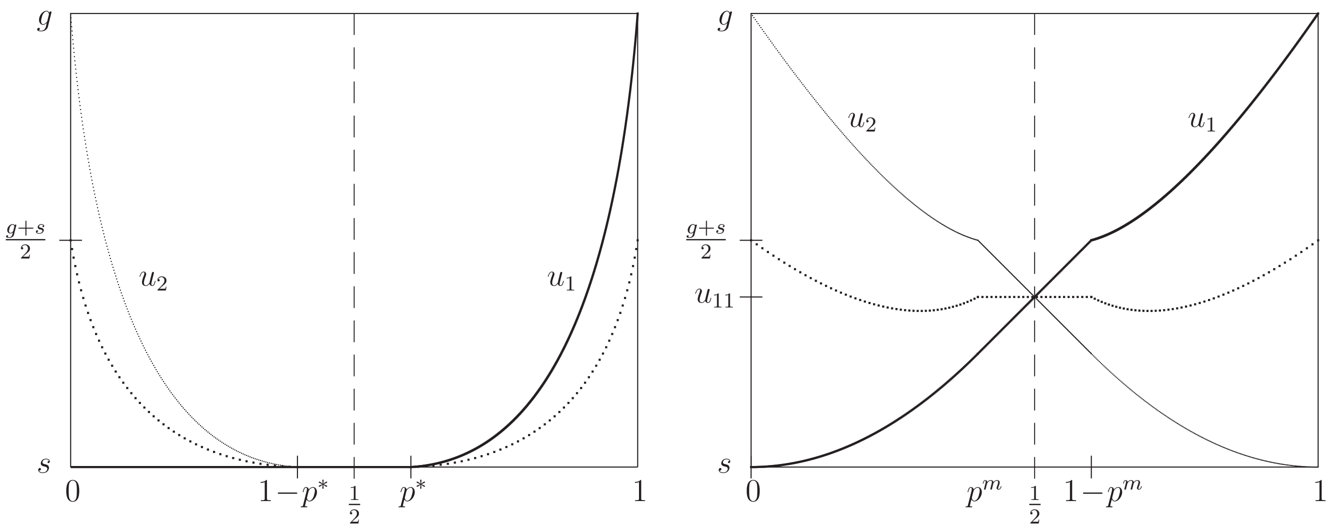

Figure 2: The equilibrium payoff functions for $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$ (left panel) and $\frac{g}{s} > 2$ (right panel). The thick solid curve depicts the payoff function of player 1, the thin solid curve that of player 2, and the dotted curve the players' average payoff function.

4.1 Low Stakes

Recall that the low-stakes case is defined by the inequality $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$. In this case, $1 - p^m < 1 - p^* < \frac{1}{2} < p^* < p^m$.

Proposition 4 (Markov perfect equilibrium for low stakes) *When $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$, the unique Markov perfect equilibrium is symmetric and coincides with the planner's solution. That is, player 1 plays risky if and only if $p > p^*$, and player 2 if and only $p < 1 - p^*$. These strategies continue to be an equilibrium in the limiting case where $\frac{g}{s} = \frac{2r+\lambda}{r+\lambda}$ and $p^* = \frac{1}{2}$.*

Proof: For $1 - p^* \leq p^*$, the cutoff strategies $k_1^{-1}(1) =]p^*, 1]$ and $k_2^{-1}(1) = [0, 1 - p^*[$ are mutually best responses by Proposition 3. For $1 - p^* < p^*$, uniqueness is proved in Appendix C. ▮

Why we should have efficiency in this case is intuitively quite clear, as the planner lets players behave as though they were single players. As $p^* > \frac{1}{2}$, there is no spillover from a player behaving like a single agent on the other player's optimization problem. Hence the latter's best response calls for behaving like a single player as well. Thus, there is no conflict between social and private incentives. The left panel of Figure 2 illustrates this result.

4.2 High Stakes

The high-stakes case is defined by the inequality $\frac{g}{s} > 2$. In this case, $p^* < p^m < \frac{1}{2} < 1 - p^m < 1 - p^*$.

Proposition 5 (Markov perfect equilibrium for high stakes) *When $\frac{g}{s} > 2$, the game has a unique Markov perfect equilibrium, which is symmetric and has both players behave myopically. That is, player 1 plays risky if and only if $p \geq p^m$, and player 2 if and only if $p \leq 1 - p^m$. These strategies also constitute the unique Markov perfect equilibrium in the limiting case where $\frac{g}{s} = 2$ and $p^m = \frac{1}{2}$.*

Proof: For $p^m \leq 1 - p^m$, the cutoff strategies $k_1^{-1}(1) = [p^m, 1]$ and $k_2^{-1}(1) = [0, 1 - p^m]$ are mutually best responses by Proposition 3. Uniqueness is proved in Appendix C. ∎

When the stakes are high, the unique equilibrium calls for both players' behaving myopically. This is best understood by recalling from our discussion above that individual optimality calls for myopic behavior whenever one's opponent is playing risky. When the stakes are high, players' myopic cutoff beliefs are more pessimistic than $p = \frac{1}{2}$, so the relevant intervals overlap.

The right panel of Figure 2 illustrates this result. Player 1's payoff function has a kink at $1 - p^m$, where player 2 changes action. Symmetrically, player 2's payoff function has a kink at p^m, where player 1 changes action. As a consequence, the average payoff function has a kink both at p^m and at $1 - p^m$. That it dips below the level u_{11} close to these kinks is evidence of the inefficiency of equilibrium.

4.3 Intermediate Stakes

This case is defined by the condition that $\frac{2r+\lambda}{r+\lambda} < \frac{g}{s} < 2$, or $p^* < \frac{1}{2} < p^m$. Equilibrium is not unique in this case; to start with, there is a continuum of equilibria in cutoff strategies, as the following proposition shows.

Proposition 6 (Intermediate stakes, equilibria in cutoff strategies) *For $\frac{2r+\lambda}{r+\lambda} < \frac{g}{s} < 2$, there is a continuum of Markov perfect equilibria in cutoff strategies, each characterized by a belief $\hat{p} \in [\max\{1 - p^m, p^*\}, \min\{p^m, 1 - p^*\}]$ such that player 1 plays risky if and only if $p \geq \hat{p}$, and player 2 if and only if $p \leq \hat{p}$. These strategies, with $\hat{p} = p^* = \frac{1}{2}$, continue to be an equilibrium in the limiting case where $\frac{g}{s} = \frac{2r+\lambda}{r+\lambda}$.*

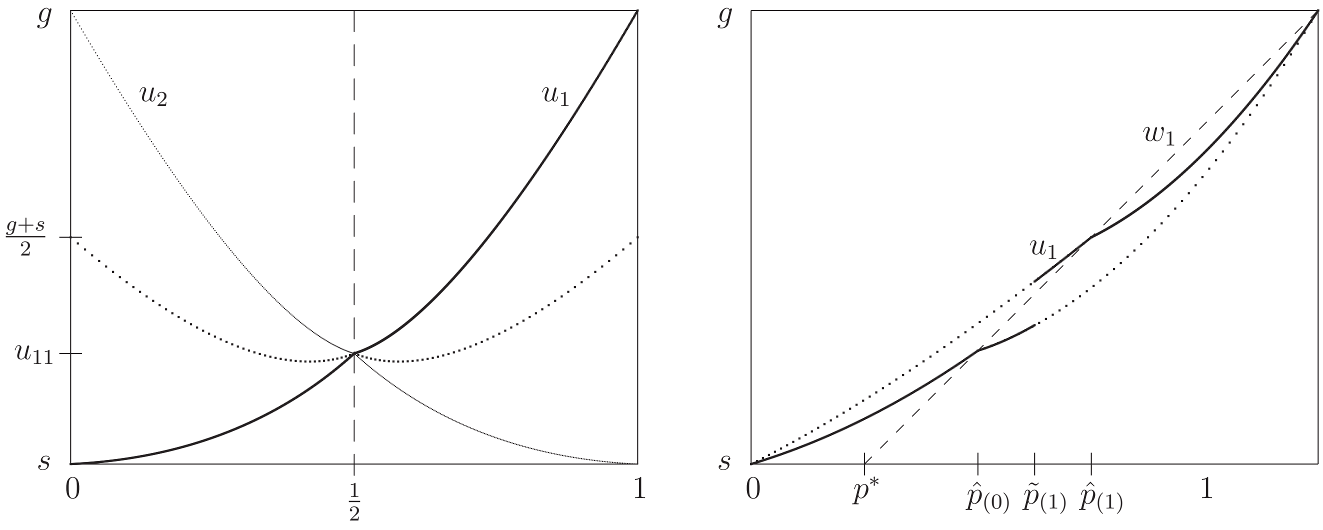

Figure 3: Equilibrium payoff functions for $\frac{2r+\lambda}{r+\lambda} < \frac{g}{s} < 2$. The left panel shows the players' payoff functions and their average in the unique symmetric equilibrium in cutoff strategies. The right panel shows the payoff function of player 1 in an equilibrium that is not in cutoff strategies (see the main text for further details).

PROOF: For $\max\{1 - p^m, p^*\} < \hat{p} < \min\{p^m, 1 - p^*\}$, the cutoff strategies $k_1^{-1}(1) = [\hat{p}, 1]$ and $k_2^{-1}(1) = [0, \hat{p}]$ are mutually best responses by Proposition 3. ∎

Amongst the continuum of equilibria characterized in Proposition 6, there is a unique symmetric one, given by $\hat{p} = \frac{1}{2}$. The left panel of Figure 3 illustrates this equilibrium. Both players' payoff functions and their average are kinked at $p = \frac{1}{2}$, where both players change action. At any belief except $p = \frac{1}{2}$, the average payoff function is below the planner's solution; if the initial belief is $p_0 = \frac{1}{2}$, however, the efficient average payoff u_{11} is achieved.

For the boundary case where $\frac{g}{s} = \frac{2r+\lambda}{r+\lambda}$ and $p^* = \frac{1}{2}$, Propositions 4 and 6 imply that both versions of the planner's solution are Markov perfect equilibria. Applying the arguments underlying the proof of Proposition 7 below, one easily shows that there are no other equilibria in this particular case.

All the equilibria exhibited so far share three features: they are in cutoff strategies; conditional on no breakthrough, posterior beliefs converge to a limit that varies continuously with the initial belief (we will return to this point in Section 4.4 below); and the players' payoff functions are continuous. For intermediate stakes, there exist further equilibria that are not in cutoff strategies. In these, the limit to which beliefs converge in the absence of a breakthrough depends discontinuously on the initial belief, and the players' payoff functions possess jump discontinuities. In combination with Proposition 6, the following result fully characterizes the set of Markov perfect equilibria for intermediate stakes.

Proposition 7 (Intermediate stakes, other equilibria) *Let* $\frac{2r+\lambda}{r+\lambda} < \frac{g}{s} < 2$*, and consider a pair of Markov strategies that are not cutoff strategies. These strategies constitute an equilibrium if and only if there exists an integer* $L \geq 1$ *and beliefs* $\hat{p}_{(0)} < \tilde{p}_{(1)} < \hat{p}_{(1)} < \ldots < \hat{p}_{(L-1)} < \tilde{p}_{(L)} < \hat{p}_{(L)}$ *in the interval* $[\max\{1-p^m, p^*\}, \min\{p^m, 1-p^*\}]$ *such that: on* $[0, \hat{p}_{(0)}[$ *and all intervals* $]\tilde{p}_{(\ell)}, \hat{p}_{(\ell)}[$*, the action profile is* $(0,1)$*; on all intervals* $]\hat{p}_{(\ell-1)}, \tilde{p}_{(\ell)}[$ *and* $]\hat{p}_{(L)}, 1]$*, the action profile is* $(1,0)$*; at all beliefs* $\hat{p}_{(\ell)}$*, the action profile is* $(1,1)$*; and at any belief* $\tilde{p}_{(\ell)}$*, the action profile is* $(0,1)$ *or* $(1,0)$*. Both players' payoff functions have jump discontinuities at all beliefs* $\tilde{p}_{(\ell)}$*.*

Proof: On the interval $[0, \tilde{p}_{(1)}[$, the players' actions and payoffs are the same as in an equilibrium in cutoff strategies with $\hat{p} = \hat{p}_{(0)}$. The same is true for each of the intervals $]\tilde{p}_{(\ell)}, \tilde{p}_{(\ell+1)}[$ (with $\hat{p} = \hat{p}_{(\ell)}$) and $]\tilde{p}_{(L)}, 1]$ (with $\hat{p} = \hat{p}_{(L)}$). So one only has to verify the mutual best-response property at the beliefs $\tilde{p}_{(\ell)}$. This is done in Appendix C. We also show there that payoffs are discontinuous at these beliefs, and that there are no other equilibria. ▮

The right panel of Figure 3 illustrates player 1's payoff function in an equilibrium with $L = 1$. The solid curve is u_1, and the dashed line the payoff w_1 that player 1 would get if both players played risky. The dotted curve starting in the lower left corner is the payoff player 1 would receive in a cutoff equilibrium with $\hat{p} = \hat{p}_{(0)}$, and the dotted curve starting in the upper right corner the payoff he would obtain in a cutoff equilibrium with $\hat{p} = \hat{p}_{(1)}$. Between $\hat{p}_{(0)}$ and $\tilde{p}_{(1)}$, beliefs drift downwards as only player 1 plays risky, and they will converge to $\hat{p}_{(0)}$ in finite time unless there is a breakthrough on player 1's risky arm. Between $\tilde{p}_{(1)}$ and $\hat{p}_{(1)}$, beliefs drift upwards as only player 2 plays risky, and they will converge to $\hat{p}_{(1)}$ in finite time unless there is a breakthrough on player 2's risky arm. Initial beliefs $\tilde{p}_{(1)} - \epsilon$ and $\tilde{p}_{(1)} + \epsilon$ thus imply very different paths of beliefs and actions. As a consequence, payoffs are discontinuous at $\tilde{p}_{(1)}$.

4.4 Asymptotics and Speed of Learning

When stakes are low and players use their single-agent cutoff strategies, the evolution of the posterior belief in the absence of a success on a risky arm is governed by

$$\dot{p} = \begin{cases} \lambda p(1-p) & \text{if } p < 1-p^*, \\ 0 & \text{if } 1-p^* \leq p \leq p^*, \\ -\lambda p(1-p) & \text{if } p > p^*. \end{cases} \tag{3}$$

The asymptotics of the learning process depend both on the true state of the world and the initial belief. Let us suppose, for example, that risky arm 1 is good. If the initial belief p_0 is lower than $1 - p^*$, the posterior belief will converge to $1 - p^*$ with probability 1 as there cannot be a breakthrough on risky arm 2. If $1 - p^* \leq p_0 \leq p^*$, the belief will remain unchanged at p_0 forever. If $p_0 > p^*$, the belief will converge either to 1 or to p^*. If t^* is the length of time needed for the belief to reach p^* conditional on there not being a breakthrough on risky arm 1, the probability that the belief will converge to p^* is $e^{-\lambda t^*}$. By Bayes' rule, we have $\frac{1-p_t}{p_t} = \frac{1-p_0}{p_0 e^{-\lambda t}}$ in the absence of a breakthrough, and so $e^{-\lambda t^*} = \frac{1-p_0}{p_0} \frac{p^*}{1-p^*}$. The belief will therefore converge to p^* with probability $\frac{1-p_0}{p_0} \frac{p^*}{1-p^*}$, and to 1 with the counter-probability. Analogous results hold when risky arm 2 is good. For low stakes, therefore, the unique (and efficient) MPE always entails a positive probability for learning to remain incomplete in the long run, that is, for the process of posterior beliefs to converge to a limit that assigns a positive probability to the false state of the world.

When stakes are high, the equilibrium dynamics of beliefs conditional on there not being a breakthrough are given by

$$\dot{p} = \begin{cases} \lambda p(1-p) & \text{if} \quad p < p^m, \\ 0 & \text{if} \quad p^m \leq p \leq 1 - p^m, \\ -\lambda p(1-p) & \text{if} \quad p > 1 - p^m. \end{cases} \tag{4}$$

Players shut down *incremental* learning on the interval $[p^m, 1 - p^m]$. Yet they still learn the true state with probability 1 in the long run because once this interval is reached, both players use their risky arm until the first success resolves all uncertainty.

When stakes are intermediate and the equilibrium is in cutoff strategies with common cutoff $\hat{p}$, the dynamics are

$$\dot{p} = \begin{cases} \lambda p(1-p) & \text{if} \quad p \in< \hat{p}, \\ 0 & \text{if} \quad p = \hat{p}, \\ -\lambda p(1-p) & \text{if} \quad p > \hat{p}, \end{cases} \tag{5}$$

and players learn the true state with probability 1 in the long run because both play risky at $\hat{p}$. When the equilibrium is not in cutoff strategies, $\dot{p} > 0$ on $[0, \hat{p}_{(0)}[$ and all intervals $]\tilde{p}_{(\ell)}, \hat{p}_{(\ell)}[$ and possibly at some of the beliefs $\tilde{p}_{(\ell)}$. Similarly, $\dot{p} < 0$ on all intervals $]\hat{p}_{(\ell-1)}, \tilde{p}_{(\ell)}[$ and $]\hat{p}_{(L)}, 1]$ and at the remaining beliefs $\tilde{p}_{(\ell)}$. Finally, $\dot{p} = 0$ at all beliefs $\hat{p}_{(\ell)}$. Starting from any prior, therefore, the dynamics conditional on no breakthrough imply convergence in finite time to some $\hat{p}_{(\ell)}$, and as both players play risky there, learning will once more be complete.

For intermediate and high stakes, learning will thus always be complete in equilibrium, exactly as it would be in the planner's solution for which (4) applies with p^m and $1 - p^m$

replaced by the cutoffs $\bar{p}$ and $1-\bar{p}$, respectively.

We summarize these findings in

Proposition 8 (Asymptotics of learning) *In any Markov perfect equilibrium of the experimentation game, the probability of learning the true state of the world as* $t \to \infty$ *is the same as in the planner's solution. It is smaller than 1 (incomplete learning) for* $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$*, and equal to 1 (complete learning) for* $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$*. For* $\frac{g}{s} = \frac{2r+\lambda}{r+\lambda}$*, both complete and incomplete learning are consistent with efficiency, and both can arise in equilibrium.*

Note that these asymptotics only depend on the position of the single-agent cutoffs. Intuitively, for both players to play safe, neither of them can be more optimistic than his single-agent cutoff. At any belief in the set $[0, 1-p^*[\cup]p^*, 1]$, therefore, at least one player must play risky and thus keep the learning process alive. This set is the entire unit interval if and only if $p^* < \frac{1}{2}$, that is, $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$.

In Keller, Rady and Cripps (2005), where players face risky arms of a common type, any Markov perfect equilibrium implies an inefficiently small probability of learning the true state in the long run. As all players become gradually more pessimistic, the incentive to free-ride makes them give up experimentation earlier than in the planner's solution. With risky arms of opposite type, by contrast, it can never be the case that both players are simultaneously very pessimistic about their individual prospects. Whenever the stakes are so high that the planner would want both players to experiment at a given belief, therefore, at least one player is willing to experiment on his own at this belief. Free-riding incentives can then delay the resolution of uncertainty relative to the social optimum, but not prevent it. The following proposition derives an upper bound on the expected delay.

Proposition 9 (Speed of learning) *For* $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$ *and any initial belief, there exists a Markov perfect equilibrium in cutoff strategies such that the expected delay in the resolution of uncertainty is less than* $\frac{1}{3}$ *of the expected time by which all uncertainty is resolved in the planner's solution.*

PROOF: See Appendix C. ▮

As we shall see next, the optimality of long-run learning outcomes and the short expected delay in the resolution of uncertainty are reflected in surprisingly good welfare properties of the Markov perfect equilibria.

4.5 Welfare

When stakes are low, the unique MPE has players use their single-agent cutoffs, which is efficient. For intermediate stakes, an efficient equilibrium *outcome* can be achieved with cutoff strategies if and only if the interval of possible equilibrium cutoffs given in Proposition 6 contains the efficient cutoff.

It is straightforward to verify that $1-p^m \leq \bar{p}$ and hence $\max\{p^*, 1-p^m\} \leq \bar{p} < 1-\bar{p} \leq \min\{p^m, 1-p^*\}$ if $\frac{2r+\lambda}{r+\lambda} < \frac{g}{s} \leq \frac{2r+\lambda}{2(r+\lambda)} + \sqrt{\frac{(2r+\lambda)^2}{4(r+\lambda)^2} + \frac{\lambda}{r+\lambda}}$. Then, if the players' initial belief is $p_0 \leq \bar{p}$, the equilibrium with cutoff $\hat{p} = \bar{p}$ achieves the efficient outcome as the only beliefs that are reached with positive probability under the equilibrium strategies are given by the set $\{0,1\} \cup [p_0, \bar{p}]$, and the equilibrium strategies prescribe the efficient actions at all of these beliefs. Similarly, for $p_0 \geq 1-\bar{p}$, the efficient outcome is achieved by the equilibrium with cutoff $\hat{p} = 1-\bar{p}$. Finally, if $\bar{p} < p_0 < 1-\bar{p}$, the efficient outcome is achieved by the equilibrium with cutoff $\hat{p} = p_0$. If $\frac{2r+\lambda}{2(r+\lambda)} + \sqrt{\frac{(2r+\lambda)^2}{4(r+\lambda)^2} + \frac{\lambda}{r+\lambda}} < \frac{g}{s} < 2$, by contrast, we have $p^* < \bar{p} < 1-p^m$ and hence $\bar{p} < \max\{p^*, 1-p^m\} < \min\{p^m, 1-p^*\} < 1-\bar{p}$. In this case, the efficient outcome can only be achieved for initial beliefs $p_0 \in \{0\} \cup [1-p^m, p^m] \cup \{1\}$.

If stakes are high, the unique MPE implies efficient behavior except on the set $[\bar{p}, p^m[\cup]1-p^m, 1-\bar{p}]$. In this case, the efficient outcome arises if and only if $p_0 \in \{0\} \cup [p^m, 1-p^m] \cup \{1\}$.

Proposition 10 (Welfare) *If $\frac{g}{s} \leq \frac{2r+\lambda}{2(r+\lambda)} + \sqrt{\frac{(2r+\lambda)^2}{4(r+\lambda)^2} + \frac{\lambda}{r+\lambda}}$, then for each initial belief, there exists a Markov perfect equilibrium in cutoff strategies that achieves the efficient outcome. If $\frac{g}{s} > \frac{2r+\lambda}{2(r+\lambda)} + \sqrt{\frac{(2r+\lambda)^2}{4(r+\lambda)^2} + \frac{\lambda}{r+\lambda}}$, there are initial beliefs under which the efficient outcome cannot be reached in any Markov perfect equilibrium. For any such belief p, there exists an equilibrium in cutoff strategies such that*

$$\frac{u(p)-s}{\bar{u}(p)-s} > \frac{1}{2},$$

where $u(p)$ and $\bar{u}(p)$ are the players' average payoffs in the equilibrium and the planner's solution, respectively.

Proof: The first two statements of the proposition follow directly from the preceding discussion. The lower bound on average payoffs is established in Appendix C. ▮

The stated lower bound is straightforward to derive from the closed-form solutions for the players' payoff functions. This bound is by no means tight, however. In fact, a numerical evaluation on a grid of pairs $(\frac{r}{\lambda}, \frac{g}{s})$ suggests that for $0 < \frac{r}{\lambda} \leq 10$ and $1 < \frac{g}{s} \leq 10$, there always exists an MPE in cutoff strategies for which the above ratio exceeds 86%. To put this number in perspective, it is worthwhile recalling from Keller, Rady and Cripps (2005) that in any Markov perfect equilibrium of the experimentation game with risky arms of a common type, there is a range of beliefs at which all players play safe while the planner would want all of them to play risky. At these beliefs, the above ratio is zero.

4.6 Discussion

We have restricted attention to what in the literature have been termed “pure strategy equilibria” (by Bolton and Harris, 1999 and 2000) or “simple equilibria” (by Keller, Rady and Cripps, 2005, and Keller and Rady, 2010). An extension of the strategy space allowing players to choose experimentation intensities from the entire unit interval would leave the planner’s solution unchanged. Moreover, as the intensity of experimentation enters linearly into a player’s Bellman equation, our simple equilibria are immune against deviations to interior intensities.

While we have assumed that players are symmetric, it is straightforward to extend our analysis to those asymmetries between players that preserve a zero value of information when both players use the risky arm. This is the case if players differ in their discount rates, safe payoff levels or average sizes of lump-sum payoffs on a good risky arm. If p^* continues to denote player 1’s single-agent cutoff, player 2’s single-agent optimum is then to play risky on an interval $[0, q^*[$ with $q^* \neq 1 - p^*$; similarly, the players’ myopic cutoffs will satisfy $q^m \neq 1 - p^m$ whenever players face different stakes. As all that matters for the planner’s solution, best responses and equilibrium is the relative position of the four cutoffs, all our results extend readily, the only difference being that typically there will be no symmetric equilibrium.

Matters become more complicated if player 1, say, has a higher innate ‘ability’ than player 2, i.e. if the risky arms are characterized by arrival rates $\lambda_1 > \lambda_2$. In this case, beliefs satisfy $\dot{p} = [\lambda_2 k_2(p) - \lambda_1 k_1(p)]\, p\,(1-p)$ up to the first breakthrough, which has two major implications. First, at any transition between the action profiles $(0,1)$ and $(1,1)$, player 1 must use the interior intensity of experimentation $k_1 = \lambda_2/\lambda_1$ both in the planner’s solution and when playing a best response. As in Presman (1990), such an interior allocation is the only way to obtain a well-defined law of motion for beliefs, and we must broaden our definition of cutoff strategies accordingly. Second, on any interval of beliefs where both players use the risky arm, $\dot{p} < 0$, which leads to convex payoff functions. So the value of information is positive and the best response against the opponent’s playing risky is given by a threshold belief more pessimistic than the myopic cutoff.

If players differ only in the arrival rates of lump-sum payoffs, for example, player 1’s best response against player 2's playing risky is to apply the cutoff $\hat{p}_1 = \frac{(r+\lambda_2)s}{(r+\lambda_1)\lambda_1 h-(\lambda_1-\lambda_2)s} < p^m$ which is pinned down by smooth pasting of player 1's payoff function with the linear lower bound obtained from the constant action profile $(\lambda_2/\lambda_1, 1)$. In the high-stakes scenario where player 2 plays risky to the right of $\hat{p}_1$, his best response must then be determined via an intermediate-value argument that enforces smooth pasting, at a cutoff $\hat{p}_2$, between the two functions that describe player 2's payoffs from the action profiles $(1, 1)$ and $(1, 0)$, respectively; $\hat{p}_2$ no longer admits a representation in closed form.[12] Still, it is straightforward to establish uniqueness and efficiency of Markov perfect equilibrium as well as incomplete learning for low stakes, and complete learning for intermediate and high stakes.

The same holds true for an extension of our model where, as in Keller and Rady (2010), even a bad risky arm has a non-zero arrival rate of lump-sum payoffs, implying that whenever a risky arm generates a success, beliefs about the quality of this arm jump up to a more optimistic level, but never to full certainty. Consequently, payoff functions solve differential-difference equations. These still admit closed-form solutions, yet it is now much harder to paste them together at those beliefs where the action profile changes. When both players use the risky arm, for instance, the continuation payoffs after both an upward and a downward jump of beliefs enter the Bellman equation, so an optimal change of action must be determined jointly with these continuation payoffs. This yields nonlinear equations for optimal cutoffs without explicit solutions. As to possible equilibria for intermediate and high stakes, the best response to an opponent using the risky arm again differs from the myopic cutoff strategy. This is because the belief held immediately after a success varies with the belief held immediately before, so that expected payoffs conditional on the true state are no longer constant over the range of beliefs where both players play risky, once more leading to convex payoff functions and a positive value of information.

5 Imperfect Negative Correlation

We now extend our model by introducing a third state of the world in which both risky arms are bad. This means that the quality of the risky arm is no longer perfectly negatively correlated across players, and introduces a dimension of collective pessimism into the game, captured by the posterior probability that neither player has a good risky arm.

There are two players, $i = 1, 2$, and three states, $\theta = 0, 1, 2$, where $\theta = i \in \{1, 2\}$ signifies that player i has the only good risky arm, while $\theta = 0$ means that both risky

[12] Details on the extensions discussed in this paragraph and the next are available from the authors upon request.

arms are bad. This structure is common knowledge. We write p_θ for the common posterior probability assigned to state θ, and use the pair (p_1, p_2) as the vector of state variables. The (subjective) correlation coefficient between the types of the two risky arms is

$$\rho = -\sqrt{\frac{p_1}{1-p_1}\frac{p_2}{1-p_2}}\,,$$

which can assume any value in the interval $[-1, 0]$.

Given time paths of actions $\{k_{i,\tau}\}_{0\leq\tau\leq t}$ for $i = 1, 2$ and no breakthrough by time t, the posterior beliefs at time t are

$$p_{i,t} \quad = \quad \frac{p_{i,0}\, e^{-\lambda \int_0^t k_{i,\tau}\, d\tau}}{1 - p_{1,0} - p_{2,0} + \sum_{j=1}^{2} p_{j,0}\, e^{-\lambda \int_0^t k_{j,\tau}\, d\tau}} \qquad (i = 1, 2).$$

The corresponding differential equations are

$$\dot{p}_i \quad = \quad \lambda p_i \left(\sum_{j=1}^{2} p_j k_j - k_i \right) \qquad (i = 1, 2).$$

We note that over any time interval where the action profile $(k_1, k_2) = (1, 1)$ is played without a success, the ratio $\frac{p_2}{p_1}$ stays constant and the beliefs (p_1, p_2) move towards the origin along a straight line, expressing an increase in collective pessimism. Under the action profile $(1, 0)$, the ratio $\frac{p_2}{1-p_1-p_2}$ stays constant and the beliefs (p_1, p_2) move along a straight line $p_2 = C\,(1 - p_1)$ with a positive constant $C < 1$, expressing increases in player 1's individual pessimism, player 2's individual optimism, and both players' collective pessimism.

Writing $\mathcal{P} = \{(p_1, p_2) \in [0, 1]^2 : p_1 + p_2 \leq 1\}$, we restrict players to Markov strategies $k_i : \mathcal{P} \to [0, 1]$ with the following properties: (i) the sets $k_i^{-1}(0)$ and $k_i^{-1}(1)$ each have a connected interior in $\mathcal{P}$; (ii) the union of the closures of $k_i^{-1}(0)$ and $k_i^{-1}(1)$ is $\mathcal{P}$; (iii) the intersection of the closures of $k_i^{-1}(0)$ and $k_i^{-1}(1)$ consists of a finite number of differentiable curves; (iv) along each of these curves, k_i varies continuously with beliefs; (v) $k_i(p_1, p_2) = 0$ if $p_i = 0$, and $k_i(p_1, p_2) = 1$ if $p_i = 1$. A Markov strategy k_i is called a cutoff strategy if there exists a continuous and piecewise differentiable function $h_i : [0, 1] \to [0, 1]$ such that $k_i(p_1, p_2) = 1$ for $p_i > h_i(p_{3-i})$ and $k_i(p_1, p_2) = 0$ for $p_i < h_i(p_{3-i})$. This function merely defines the switching boundary where a player changes from one action to the other. The behavior along the boundary needs to be specified separately so as to ensure a well-defined evolution of beliefs. In some cases, this will require interior intensities of experimentation.

A pair of Markov strategies is called symmetric if $k_1(p, q) = k_2(q, p)$ for all $(p, q) \in \mathcal{P}$. For cutoff strategies, symmetry means $h_1 = h_2$. The definition of admissibility is analogous to the benchmark model of perfect negative correlation.

Player i's Bellman equation is

$$u_i(p_1, p_2) = s + k_{3-i}(p_1, p_2)\beta_i(p_1, p_2, u_i) + \max_{k_i \in [0,1]} k_i \left[b_i(p_1, p_2, u_i) - c_i(p_i)\right],$$

where $b_i(p_1, p_2, u_i) = \frac{\lambda}{r} p_i[g - u_i - (1 - p_i)\frac{\partial u_i}{\partial p_i} + p_{3-i}\frac{\partial u_i}{\partial p_{3-i}}]$, $\beta_i(p_1, p_2, u_i) = \frac{\lambda}{r} p_{3-i}[s - u_i - (1 - p_{3-i})\frac{\partial u_i}{\partial p_{3-i}} + p_i \frac{\partial u_i}{\partial p_i}]$ and $c_i(p_i) = s - p_i g$. In Appendix A.3, we use the method of characteristic curves to derive explicit expressions for the players' payoffs from the action profiles $(1, 1)$, $(1, 0)$ and $(0, 1)$. This allows us to derive the following result.

Proposition 11 (Imperfect correlation) *There always exists a symmetric Markov perfect equilibrium in cutoff strategies.*

PROOF: The proof is by construction; see the specification of equilibrium strategies below and the verification of the best-response property in Appendix C. ▮

For $\frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$, and hence $p^* \geq \frac{1}{2}$, the common equilibrium cutoff can be taken to be constant and equal to the single-agent cutoff p^*, with either player playing safe at the cutoff itself. This equilibrium is illustrated in the left panel of Figure 4, with the labels "00", "01" and "10" standing for the action profiles $(0, 0)$, $(0, 1)$ and $(1, 0)$, respectively. The intuition for this equilibrium carries over from the case of perfect negative correlation.

For $\frac{2r+\lambda}{r+\lambda} < \frac{g}{s} \leq 2$, and hence $p^* < \frac{1}{2} \leq p^m$, equilibrium cutoffs can be defined by the function $h(p) = \max\{p^*, p\}$. Along the switching boundary, player i plays safe when $p_i = p^* \geq p_{3-i}$ and risky when $p_i = p_{3-i} > p^*$. This equilibrium is illustrated in the middle panel of Figure 4. Fix a prior in the interior of the 10 region. If this prior lies below the line joining the belief (p^*, p^*) with the belief $(1, 0)$, player 1 plays risky until either a breakthrough occurs or beliefs reach the vertical segment $\{p^*\} \times [0, p^*]$, where player 1 gives up and all learning stops. In this scenario, the increase in player 2's optimism as player 1 fails to have a breakthrough is not enough to entice him to experiment himself. This is different if the prior lies above the line joining the belief (p^*, p^*) with the belief $(1, 0)$. In the absence of a breakthrough on player 1's risky arm, beliefs now move to the 45 degree line, where player 2 joins player 1 in playing risky. From that point on, beliefs move down the 45 degree line and, in the absence of a breakthrough, become stationary in the point (p^*, p^*) where both players play safe. Along the part of the 45 degree line where both players play risky, their payoff functions are kinked and the best-response property follows from the restrictions imposed by admissibility of the players' strategies, exactly as in the symmetric MPE in cutoff strategies under perfect negative correlation (corresponding to the upper right edge of the triangle).

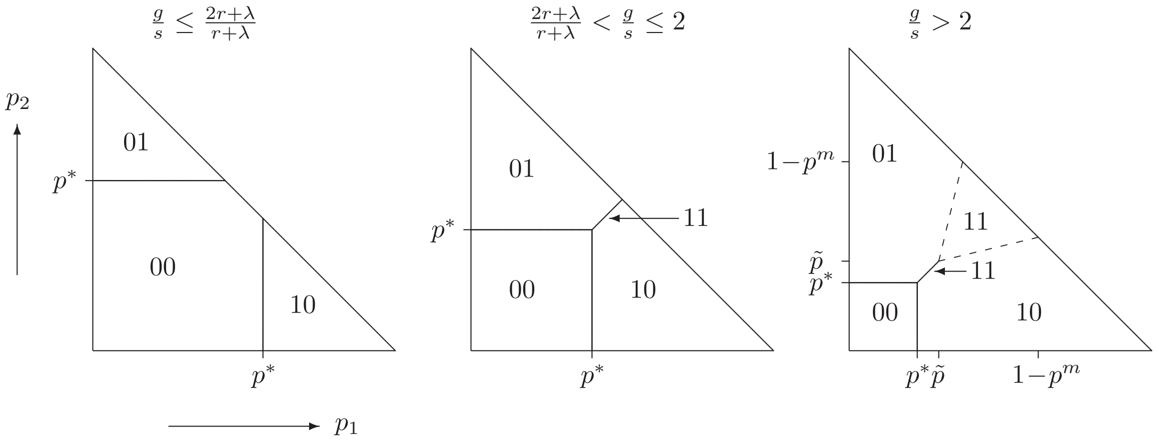

Figure 4: Cutoff equilibria of the experimentation game with imperfect negative correlation between the types of the risky arms.

For $\frac{g}{s} > 2$, and hence $p^m < \frac{1}{2}$, we define $\tilde{p} = \frac{rs}{(r+\lambda)g-2\lambda s}$, which lies between p^* and p^m. An equilibrium cutoff function is then given by $h(p) = \max\{p^*, p\}$ for $p \leq \tilde{p}$, and

$$h(p) = \frac{(r+\lambda p)s}{(r+\lambda)g - \lambda s}$$

for $p > \tilde{p}$. As to the actions chosen along the switching boundary, player i plays safe when $p_i = p^* \geq p_{3-i}$, plays risky when $p^* < p_i = p_{3-i} \leq \tilde{p}$, and sets

$$k_i = \frac{(r+\lambda)g - \lambda s}{g-s} \, \frac{p_{3-i}}{r+\lambda p_{3-i}}$$

when $p_i = h(p_{3-i}) > \tilde{p}$. This equilibrium is illustrated in the right panel of Figure 4. As we move down along player 2's switching boundary from the belief $(1-p^m, p^m)$ to the belief $(\tilde{p}, \tilde{p})$, player 2's intensity of experimentation monotonically falls from 1 to $\frac{s}{g-s}$.[13] This interior intensity is precisely the one that keeps posterior beliefs on the boundary as long as no breakthrough occurs. The boundary itself is pinned down by the requirement that given $k_1 = 1$, player 2 must have $b_2 > c_2$ above the boundary and $b_2 < c_2$ below it.[14] Once a belief on the diagonal line segment between (p^*, p^*) and $(\tilde{p}, \tilde{p})$ is reached, the evolution of beliefs and actions is the same as in the MPE for intermediate stakes.

For low stakes, the equilibrium described above is efficient. For intermediate and high stakes, the planner's solution is given by the cutoff function

$$h(p) = \max\left\{p^*, \frac{(r+\lambda p)s}{(r+\lambda)g}\right\}.$$

[13] In this and the following figure, boundaries along which some player uses an interior intensity of experimentation are shown as dashed lines.

[14] See Lemmas A.2 and A.3 in the Appendix.

The increasing part of player 2's efficient switching boundary is thus a straight line joining the beliefs (p^*, p^*) and $(1-\bar{p}, \bar{p})$, where $\bar{p}$ is the efficient cutoff for perfect negative correlation.[15] For intermediate and high stakes, therefore, the equilibria that we constructed are inefficient in that the set of beliefs at which both players use the risky arm is smaller than in the planner's solution. The set of beliefs at which learning stops is the same as in the planner's solution, though. This yields the following counterpart to Proposition 8.

Proposition 12 (Imperfect correlation, asymptotics of learning) *There always exists a Markov perfect equilibrium in which the probability of learning the true state of the world as $t \to \infty$ is the same as in the planner's solution.*

PROOF: As the equilibrium constructed for $\frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$ is efficient, we can assume that $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$ and hence $p^* < \frac{1}{2}$. By symmetry of the planner's solution and the above equilibria in cutoff strategies, we can restrict ourselves to initial beliefs $(p_{1,0}, p_{2,0})$ with $p_{1,0} \geq p_{2,0}$. By Proposition 8, we can further assume that $p_{1,0} + p_{2,0} < 1$. If this initial belief satisfies $p_{1,0} \leq p^*$ or $p_{2,0} \leq \frac{p^*}{1-p^*}(1-p_{1,0})$, the equilibrium outcome is the same as the efficient outcome. Suppose therefore that $p_{1,0} > p^*$ and $p_{2,0} > \frac{p^*}{1-p^*}(1-p_{1,0})$. In the absence of a breakthrough, both the efficient and the equilibrium paths of play then lead to the posterior belief (p^*, p^*) in finite time.

Consider any time paths of actions $\{k_{i,\tau}\}_{0 \leq \tau \leq t}$ for $i = 1, 2$ that in the absence of a breakthrough lead to the belief (p^*, p^*) by time t. Bayes' law then implies

$$p^* = \frac{p_{i,0}\, e^{-\lambda \int_0^t k_{i,\tau}\, d\tau}}{1 - p_{1,0} - p_{2,0} + \sum_{j=1}^2 p_{j,0}\, e^{-\lambda \int_0^t k_{j,\tau}\, d\tau}}$$

for $i = 1, 2$. This is a system of two linear equations in $P_i = e^{-\lambda \int_0^t k_{i,\tau}\, d\tau}$ that is easily seen to have a unique solution (P_1, P_2) for $p^* \neq \frac{1}{2}$. As P_i is the probability of no breakthrough on player i's risky arm up to time t conditional on this arm being good, the efficient and the equilibrium paths of play imply the same conditional probability of a breakthrough before all learning stops, and hence the same conditional probability of learning the true state. ∎

As in the case of perfect negative correlation, therefore, strategic interaction between the players need not lead to an inefficiently high probability of incomplete learning.

[15] The derivation of the planner's solution is very similar to the construction of the high-stakes MPE, and can be based on straightforward adaptations of Lemmas A.2 and A.3.

6 More Than Two Players

The last extension that we explore has $N \geq 3$ players, $i = 1, \ldots, N$, each playing a bandit of the exponential type. It is common knowledge that exactly one of them has a good risky arm. We write p_i for the common posterior probability that player i's risky arm is the good one.

The three-player case is easiest to visualize, and its analysis very similar to that of the two-player game with imperfect negative correlation, so we focus on this case, returning to general N only briefly at the end of the section. With $N = 3$, we can again use the pair (p_1, p_2) as the vector of state variables. Markov and cutoff strategies can be defined along the same lines as in the previous section.

Player i's Bellman equation is now

$$u_i(p_1, p_2) = s + \sum_{j \neq i} k_j \beta_{ij}(p_1, p_2, u_i) + \max_{k_i \in \{0,1\}} k_i [b_i(p_1, p_2, u_i) - c_i(p_1, p_2)].$$

Here, the learning benefits from a player's own experimentation are $b_i(p_1, p_2, u_i) = \frac{\lambda}{r} p_i [g - u_i - (1 - p_i) \frac{\partial u_i}{\partial p_i} + p_{3-i} \frac{\partial u_i}{\partial p_{3-i}}]$ for $i = 1, 2$ and $b_3(p_1, p_2, u_3) = \frac{\lambda}{r}(1 - p_1 - p_2)[g - u_3 + p_1 \frac{\partial u_3}{\partial p_1} + p_2 \frac{\partial u_3}{\partial p_2}]$. The learning benefits that accrue to player i when player $j \neq i$ uses the risky arm are $\beta_{ij}(p_1, p_2, u_i) = \frac{\lambda}{r} p_j [s - u_i - (1 - p_j) \frac{\partial u_i}{\partial p_j} + p_{3-j} \frac{\partial u_i}{\partial p_{3-j}}]$ for $j = 1, 2$ and $\beta_{i3}(p_1, p_2, u_i) = \frac{\lambda}{r}(1 - p_1 - p_2)[s - u_i + p_1 \frac{\partial u_i}{\partial p_1} + p_2 \frac{\partial u_i}{\partial p_2}]$. The opportunity costs of experimentation are $c_i(p_1, p_2) = s - p_i g$ for $i = 1, 2$, and $c_3(p_1, p_2) = s - (1 - p_1 - p_2) g$.

If the prevailing action profile is $(0, 0, 0)$, each player's payoff function equals $u_i = s$. If $(1, 1, 1)$ prevails, the payoff functions are linear, exactly as in the two-player model: $u_i = p_i g + (1 - p_i) \frac{\lambda}{r+\lambda} s$. Explicit expressions for the players' payoffs from all other action profiles can again be derived as in Appendix A.3. An equilibrium transition between the action profiles $(1, 0, 0)$ and $(0, 0, 0)$ is easily seen to require $p_1 = p^*$, while a transition between $(1, 1, 1)$ and $(0, 1, 1)$ requires $p_1 = p^m$; the intuition for these findings is exactly the same as in the two-player model with perfect negative correlation.

For $\frac{g}{s} < \frac{3r+\lambda}{r+\lambda}$, and hence $p^* > \frac{1}{3}$, the equilibria that we constructed for the two-player game with imperfect negative correlation translate one-to-one into equilibria of the three-player game. To see this, consider the triangle $\mathcal{T}$ with the corners $(\frac{1}{3}, \frac{1}{3})$, $(\frac{1}{2}, \frac{1}{2})$ and $(1, 0)$ in the (p_1, p_2)-plane; in the three-player game, this corresponds to the set of all beliefs such that $p_1 \geq p_2 \geq p_3$. On $\mathcal{T}$, let players 1 and 2 play the same strategies as in the two-player MPE constructed in the previous section, and let player 3 play safe. Given a prior belief in $\mathcal{T}$, posterior beliefs then remain in $\mathcal{T}$ unless there is a success on player 1's or, if he gets to experiment at all, player 2's risky arm. As player 3 never experiments, players 1 and 2 are facing exactly the same situation as in a two-player game between them, and thus are playing best responses. Player 3's payoff on $\mathcal{T}$ is $u_3 = s$, so $b_3 < c_3$ if and only if $p_3 < p^*$, which is obviously the case here because the inequalities $p_1 \geq p_2 \geq p_3$ imply $p_3 \leq \frac{1}{3}$. There is now a unique way to extend the players' strategies on $\mathcal{T}$ to a symmetric strategy profile on the entire state space; this strategy profile clearly constitutes an equilibrium.

For the following proposition, therefore, only parameter constellations such that $\frac{g}{s} \geq \frac{3r+\lambda}{r+\lambda}$, and hence $p^* \leq \frac{1}{3}$, require further work.

Proposition 13 (Three players) *There always exists a symmetric Markov perfect equilibrium in cutoff strategies.*

PROOF: The proof is again by construction. Equilibrium strategies for $\frac{g}{s} \geq \frac{3r+\lambda}{r+\lambda}$ are illustrated in Figure 5 below. The verification of the best-response property proceeds along the same lines as in the proof of Proposition 11. Details are available from the authors upon request. ∎

Figure 5 illustrates equilibrium strategies in the four cases that need to be distinguished when $p^* \leq \frac{1}{3}$. To emphasize the symmetry of these equilibria, beliefs are represented as elements of a standard 2-simplex. Its vertices correspond to the three possible degenerate beliefs about the state of the world; the vertex marked "1", for instance, corresponds to subjective certainty that player 1 has the good risky arm. The probability p_i that player i has the good risky arm is constant along any line running parallel to the edge that lies opposite vertex "i".

In each of the four panels of Figure 5, all players use the risky arm at the center of the simplex, where $p_1 = p_2 = p_3 = \frac{1}{3}$. When $\frac{g}{s} \leq 3$, and hence $p^m \geq \frac{1}{3}$, this is the only belief at which the action profile $(1,1,1)$ is played; when $\frac{g}{s} > 3$, and hence $p^m < \frac{1}{3}$, this profile is played at all beliefs such that $\min\{p_1, p_2, p_3\} \geq p^m$.

As to the solid lines that end in the center of the simplex in the upper two panels of Figure 5, the two players who experiment individually on either side of such a line, experiment jointly along it. In the lower left panel, the action profile along any such line is the same as in the region from where the line emanates, so that just one player experiments along it. The same goes, in the lower right panel, for the solid lines ending in the triangle where the profile $(1,1,1)$ is played. In each case, the verification of the best-response property along a solid line rests on restrictions imposed by admissibility of the players' strategies.

The dashed lines in the upper right and the two lower panels are switching boundaries of exactly the same type as in the high-stakes MPE of the two-player game with imperfect

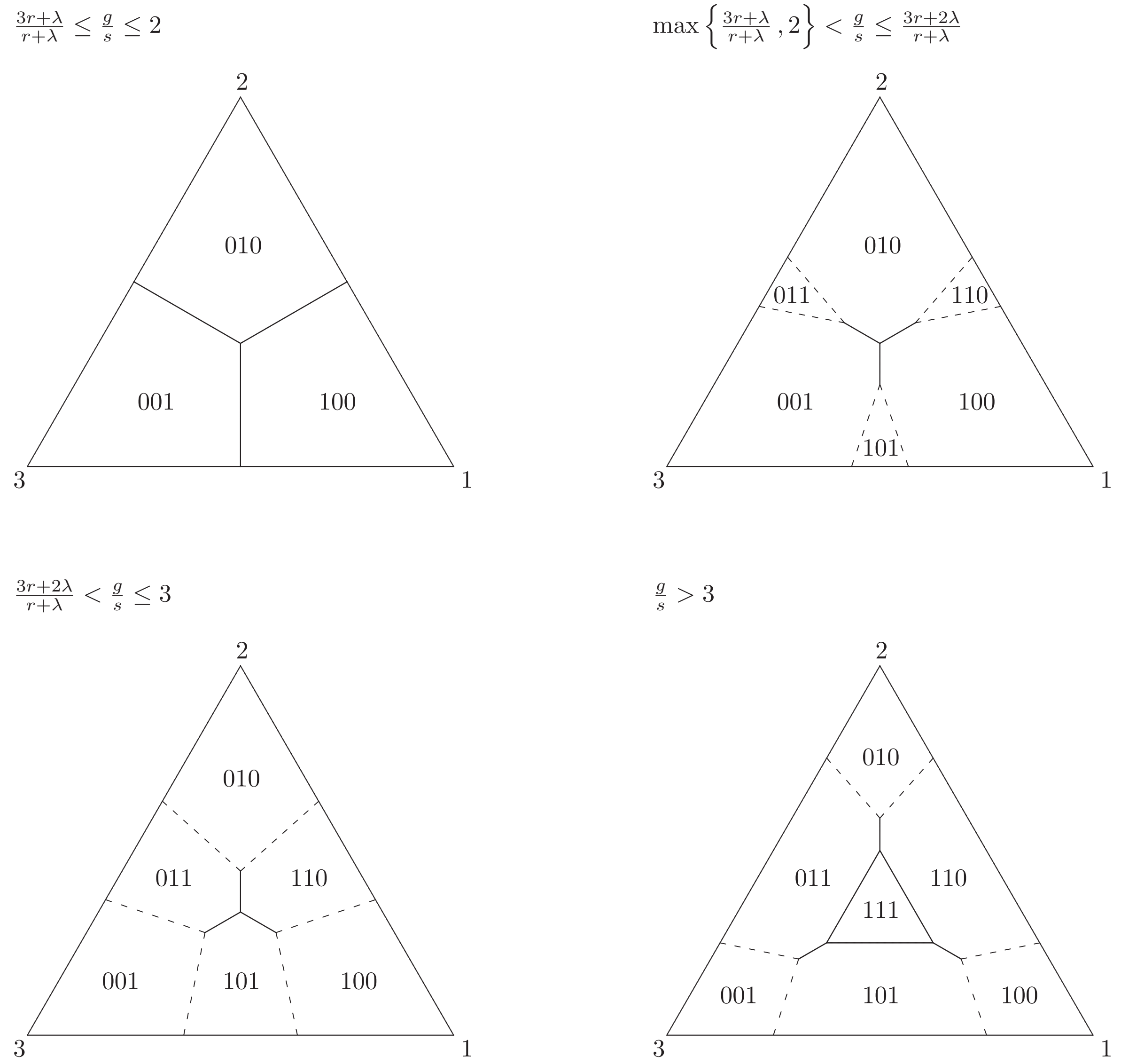
$\frac{3r+\lambda}{r+\lambda} \leq \frac{g}{s} \leq 2$
$\max\left\{\frac{3r+\lambda}{r+\lambda}, 2\right\} < \frac{g}{s} \leq \frac{3r+2\lambda}{r+\lambda}$
$\frac{3r+2\lambda}{r+\lambda} < \frac{g}{s} \leq 3$
$\frac{g}{s} > 3$
1
2
3
010
011
110
001
101
100
111

Figure 5: Cutoff equilibria of the experimentation game with three players when $p^* \leq \frac{1}{3}$.

negative correlation. The player who plays safe on one side of the boundary, and risky on the other, chooses an interior intensity of experimentation at the boundary itself, making posterior beliefs move along it as long as no breakthrough occurs.

Clearly, learning will be complete in any of the equilibria depicted in Figure 5. For $p^* \leq \frac{1}{3}$, complete learning is also efficient because the action profile $(1,1,1)$ weakly dominates the profile $(0,0,0)$ in terms of the three players' expected average payoff, so the planner has no reason ever to stop learning. For $p^* > \frac{1}{3}$, we can exploit symmetry of our equilibria as well as of the planner's solution and restrict our attention to beliefs in the set $\mathcal{T}$ defined above. On this set, the planner asks player 3 to use the safe arm, and players 1 and 2 to follow the strategies that are efficient in the two-player game with imperfect negative correlation. Invoking Proposition 12, we thus obtain

Proposition 14 (Three players, asymptotics of learning) *There always exists a Markov perfect equilibrium in which the probability of learning the true state of the world as $t \to \infty$ is the same as in the planner's solution.*

While the construction of Markov perfect equilibria becomes increasingly complex as the number of players grows, it is clear that for $\frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$, and hence $p^* \geq \frac{1}{2}$, the planner's solution remains an equilibrium for arbitrary N; as before, it lets all players use the single-agent cutoff and implies incomplete learning. More generally, a necessary condition for all N players to play safe on some non-empty open set of beliefs, be it in the planner's solution or in equilibrium, is that all elements of this set satisfy $\max\{p_1, \ldots, p_N\} \leq p^*$. For $\frac{g}{s} > \frac{Nr+\lambda}{r+\lambda}$, and hence $p^* < \frac{1}{N}$, this means that the planner's solution as well as any MPE must lead to complete learning. For $\frac{2r+\lambda}{r+\lambda} < \frac{g}{s} \leq \frac{Nr+\lambda}{r+\lambda}$, it is optimal for the planner to let all N players use the safe arm if and only if $\max\{p_1, \ldots, p_N\} \leq p^*$. We conjecture that there exist equilibria in which learning stops on the exact same set of beliefs.

7 Concluding Remarks

We have analyzed games of strategic experimentation in continuous time where players' expected risky payoffs are negatively correlated. Our first set of results concerns a game with two players and common knowledge that exactly one of them has a bandit with a good risky arm. In sharp contrast to the situation where players face risky arms of a common quality, this game always admits equilibria of the cutoff type, and equilibrium is unique and symmetric in two subsets of the parameter space. When the stakes are low, players behave as if they were single players experimenting in isolation, and this is efficient. When the stakes are high, players behave as if they were myopic. Finally, learning will be complete in equilibrium if and only if efficiency requires complete learning.

This analysis naturally raises the question under what circumstances two players would *choose* to play a strategic experimentation game with bandits of opposite, rather than common, type. To analyze this question, we can extend our model by letting players first decide sequentially whether they want to experiment with risky arm 1, whose prior probability of being good is p_0, or with risky arm 2, whose corresponding probability is $1 - p_0$. They then play the strategic experimentation game with either perfectly positively or perfectly negatively correlated bandits, as the case might be. Using the fact that in the experimentation game in Keller, Rady and Cripps (2005), no player can obtain an equilibrium payoff higher than twice the planner's solution minus the single-agent solution, it is straightforward to derive a condition on the model parameters under which equilibrium of the extended game uniquely predicts that players choose different risky arms for all priors p_0 in a neighborhood of $\frac{1}{2}$. It is easy to find parameter combinations that satisfy this condition; for instance, $\frac{r}{\lambda} = 2$ and $\frac{g}{s} = 3$ will do. Given any r and λ, moreover, the condition will always be fulfilled if the stakes $\frac{g}{s}$ are large enough.

While this extension of the model with perfect negative correlation merely allows for one irreversible project choice, Chapter 2 analyzes a variant of this setup where, akin to Chatterjee and Evans (2004), both players have access to both risky arms and can switch between them at will. This requires the players to solve identical three-armed bandit problems with a safe arm and two risky arms that are known to be of opposite types. In contrast to our setting, where the planner's solution is incentive-compatible if and only if the stakes are low enough, he finds that the planner's solution is incentive-compatible if and only if the stakes are *high* enough. This is because for sufficiently high stakes, the safe arm becomes so unattractive that the players are willing to explore the risky arm that momentarily appears more promising given that the opponent also explores the arm, which is exactly what efficiency requires.

Our second set of results concerns experimentation games with imperfect negative correlation of the type of risky arm across players. In the model with two players and a third state of the world in which neither has a good risky arm, there always exists a symmetric Markov perfect equilibrium in cutoff strategies. Although the state space is a two-dimensional simplex and the players' payoff functions solve partial differential equations, we have been able to compute players' equilibrium strategies and payoffs in closed form. Imperfect correlation introduces a dimension of collective pessimism into the model, captured by the posterior probability that both risky arms are bad. As a consequence, the planner's solution involves a set of beliefs where both players use the safe arm, so efficient learning is necessarily incomplete. In the equilibria we construct, the set of beliefs where both players play safe is the same as in the planner's solution, so learning does not stop inefficiently early; in fact, the probability of learning the true state of the world in the long run is the same as in the planner's solution. Even with imperfect negative correlation, therefore, strategic interaction does not make learning inefficient in the long run. We obtain quite similar results in a three-player game in which it is common knowledge that exactly one player has a good risky arm.

In our setting, the definition of admissible strategies turned out to be more involved than in the case of perfect positive correlation. As a matter of fact, this difficulty arises as soon as the (binary) type of the risky arm is not perfectly positively correlated across players, that is, as soon as there is a positive probability that they might have risky arms of opposite type. To see this in a two-player setting, consider the four possible states of the world: $\theta = 0$ (no player has a good risky arm), $\theta = 1$ (player 1 has the only good risky arm), $\theta = 2$ (player 2 has the only good risky arm), and $\theta = 3$ (both players have a good risky arm), and write p_θ for the probability that the players assign to state θ. As long as there is no breakthrough, we then have $\dot{p}_3 = -\lambda p_3 \{(1-p_3)(k_1+k_2) - p_1 k_1 - p_2 k_2\} \leq 0$; for $p_1 = p_2 = 0$, this reduces to the dynamics in Keller, Rady & Cripps (2005), where each pair of strategies that are left-continuous in p_3 is admissible. For $i = 1, 2$, on the other hand, we have $\dot{p}_i = \lambda p_i \{p_i k_i + p_{3-i} k_{3-i} - k_i + p_3(k_1+k_2)\}$; for $p_3 = 0$, this reduces to the dynamics in the imperfect-correlation version of our model. In particular, $\dot{p}_1 = -\lambda p_1(1-p_1-p_3)$ when the action profile is $(k_1, k_2) = (1, 0)$, $\dot{p}_1 = \lambda p_1(p_2+p_3)$ when the action profile is $(0, 1)$, and similarly for $\dot{p}_2$. Whenever $p_i > 0$, therefore, the sign of $\dot{p}_i$ depends on the action profile, which means that, as in our model, player i's admissible strategies cannot be defined without reference to the other player's strategy. This also applies to a scenario of imperfect positive correlation obtained for p_1 and p_2 positive but small. Thus, the admissibility issues showing up in our model are the "generic" phenomenon, while the case of perfect positive correlation is truly exceptional because it is one of only two cases in which the space of admissible strategy pairs is a product set, the other being the trivial case of independent types.

Throughout our analysis, we have maintained the assumption that both actions and outcomes were publicly observable at all times. Bonatti & Hörner (2010) investigate varying correlations of bandit types between players under the assumption of private actions and publicly observable outcomes, but in their setup everybody switches to playing safe at the myopic cutoff. The effect of allowing for private actions when there is an incentive to play risky beyond the myopic cutoff has not been investigated yet. One of our main conclusions appears robust to such an extension of our model: for sufficiently high stakes, it cannot be common knowledge that all players have stopped using the risky arm, so there must be complete learning in equilibrium. We leave a full analysis of such a model to future work.

Appendix

A Payoff Functions

For $p \in [0,1]$, we define

$$w_1(p) = pg + (1-p)\frac{\lambda}{r+\lambda}s \quad \text{and} \quad w_2(p) = (1-p)g + p\frac{\lambda}{r+\lambda}s = w_1(1-p).$$

These are the players' payoff functions when both are playing risky. For the explicit representation of other payoff functions, it will be convenient to define

$$u_0(p) = (1-p)\left(\frac{1-p}{p}\right)^{\frac{r}{\lambda}}.$$

Note that

$$u_0'(p) = -\frac{\frac{r}{\lambda}+p}{p(1-p)}\,u_0(p)$$

and $u_0'' > 0$.

A.1 Explicit Solutions for Perfect Negative Correlation

On any open interval where $k_1(p) = 1$ and $k_2(p) = 0$, u_1 and u_2 satisfy the ODEs

$$\begin{aligned} \lambda p(1-p)u_1'(p) + (r+\lambda p)u_1(p) &= (r+\lambda)pg, \\ \lambda p(1-p)u_2'(p) + (r+\lambda p)u_2(p) &= (r+\lambda p)s, \end{aligned}$$

which have the solutions $u_1(p) = pg + C_1 u_0(p)$ and $u_2(p) = s + C_2 u_0(p)$ with constants C_1 and C_2.

Finally, on any open interval where $k_1(p) = 0$ and $k_2(p) = 1$, u_1 and u_2 solve

$$\begin{aligned} \lambda p(1-p)u_1'(p) - [r+\lambda(1-p)]u_1(p) &= -[r+\lambda(1-p)]s, \\ \lambda p(1-p)u_2'(p) - [r+\lambda(1-p)]u_2(p) &= -(r+\lambda)(1-p)g, \end{aligned}$$

hence $u_1(p) = s + D_1 u_0(1-p)$ and $u_2(p) = (1-p)g + D_2 u_0(1-p)$ with constants D_1 and D_2.

Note that each of the above closed-form solutions is the sum of one term that expresses the expected payoff from committing to a particular action and another term that captures the option value of being able to switch to the other action.

A.2 An Auxiliary Result for Perfect Negative Correlation

The following lemma will be useful in the proofs of Lemma B.4 and Proposition 3.

Lemma A.1 *On any open interval of beliefs where the payoff function of player i satisfies $u_i(p) = s + \beta_i(p, u_i)$, the sign of $b_i(p, u_i) - c_i(p)$ coincides with the sign of $w_i(p) - u_i(p)$.*

Proof: We first note that $b_i(p,u_i) + \beta_i(p,u_i) = \frac{\lambda}{r}\left[\overline{u}_i(p) - u_i(p)\right]$ where $\overline{u}_1(p) = pg + (1-p)s$ and $\overline{u}_2(p) = \overline{u}_1(1-p)$ are the players' expected full-information payoffs. As $\beta_i(p,u_i) = u_i(p) - s$, this implies $b_i(p,u_i) - c_i(p) = \frac{\lambda}{r}\left[\overline{u}_i(p) - u_i(p)\right] - u_i(p) + s - c_i(p) = \frac{r+\lambda}{r}\left[w_i(p) - u_i(p)\right]$. ∎

A.3 Explicit Solutions in the Three-State Model

The laws of motion for p_1 and p_2 under the action profile $(1,0)$ are $\dot{p}_1 = -\lambda p_1(1-p_1)$ and $\dot{p}_2 = \lambda p_1 p_2$. The resulting partial differential equation for player 1's payoff function is

$$\lambda p_1(1-p_1)\frac{\partial u_1}{\partial p_1} - \lambda p_1 p_2 \frac{\partial u_1}{\partial p_2} + (r+\lambda p_1)u_1 = (r+\lambda)p_1 g.$$

A particular solution is $u(p_1,p_2) = p_1 g$, that is, the payoff from committing to $(1,0)$ forever.

We look for solutions to the homogeneous PDE of the form $u_1(p_1,p_2) = (1-p_1)v(p_1,p_2)$, so that v must solve the PDE

$$\lambda p_1(1-p_1)\frac{\partial v}{\partial p_1} - \lambda p_1 p_2 \frac{\partial v}{\partial p_2} + rv = 0.$$

Along a trajectory $(p_{1,t}, p_{2,t})_{t\geq 0}$, this implies

$$\frac{d}{dt}v(p_{1,t}, p_{2,t}) = rv(p_{1,t}, p_{2,t})$$

and hence

$$v(p_{1,t}, p_{2,t}) = e^{rt}v(p_{1,0}, p_{2,0}).$$

We now note that under the action profile $(1,0)$, the posterior probability for player 1's arm being good in the absence of a breakthrough is

$$p_{1,t} = \frac{p_{1,0}e^{-\lambda t}}{p_{1,0}e^{-\lambda t} + 1 - p_{1,0}}\ ,$$

implying

$$e^{rt} = \left(\frac{1-p_{1,0}}{p_{1,0}}\right)^{-\frac{r}{\lambda}} \left(\frac{1-p_{1,t}}{p_{1,t}}\right)^{\frac{r}{\lambda}}.$$

Therefore,

$$v(p_{1,t}, p_{2,t})\left(\frac{1-p_{1,t}}{p_{1,t}}\right)^{-\frac{r}{\lambda}}$$

is constant along the trajectory. As each trajectory is uniquely described by its slope $\frac{p_2}{1-p_1}$, we thus have

$$v(p_1,p_2) = f_{10}^1\!\left(\frac{p_2}{1-p_1}\right)\left(\frac{1-p_1}{p_1}\right)^{\frac{r}{\lambda}}$$

with some differentiable univariate function f_{10}^1. This yields the following general form for player 1's payoff function under $(1,0)$:

$$u_1(p_1,p_2) \quad = \quad p_1 g + f_{10}^1\!\left(\frac{p_2}{1-p_1}\right) u_0(p_1).$$

Player 2's payoff function under the action profile $(1,0)$ satisfies

$$\lambda p_1(1-p_1)\frac{\partial u_2}{\partial p_1} - \lambda p_1 p_2 \frac{\partial u_2}{\partial p_2} + (r+\lambda p_1)u_2 = (r+\lambda p_1)s,$$

for which the same steps as above yield the general solution

$$u_2(p_1,p_2) = s + f_{10}^2\left(\frac{p_2}{1-p_1}\right) u_0(p_1).$$

Straightforward computations show that the corresponding benefit of experimentation is

$$\begin{aligned} b_2(p_1,p_2,u_2) &= \frac{\lambda}{r}\, p_2\,(g-s) \\ &\quad - \left[\frac{r+\lambda}{r}\,\frac{p_2}{1-p_1}\, f_{10}^2\left(\frac{p_2}{1-p_1}\right) + \frac{\lambda}{r}\,\frac{p_2\,(1-p_1-p_2)}{(1-p_1)^2}\,(f_{10}^2)'\left(\frac{p_2}{1-p_1}\right)\right] u_0(p_1). \end{aligned}$$

Under the action profile $(1,1)$, the PDE for player i's payoff function,

$$\lambda p_i(1-p_1-p_2)\frac{\partial u_i}{\partial p_i} + \lambda p_{3-i}(1-p_1-p_2)\frac{\partial u_i}{\partial p_{3-i}} + (r+\lambda(p_1+p_2))u_i = (r+\lambda)p_i g + \lambda p_{3-i} s,$$

has the general solution

$$u_i(p_1,p_2) = p_i g + p_{3-i}\frac{\lambda}{\lambda+r}s + f_{11}^i\left(\frac{p_2}{p_1}\right) u_0(p_1+p_2).$$

Here, one finds

$$b_2(p_1,p_2,u_2) = \left[\frac{p_2}{p_1+p_2} f_{11}^2\left(\frac{p_2}{p_1}\right) - \frac{\lambda}{r}\,\frac{p_2}{p_1}(f_{11}^2)'\left(\frac{p_2}{p_1}\right)\right] u_0(p_1+p_2).$$

A.4 Auxiliary Results for the Three-State Model

The following lemma will be helpful in constructing Markov perfect equilibria in cutoff strategies. It is a simple consequence of value matching.

Lemma A.2 *Consider an interval $I =]p_\ell, p_r[$ with $0 < p_\ell < p_r < 1$ and a differentiable function $h\colon \to]0,1[$ with $h(p_1) < 1-p_1$ and $-\frac{h(p_1)}{1-p_1} \leq h'(p_1) \leq \frac{h(p_1)}{p_1}$. Assume that at any belief (p_1,p_2) on the graph $\mathcal{H}$ of h, player 1 sets $k_1 = 1$ while player 2 chooses*

$$k_2(p_1) = \frac{p_1}{h(p_1)}\,\frac{h(p_1)+(1-p_1)h'(p_1)}{1-h(p_1)+p_1 h'(p_1)}.$$

Starting at a belief $(p_1,p_2) \in \mathcal{H}$, posterior beliefs move along $\mathcal{H}$ until either a breakthrough occurs or the belief $(p_\ell, h(p_\ell))$ is reached; fixing a continuation payoff at that belief, let $U(p_1)$ be player 2's resulting payoff. For small $\epsilon > 0$, let $u_2^\uparrow$ be the solution to the equation $u_2 = s + \beta_2$ on $\{(p_1,p_2)\colon p_\ell < p_1 < p_r,\ p_2 < h(p_1)+\epsilon\}$ with $u_2^\uparrow(p_1,h(p_1)) = U(p_1)$ on I, and $u_2^\downarrow$ the solution to the equation $u_2 = s+\beta_2+b_2-c_2$ on $\{(p_1,p_2)\colon p_\ell < p_1 < p_r,\ p_2 > h(p_1)-\epsilon\}$ with $u_2^\downarrow(p_1,h(p_1)) = U(p_1)$ on I. Then,

$$k_2(p_1)\left[b_2(p_1,h(p_1),u_2^\uparrow) - c_2(h(p_1))\right] = [1-k_2(p_1)]\left[b_2(p_1,h(p_1),u_2^\downarrow) - c_2(h(p_1))\right] = 0$$

on I.

PROOF: We suppress the argument p_1 whenever this is expedient. The bounds on h' ensure that $0 \leq k_2 \leq 1$. Moreover,

$$\dot{p_2} = \lambda p_2[p_1 + p_2 k_2 - k_2] = h'\lambda p_1[p_1 + p_2 k_2 - 1] = h'\dot{p_1},$$

which means that starting from $(p_1, p_2) \in \mathcal{H}$, the action profile $(1, k_2)$ makes posterior beliefs move along $\mathcal{H}$ until a breakthrough occurs or the left endpoint of $\mathcal{H}$ is reached.

On I, the payoff U satisfies the ODE

$$rU = r\{k_2 h g + [1-k_2]s\} + \lambda p_1[s-U] + \lambda k_2 h[g-U] - \lambda p_1[1-p_1-k_2 h]U'.$$

As $U(p_1) = u_2^{\uparrow}(p_1, h(p_1))$ on I, we have

$$U'(p_1) = \frac{\partial u_2^{\uparrow}}{\partial p_1}(p_1, h(p_1)) + \frac{\partial u_2^{\uparrow}}{\partial p_2}(p_1, h(p_1))h'(p_1).$$

Suppressing the argument $(p_1, h(p_1))$ in $u_2^{\uparrow}$ and its derivatives, we can thus rewrite the ODE for U as

$$ru_2^{\uparrow} = r\{k_2 h g + [1-k_2]s\} + \lambda p_1[s-u_2^{\uparrow}] + \lambda k_2 h[g-u_2^{\uparrow}] - \lambda p_1[1-p_1-k_2 h]\left(\frac{\partial u_2^{\uparrow}}{\partial p_1} + \frac{\partial u_2^{\uparrow}}{\partial p_2}h'\right).$$

As $p_1[1-p_1-k_2 h]h' = h[(1-h)k_2 - p_1]$, the previous equation is easily seen to transform into

$$u_2^{\uparrow}(p_1, h(p_1)) = s + \beta_2(p_1, h(p_1), u_2^{\uparrow}) + k_2(p_1)\left[b_2(p_1, h(p_1), u_2^{\uparrow}) - c_2(h(p_1))\right].$$

For $p_2 < h(p_1)$, however, $u_2^{\uparrow}(p_1, p_2) = s + \beta_2(p_1, p_2, u_2^{\uparrow})$. Continuity of $u_2^{\uparrow}$ and its derivatives implies $k_1(p_1)\left[b_2(p_1, h(p_1), u_2^{\uparrow}) - c_2(h(p_1))\right] = 0$.

Arguing exactly as above, one also shows that

$$u_2^{\downarrow}(p_1, h(p_1)) = s + \beta_2(p_1, h(p_1), u_2^{\downarrow}) + k_2(p_1)\left[b_2(p_1, h(p_1), u_2^{\downarrow}) - c_2(h(p_1))\right].$$

For $p_2 > h(p_1)$, we now have $u_2^{\downarrow}(p_1, p_2) = s + \beta_2(p_1, p_2, u_2^{\downarrow}) + b_2(p_1, h(p_1), u_2^{\downarrow}) - c_2(h(p_1))$. So continuity of $u_2^{\downarrow}$ and its derivatives implies $[1-k_2(p_2)]\left[b_2(p_1, h(p_1), u_2^{\downarrow}) - c_2(h(p_1))\right] = 0$. ∎

While the previous result applies to all possible switching boundaries, the next lemma uses necessary and sufficient conditions for optimality to derive constraints on the location of such a boundary in equilibrium.

Lemma A.3 *Let the players use an admissible strategy pair, giving player 2 the payoff function* u_2. *Fix a belief* $(\hat{p}_1, \hat{p}_2)$ *with* $\hat{p}_1 > 0$, $\hat{p}_2 > 0$ *and* $0 < \hat{p}_1 + \hat{p}_2 < 1$, *and define the rays* $\mathcal{R}_{11} = \left\{(p_1, p_2)\colon \hat{p}_1 < p_1 < \frac{\hat{p}_1}{\hat{p}_1 + \hat{p}_2},\ p_2 = \frac{\hat{p}_2}{\hat{p}_1} p_1\right\}$ *and* $\mathcal{R}_{10} = \left\{(p_1, p_2)\colon \hat{p}_1 < p_1 < 1,\ p_2 = \frac{\hat{p}_2}{1-\hat{p}_1}(1-p_1)\right\}$.

(1) Suppose that both players use the risky arm on $\mathcal{R}_{11}$, *and that* $b_2(p_1, p_2, u_2)$ *converges to* $c_2(\hat{p}_2)$ *as* (p_1, p_2) *approaches* $(\hat{p}_1, \hat{p}_2)$ *along* $\mathcal{R}_{11}$. *Then player 2 is playing a best response on* $\mathcal{R}_{11}$ *if and only if* $\hat{p}_2 \geq \frac{(r+\lambda\hat{p}_1)s}{(r+\lambda)g - \lambda s}$.

(2) Suppose that player 1 uses the risky arm, and player 2 the safe arm, on $\mathcal{R}_{10}$, *and that* $b_2(p_1, p_2, u_2)$ *converges to* $c_2(\hat{p}_2)$ *as* (p_1, p_2) *approaches* $(\hat{p}_1, \hat{p}_2)$ *along* $\mathcal{R}_{10}$. *Then player 2 is playing a best response on* $\mathcal{R}_{10}$ *if and only if* $\hat{p}_2 \leq \frac{(r+\lambda\hat{p}_1)s}{(r+\lambda)g - \lambda s}$.

PROOF: (1) Writing $\gamma = \frac{\hat{p}_2}{\hat{p}_1}$, we have

$$b_2(p_1, p_2, u_2) = \left[\frac{\gamma}{1+\gamma} f_{11}^2(\gamma) - \frac{\lambda\gamma}{r}(f_{11}^2)'(\gamma)\right] u_0([1+\gamma]p_1)$$

on $\mathcal{R}_{11}$. By assumption, this converges to $c_2(\hat{p}_2)$ as $p_1 \downarrow \hat{p}_1$, so we have

$$b_2(p_1, p_2, u_2) = \frac{c_2(\hat{p}_2)}{u_0(\hat{p}_1 + \hat{p}_2)}\, u_0([1+\gamma]p_1)$$

on $\mathcal{R}_{11}$. If $\hat{p}_2 < p^m$, and hence $c_2(\hat{p}_2) > 0$, convexity of $u_0([1+\gamma]p_1)$ and linearity of $c_2(\gamma p_1)$ imply that $b_2 \geq c_2$ on $\mathcal{R}_{11}$ if and only if

$$\frac{c_2(\hat{p}_2)}{u_0(\hat{p}_1 + \hat{p}_2)} u_0'(\hat{p}_1 + \hat{p}_2)[1+\gamma] \geq -\gamma g.$$

This condition is easily seen to be equivalent to the inequality $\hat{p}_2 \geq \frac{(r+\lambda\hat{p}_1)s}{(r+\lambda)g-\lambda s}$. If $\hat{p}_2 \geq p^m$, then b_2 is constant or increasing in p_1 along $\mathcal{R}_{11}$, whereas c_2 is decreasing, which implies $b_2 \geq c_2$. We complete the proof by noting that $\frac{(r+\lambda\hat{p}_1)s}{(r+\lambda)g-\lambda s} < p^m$ for all $\hat{p}_1 < 1 - p^m$.

(2) Writing $\eta = \frac{\hat{p}_2}{1-\hat{p}_1}$, we have

$$b_2(p_1, p_2, u_2) = \frac{\lambda\eta}{r}(g-s)(1-p_1) - \eta\left[\frac{r+\lambda}{r} f_{10}^2(\eta) + \frac{\lambda}{r}(1-\eta)(f_{10}^2)'(\eta)\right] u_0(p_1)$$

on $\mathcal{R}_{10}$. By assumption, this converges to $c_2(\hat{p}_2)$ as $p_1 \downarrow \hat{p}_1$, so we have

$$b_2(p_1, p_2, u_2) = \frac{\lambda\eta}{r}(g-s)(1-p_1) + \left[c_2(\hat{p}_2) - \frac{\lambda}{r}(g-s)\hat{p}_2\right]\frac{u_0(p_1)}{u_0(\hat{p}_1)}$$

on $\mathcal{R}_{10}$. If $\hat{p}_2 > p^*$, we have $c_2(\hat{p}_2) < \frac{\lambda}{r}(g-s)\hat{p}_2$, so convexity of u_0 and linearity of c_2 imply that $b_2 \leq c_2$ on $\mathcal{R}_{10}$ if and only if

$$-\frac{\lambda\eta}{r}(g-s) + \left[c_2(\hat{p}_2) - \frac{\lambda}{r}(g-s)\hat{p}_2\right]\frac{u_0'(\hat{p}_1)}{u_0(\hat{p}_1)} \leq \eta g.$$

This is easily seen to be equivalent to the inequality $\hat{p}_2 \leq \frac{(r+\lambda\hat{p}_1)s}{(r+\lambda)g-\lambda s}$. If $\hat{p}_2 \leq p^*$, we have $c_2(\hat{p}_2) \geq \frac{\lambda}{r}(g-s)\hat{p}_2$, so b_2 is constant or decreasing in p_1 along $\mathcal{R}_{10}$, whereas c_2 is increasing, which implies $b_2 \leq c_2$. We complete the proof by noting that $\frac{(r+\lambda\hat{p}_1)s}{(r+\lambda)g-\lambda s} > p^*$ for all $\hat{p}_1 > 0$. ∎

B Admissible Pairs of Markov Strategies

We start with three examples.

Example 1: Suppose that player 1 plays risky at all beliefs $p > \frac{1}{2}$ and safe otherwise, while player 2 plays risky at all beliefs $p \leq \frac{1}{2}$ and safe otherwise. Then there is no continuous function $t \mapsto p_t$ with $p_0 = \frac{1}{2}$ that satisfies equation (1) at all $t \geq 0$. Suppose to the contrary that there exists such a solution. If there is a time t such that $p_t > p_0$, then continuity implies that there exists a $t' < t$ such that $\frac{1}{2} < p_{t'} < p_t$ and $p_\tau > \frac{1}{2}$ for all τ in $[t', t]$. Yet, $k_1(p_\tau) = 1$ and $k_2(p_\tau) = 0$ on $[t', t]$, so (1) implies $p_t < p_{t'}$, a contradiction. The symmetric argument rules out the existence of a time t such that $p_t < p_0$. So the only candidate for a solution to (1) is the constant function $p_t \equiv \frac{1}{2}$. With this function, $k_1(p_t) = 0$ and $k_2(p_t) = 1$ at all t, but then p_t must be increasing by (1), another contradiction. Starting from the prior belief $p_0 = \frac{1}{2}$, therefore, there is no solution to the law of motion for beliefs consistent with the above strategies, which means that these strategies do not pin down the players' actions.

Example 2: Suppose that player 1 plays risky whenever $\frac{1}{4} \leq p < \frac{1}{2}$ and safe whenever $\frac{1}{2} \leq p \leq \frac{3}{4}$; his behavior at other beliefs is irrelevant for this example. Player 2 always plays safe. For each $T \in [0, \infty]$, the continuous function $t \mapsto p_t$ given by

$$p_t = \begin{cases} \frac{1}{2} & \text{for} \quad 0 \leq t \leq T, \\ \frac{e^{-\lambda(t-T)}}{e^{-\lambda(t-T)}+1} & \text{for} \quad \quad t > T \end{cases}$$

then satisfies (1) up to the time $T + \tau$ at which it reaches the belief $\frac{1}{4}$. This means that the given Markov strategies are consistent with a continuum of different solutions to the law of motion of beliefs in continuous time. If we discretize time with fixed increment $\Delta t > 0$ and approximate (2) by

$$p_{t+\Delta t} - p_t = \lambda \left[k_2(p_t) - k_1(p_t)\right] p_t \left(1 - p_t\right) \Delta t$$

for $t = 0, \Delta t, 2\Delta t, \ldots$, however, there is a unique solution with $p_0 = \frac{1}{2}$, namely $p_t = \frac{1}{2}$ for all $t = n\Delta t$. The only continuous-time solution that can be approximated by the discrete-time solution as $\Delta t \downarrow 0$ is the constant function $p_t \equiv \frac{1}{2}$, corresponding to $T = \infty$.

Example 3: Suppose that player 1 plays risky and player 2 plays safe whenever $\frac{1}{4} \leq p < \frac{1}{2}$, while player 1 plays safe and player 2 plays risky whenever $\frac{1}{2} \leq p \leq \frac{3}{4}$. Behavior at other beliefs is again irrelevant. Then there are two different solutions to (1) starting in $p_0 = \frac{1}{2}$,

$$p_t = \frac{e^{-\lambda t}}{e^{-\lambda t} + 1} \quad \text{and} \quad p_t = \frac{e^{\lambda t}}{e^{\lambda t} + 1} \, .$$

Only the latter is consistent with a discrete-time approximation as in Example 2.

In Examples 1 and 2, existence and uniqueness of solutions to the law of motion of beliefs in a neighborhood of $\frac{1}{2}$ can be restored by imposing specific one-sided continuity requirements on the players' strategies. In the first example, it suffices to make player 1's strategy right-continuous at the belief $\frac{1}{2}$, and in the second example, left-continuous. The appropriate one-sided continuity requirement in these examples thus depends on the opponent's strategy. In Example 3, moreover, no combination of one-sided continuity requirements on the two players' strategies can ensure uniqueness. We therefore do not require uniqueness of the law of motion of beliefs in our definition of admissible strategy pairs. Instead, whenever there are multiple continuous-time solutions, we shall select the solution that is obtained in the limit of discrete-time approximations.

The following result shows that the problem of non-existence of solutions to the law of motion of beliefs in continuous time would arise even if we were to restrict the space of strategies to less complex functions such as cutoff strategies. It also establishes that the set of admissible strategy pairs is not a product set.

Lemma B.1 *There exist admissible pairs of cutoff strategies (k_1, k_2) and $(\tilde{k}_1, \tilde{k}_2)$ such that $(k_1, \tilde{k}_2)$ is inadmissible.*

PROOF: We take $k_1^{-1}(1) = [\frac{1}{4}, 1]$, $k_2^{-1}(1) = [0, \frac{3}{4}]$, $\tilde{k}_1^{-1}(1) =]\frac{2}{3}, 1]$, and $\tilde{k}_2^{-1}(1) = [0, \frac{1}{3}[$. Then each of the pairs (k_1, k_2) and $(\tilde{k}_1, \tilde{k}_2)$ implies a unique solution to the law of motion of beliefs from any starting point, whereas for $(k_1, \tilde{k}_2)$, non-existence of a solution starting from $p_0 = \frac{1}{3}$ follows exactly as in Example 1. ∎

B.1 Admissible Transitions

We say that the *transition* (k_1^-, k_2^-)—(k_1, k_2)—(k_1^+, k_2^+) occurs at the belief $\hat{p} \in]0, 1[$ if $\lim_{p \uparrow \hat{p}}(k_1(p), k_2(p)) = (k_1^-, k_2^-)$, $(k_1(\hat{p}), k_2(\hat{p})) = (k_1, k_2)$, $\lim_{p \downarrow \hat{p}}(k_1(p), k_2(p)) = (k_1^+, k_2^+)$, and at least one of the sets $\{k_1^-, k_1, k_1^+\}$ and $\{k_2^-, k_2, k_2^+\}$ contains more than one element. Given our definition of strategies, each MPE has a finite number of transitions. We call a transition *admissible* if it can arise under an admissible pair of Markov strategies.

We can rewrite (1) as

$$p_t = \left[1 + \frac{1 - p_0}{p_0} e^{-\lambda \int_0^t \Delta(p_\tau)\, d\tau}\right]^{-1}, \tag{B.1}$$

with $\Delta(p) = k_2(p) - k_1(p)$. For any belief $\hat{p}$ in the open unit interval, we define $\Delta(\hat{p}-) = \lim_{p \uparrow \hat{p}} \Delta(p)$ and $\Delta(\hat{p}+) = \lim_{p \downarrow \hat{p}} \Delta(p)$. For the purposes of this section, we shall consider two transitions at the beliefs $\hat{p}$ and $\tilde{p}$ as equivalent if $\Delta(\hat{p}-) = \Delta(\tilde{p}-)$, $\Delta(\hat{p}) = \Delta(\tilde{p})$, and $\Delta(\hat{p}+) = \Delta(\tilde{p}+)$. For the remainder of this section, we shall only be concerned with the so defined equivalence classes of transitions which we denote by triplets $(\Delta(\hat{p}-), \Delta(\hat{p}), \Delta(\hat{p}+))$. Since $\Delta(p) \in \{-1, 0, 1\}$ for all $p \in [0, 1]$, there are 27 such triplets. Two of them, $(-1, -1, -1)$ and $(1, 1, 1)$, do not correspond to any change in action profile. A third one, $(0, 0, 0)$, corresponds to a transition if and only if players switch between $(1, 1)$ and $(0, 0)$; the associated dynamics are trivial. A further eight classes, $(1, 0, 1)$, $(1, -1, 1)$, $(1, -1, 0)$, $(0, 1, -1)$, $(0, -1, 1)$, $(-1, 1, 0)$, $(-1, 1, -1)$ and $(-1, 0, -1)$, are ruled out by our requirement that both $k_i^{-1}(0)$ and $k_i^{-1}(1)$ be disjoint unions of a finite number of non-degenerate intervals. For each of the remaining classes, we are interested in solutions to (B.1) with initial condition $p_0 = \hat{p}$ (the belief at which the transition occurs).

No Solution

Arguing as in Example 1, it is straightforward to establish that there is no solution to (B.1) with $p_0 = \hat{p}$ for the following classes:

- $(1, 1, -1)$, $(1, -1, -1)$, $(1, 1, 0)$, $(0, 1, 0)$, $(0, -1, 0)$, $(0, -1, -1)$.

A Continuum of Solutions

As in Example 2, there exists a continuum of solutions to (B.1) with $p_0 = \hat{p}$ for each of the following classes:

- $(0,0,1)$, $(-1,0,1)$, $(-1,0,0)$.

We select the constant solution $p_t \equiv \hat{p}$ because this is the one obtained as the limit of any discrete-time approximation.

Exactly Two Solutions

The logic of Example 3 applies to both the following classes:

- $(-1,1,1)$ and $(-1,-1,1)$.

Consistency with a discrete-time approximation leads us to select the solution $p_t = \left[1 + \frac{1-\hat{p}}{\hat{p}}\, e^{-\lambda t}\right]^{-1}$ for $(-1,1,1)$ and the solution $p_t = \left[1 + \frac{1-\hat{p}}{\hat{p}}\, e^{\lambda t}\right]^{-1}$ for $(-1,-1,1)$.

A Unique Solution

Each of the remaining five classes implies a unique continuous-time solution to the law of motion of beliefs (equal to the limit of any discrete-time approximation) in a neighborhood of $\hat{p}$:

- $(1,0,0)$, $(0,1,1)$, $(-1,-1,0)$, $(0,0,-1)$, $(1,0,-1)$.

Admissible Classes and Transitions

The following table lists the admissible classes and the transitions that they represent.

Class	Transitions
$(1,0,0)$	$(0,1)$—$(1,1)$—$(1,1)$, $(0,1)$—$(0,0)$—$(0,0)$
$(1,0,-1)$	$(0,1)$—$(0,0)$—$(1,0)$, $(0,1)$—$(1,1)$—$(1,0)$
$(0,1,1)$	$(0,0)$—$(0,1)$—$(0,1)$, $(1,1)$—$(0,1)$—$(0,1)$
$(0,0,1)$	$(0,0)$—$(0,0)$—$(0,1)$, $(1,1)$—$(1,1)$—$(0,1)$
$(0,0,0)$	$(0,0)$—$(0,0)$—$(1,1)$, $(0,0)$—$(1,1)$—$(1,1)$, $(1,1)$—$(0,0)$—$(0,0)$, $(1,1)$—$(1,1)$—$(0,0)$
$(0,0,-1)$	$(0,0)$—$(0,0)$—$(1,0)$, $(1,1)$—$(1,1)$—$(1,0)$
$(-1,1,1)$	$(1,0)$—$(0,1)$—$(0,1)$
$(-1,0,1)$	$(1,0)$—$(0,0)$—$(0,1)$, $(1,0)$—$(1,1)$—$(0,1)$
$(-1,0,0)$	$(1,0)$—$(0,0)$—$(0,0)$, $(1,0)$—$(1,1)$—$(1,1)$
$(-1,-1,1)$	$(1,0)$—$(1,0)$—$(0,1)$
$(-1,-1,0)$	$(1,0)$—$(1,0)$—$(0,0)$, $(1,0)$—$(1,0)$—$(1,1)$

Table 1: Admissible transitions

This table yields the following characterization of admissible strategy pairs.

Lemma B.2 *A pair of Markov strategies (k_1, k_2) is admissible if and only if all its finitely many transitions appear in Table 1. Starting from a prior belief p_0 equal to a transition point $\hat{p}$, the evolution of beliefs is fully determined by $\Delta(\hat{p}) = k_2(\hat{p}) - k_1(\hat{p})$: $p_t \equiv \hat{p}$ if $\Delta(\hat{p}) = 0$; $p_t = \left[1 + \frac{1-\hat{p}}{\hat{p}}\, e^{-\lambda t}\right]^{-1}$ if $\Delta(\hat{p}) = 1$; and $p_t = \left[1 + \frac{1-\hat{p}}{\hat{p}}\, e^{\lambda t}\right]^{-1}$ if $\Delta(\hat{p}) = -1$. These solutions are valid as long as there is no breakthrough on a risky arm and no other transition is reached.*

Remarks

The six classes that do not admit a solution in continuous time would either lead to a short "blip" in discrete time before reaching an absorbing state, as is the case with the classes $(1, 1, 0)$, $(0, 1, 0)$, $(0, -1, 0)$ and $(0, -1, -1)$), or to an oscillating solution, which, as we reduce period length, leads to stasis in the limit, as is the case with the classes $(1, 1, -1)$ and $(1, -1, -1)$. While we rule these classes out, the limits of their discrete-time solutions are still available through other admissible strategy pairs. The continuous-time limit of the discrete-time solutions associated with the class $(1, 1, 0)$, for instance, is captured by the admissible class $(1, 0, 0)$. Similarly, the limit of the discrete-time oscillations implied by the classes $(1, 1, -1)$ or $(1, -1, -1)$ is captured by the admissible class $(1, 0, -1)$.

Each inadmissible strategy pair has but a finite number of inadmissible transitions. Each of these can be made admissible by changing one player's action at the belief where the transition occurs. This means that for each inadmissible strategy pair (k_1, k_2), there exists an admissible pair $(\tilde{k}_1, \tilde{k}_2)$ such that $\tilde{k}_i$ differs from k_i at finitely many points.

B.2 Locating Admissible Transitions

We first consider those admissible transitions in which one player's action remains fixed.

Lemma B.3 *The following statements hold for all Markov perfect equilibria:*

(i) $(0, 0)$—$(0, 0)$—$(1, 0)$ *can only occur at the belief* p^*; $(1, 0)$—$(1, 0)$—$(0, 0)$ *and* $(1, 0)$—$(0, 0)$—$(0, 0)$ *can only occur if* $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$ *and only at beliefs in* $[1 - p^*, p^*[$.

(ii) $(0, 1)$—$(0, 0)$—$(0, 0)$ *can only occur at the belief* $1 - p^*$; $(0, 0)$—$(0, 1)$—$(0, 1)$ *and* $(0, 0)$—$(0, 0)$—$(0, 1)$ *can only occur if* $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$ *and only at beliefs in* $]1 - p^*, p^*]$.

(iii) $(0, 1)$—$(1, 1)$—$(1, 1)$ *can only occur at the belief* p^m; $(1, 1)$—$(0, 1)$—$(0, 1)$ *and* $(1, 1)$—$(1, 1)$—$(0, 1)$ *can only occur if* $\frac{g}{s} > 2$ *and only at beliefs in* $]p^m, 1 - p^m]$.

(iv) $(1, 1)$—$(1, 1)$—$(1, 0)$ *can only occur at the belief* $1 - p^m$; $(1, 0)$—$(1, 0)$—$(1, 1)$ *and* $(1, 0)$—$(1, 1)$—$(1, 1)$ *can only occur if* $\frac{g}{s} > 2$ *and only at beliefs in* $[p^m, 1 - p^m[$.

Proof: Suppose the transition $(0, 0)$—$(0, 0)$—$(1, 0)$ occurs at $\hat{p}$. Starting to the immediate right of $\hat{p}$, the dynamics of beliefs in the absence of a breakthrough converge to $\hat{p}$, so u_1 is continuous at this belief. If $\hat{p} > p^*$, then $u_1 = s$ implies $b_1 > c_1$ to the immediate left of $\hat{p}$, so player 1 would deviate to playing risky there. If $\hat{p} < p^*$, then the solution to the ODE $u_1 = s + b_1 - c_1$ with $u_1(\hat{p}) = s$ has $u_1'(\hat{p}+) < 0$, so player 1 would deviate to playing safe to the immediate right of $\hat{p}$. So we must have $\hat{p} = p^*$.

Now, suppose the transition $(1,0)$—$(1,0)$—$(0,0)$ occurs at $\hat{p}$. If $\hat{p} \geq p^*$, then $u_1 = s$ implies $b_1 > c_1$ to the immediate right of $\hat{p}$, so player 1 would deviate to playing risky there. If $\hat{p} < 1 - p^*$, then $u_2 = s$ implies $b_2 > c_2$ to the immediate right of $\hat{p}$, so player 2 would deviate to playing risky there. So we must have $1 - p^* \leq \hat{p} < p^*$, which requires $p^* > \frac{1}{2}$, that is, $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$.

The same arguments apply to the transition $(1,0)$—$(0,0)$—$(0,0)$. This proves part (i). Part (ii) is the mirror image of part (i) with the players' roles reversed.

As to part (iii), suppose the transition $(0,1)$—$(1,1)$—$(1,1)$ occurs at $\hat{p}$. Starting to the immediate left of $\hat{p}$, the dynamics of beliefs in the absence of a breakthrough converge to $\hat{p}$, so u_1 is continuous at this belief. If $\hat{p} < p^m$, then $u_1 = w_1$ implies $b_1 = 0 < c_1$ to the immediate right of $\hat{p}$, so player 1 would deviate to playing safe there. If $\hat{p} > p^m$, then the solution to the ODE $u_1 = s + \beta_1$ with $u_1(\hat{p}) = w_1(\hat{p})$ has $u_1'(\hat{p}-) > w_1'(\hat{p}-)$, so player 1 would deviate to playing risky to the immediate left of $\hat{p}$. So we must have $\hat{p} = p^m$ (with smooth pasting).

Next, suppose the transition $(1,1)$—$(0,1)$—$(0,1)$ occurs at $\hat{p}$. If $\hat{p} \leq p^m$, then $u_1 = w_1$ implies $b_1 = 0 < c_1$ to the immediate left of $\hat{p}$, so player 1 would deviate to playing safe there. If $\hat{p} > 1 - p^m$, then $u_2 = w_2$ implies $b_2 = 0 < c_2$ to the immediate left of $\hat{p}$, so player 2 would deviate to playing safe there. So we must have $p^m < \hat{p} \leq 1 - p^m$, which requires $p^m < \frac{1}{2}$, that is, $\frac{g}{s} > 2$.

The same arguments apply to the transition $(1,1)$—$(1,1)$—$(0,1)$. This proves part (iii) and, by symmetry, part (iv). ∎

Next, we pin down the conditions under which the admissible transitions in the classes $(1,0,-1)$ and $(-1,0,1)$ may occur in equilibrium.

Lemma B.4 *The following statements hold for all Markov perfect equilibria. (i) The transition* $(0,1)$—$(0,0)$—$(1,0)$ *can only occur if* $\frac{g}{s} = \frac{2r+\lambda}{r+\lambda}$ *and only at belief* $\frac{1}{2}$*. (ii) The transition* $(0,1)$—$(1,1)$—$(1,0)$ *can only occur if* $\frac{2r+\lambda}{r+\lambda} \leq \frac{g}{s} \leq 2$ *and only at beliefs in* $[\max\{1 - p^m, p^*\}, \min\{p^m, 1 - p^*\}]$*. (iii) The transition* $(1,0)$—$(0,0)$—$(0,1)$ *can only occur if* $\frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$ *and only at beliefs in* $[1 - p^*, p^*]$*. (iv) The transition* $(1,0)$—$(1,1)$—$(0,1)$ *can only occur if* $\frac{g}{s} \geq 2$ *and only at beliefs in* $[p^m, 1 - p^m]$.

PROOF: Suppose the transition $(0,1)$—$(0,0)$—$(1,0)$ occurs at belief $\hat{p}$. As the dynamics of beliefs are convergent, the players' payoff functions are continuous at $\hat{p}$ with $u_1(\hat{p}) = u_2(\hat{p}) = s$. If $\hat{p} < p^*$, then the solution to the ODE $u_1 = s + b_1 - c_1$ with $u_1(\hat{p}) = s$ has $u_1'(\hat{p}+) < 0$, so player 1 would deviate to playing safe to the immediate right of $\hat{p}$. So we must have $\hat{p} \geq p^*$. Immediately to the left of $\hat{p}$, $u_1 \geq w_1$ by Lemma A.1. If $\hat{p} > p^*$, continuity implies $u_1(\hat{p}) \geq w_1(\hat{p}) > s$, which is a contradiction. This shows that $\hat{p} = p^*$. The analogous steps for player 2 establish that $\hat{p} = 1 - p^*$. So we must have $p^* = 1 - p^* = \frac{1}{2}$, which requires $\frac{g}{s} = \frac{2r+\lambda}{r+\lambda}$. This proves statement (i).

Suppose now that the transition $(0,1)$—$(1,1)$—$(1,0)$ occurs at belief $\hat{p}$. As the dynamics of beliefs are convergent, the players' payoff functions are continuous at $\hat{p}$ with $u_i(\hat{p}) = w_i(\hat{p})$. If

$\hat{p} > p^m$, then the solution to the ODE $u_1 = s + \beta_1$ with $u_1(\hat{p}) = w_1(\hat{p})$ has $u_1'(\hat{p}-) > w_1'(\hat{p})$, so player 1 would deviate to playing risky to the immediate left of $\hat{p}$. If $\hat{p} < p^*$, then $w_1(\hat{p}) < s$ and, by continuity, $u_1 < s$ to the immediate right of $\hat{p}$, so player 1 would deviate to playing safe there. This shows that $p^* \leq \hat{p} \leq p^m$. The analogous steps for player 2 establish that $1 - p^m \leq \hat{p} \leq 1 - p^*$. So we must have $p^* \leq 1 - p^*$, which requires $p^* \leq \frac{1}{2}$, that is, $\frac{g}{s} \geq \frac{2r+\lambda}{r+\lambda}$. On the other hand, we must have $1 - p^m \leq p^m$, which requires $p^m > \frac{1}{2}$, that is, $\frac{g}{s} < 2$. This proves statement (ii).

Next, suppose the transition $(1,0)$—$(0,0)$—$(0,1)$ occurs at belief $\hat{p}$. This implies $u_1(\hat{p}) = u_2(\hat{p}) = s$. Starting close to $\hat{p}$, the dynamics of beliefs in the absence of a breakthrough are divergent, so u_1 and u_2 need not be continuous at this belief. We can establish one-sided continuity, though. If $u_1(\hat{p}-) > s$, player 1 would deviate to playing risky at $\hat{p}$ (note that this deviation yields an admissible transition again). If $u_1(\hat{p}-) < s$, player 1 would deviate to playing safe immediately to the left of $\hat{p}$. So u_1 must be left-continuous at this belief. By symmetry, u_2 must be right-continuous. Now, if $\hat{p} > p^*$, then the solution to the ODE $u_1 = s + b_1 - c_1$ with $u_1(\hat{p}) = s$ has $u_1'(\hat{p}-) > 0$, so player 1 would deviate to playing safe to the immediate left of $\hat{p}$. This implies $\hat{p} \leq p^*$. The analogous argument for player 2 yields $\hat{p} \geq 1 - p^*$. So we must have $p^* \geq \frac{1}{2}$, that is, $\frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$. This proves statement (iii).

Finally, suppose the transition $(1,0)$—$(1,1)$—$(0,1)$ occurs at belief $\hat{p}$, so that $u_1(\hat{p}) = w_1(\hat{p})$ and $u_2(\hat{p}) = w_2(\hat{p})$. If we had $u_1(\hat{p}+) > w_1(\hat{p})$, player 1 would deviate to playing safe at $\hat{p}$ (yielding another admissible transition again), so we must have $u_1(\hat{p}+) \leq w_1(\hat{p})$. But play of $(0,1)$ to the immediate right of $\hat{p}$ requires $u_1 \geq w_1$ there by Lemma A.1, hence $u_1(\hat{p}+) \leq w_1(\hat{p})$. So u_1 is right-continuous at $\hat{p}$, and u_2 left-continuous by symmetry. Now, if $\hat{p} < p^m$, then the solution to the ODE $u_1 = s + \beta_1$ with $u_1(\hat{p}) = w_1(\hat{p})$ has $u_1'(\hat{p}+) < w_1'(\hat{p})$, so player 1 would deviate to playing risky to the immediate right of $\hat{p}$. This implies $\hat{p} \geq p^m$. The analogous argument for player 2 yields $\hat{p} \leq 1 - p^m$. So we must have $p^m \leq \frac{1}{2}$, that is, $\frac{g}{s} \geq 2$. This proves statement (iv). ▌

Finally, we show that transitions in the class $(0,0,0)$ cannot arise in equilibrium.

Lemma B.5 *The transitions* $(0,0)$—$(0,0)$—$(1,1)$, $(0,0)$—$(1,1)$—$(1,1)$, $(1,1)$—$(0,0)$—$(0,0)$ *and* $(1,1)$—$(1,1)$—$(0,0)$ *do not occur in any Markov perfect equilibrium.*

Proof: By symmetry, it is enough to establish the claim for the transitions $(1,1)$—$(0,0)$—$(0,0)$ and $(1,1)$—$(1,1)$—$(0,0)$. Suppose that the former occurs at $\hat{p}$, so that $u_1(\hat{p}) = w_1(\hat{p})$ and $u_2(\hat{p}) = w_2(\hat{p})$. If $\hat{p} > p^*$, then $w_1(\hat{p}) > s$ and player 1 has an incentive to deviate to playing risky at $\hat{p}$ (which yields an admissible transition again). If $\hat{p} < p^*$, then $w_1(\hat{p}) < s$ and player 1 has an incentive to deviate to playing safe immediately to the left of $\hat{p}$. So we must have $\hat{p} = p^*$. But then player 1 has an incentive to deviate to playing risky immediately to the right of $\hat{p}$. An analogous argument rules out the transition $(1,1)$—$(1,1)$—$(0,0)$. ▌

The only admissible transitions that Lemmas B.3–B.5 do not cover are $(1,0)$—$(0,1)$—$(0,1)$ and $(1,0)$—$(1,0)$—$(0,1)$. We shall see in the proofs of Propositions 4–7 that they can only occur in those equilibria for intermediate stakes that involve jump discontinuities in the players' value functions.

C Proofs

Proof of Proposition 1

The policy (k_1, k_2) implies a well-defined law of motion for the posterior belief. The planner's payoff function from this policy is

$$u(p) = \begin{cases} \frac{1}{2}\left[s + (1-p)g + (s - p^*g)\frac{u_0(1-p)}{u_0(1-p^*)}\right] & \text{if} \quad p \leq 1 - p^*, \\ s & \text{if} \quad 1 - p^* \leq p \leq p^*, \\ \frac{1}{2}\left[s + pg + (s - p^*g)\frac{u_0(p)}{u_0(p^*)}\right] & \text{if} \quad p \geq p^*. \end{cases}$$

This function satisfies value matching and smooth pasting at p^* and $1-p^*$, hence is of class C^1. It is decreasing on $[0, 1-p^*]$ and increasing on $[p^*, 1]$. Moreover, $u = s + B_2 - \frac{c_2}{2}$ on $[0, 1-p^*]$, $u = s$ on $[1-p^*, p^*]$, and $u = s + B_1 - \frac{c_1}{2}$ on $[p^*, 1]$ (we drop the arguments for simplicity).

To show that u and the policy (k_1, k_2) solve the planner's Bellman equation, and hence that (k_1, k_2) is optimal, it is enough to establish that $B_1 < \frac{c_1}{2}$ and $B_2 > \frac{c_2}{2}$ on $]0, 1-p^*[$, $B_1 < \frac{c_1}{2}$ and $B_2 < \frac{c_2}{2}$ on $]1-p^*, p^*[$, and $B_1 > \frac{c_1}{2}$ and $B_2 < \frac{c_2}{2}$ on $]p^*, 1[$. Consider this last interval. There, $u = s + B_1 - \frac{c_1}{2}$ and $u > s$ (by monotonicity of u) immediately imply $B_1 > \frac{c_1}{2}$. Next, $B_2 = \frac{\lambda}{r}[\frac{g+s}{2} - u] - B_1 = \frac{\lambda}{r}[\frac{g+s}{2} - u] - u + s - \frac{c_1}{2}$; this is smaller than $\frac{c_2}{2}$ if and only if $u > u_{11}$, which holds here since $u > s$ and $s > u_{11}$. The other two intervals are treated in a similar way. $\blacksquare$

Proof of Proposition 3

Suppose $k_2^{-1}(1) = [0, \hat{p}_2[$ with $\hat{p}_2 \leq p^*$. Then, player 1's payoff from the strategy $k_1^{-1}(1) =]p^*, 1]$ is his single-agent payoff $u_1 = u_1^*$, that is, $u_1 = s$ on $[0, p^*]$ and $u_1 = s + b_1 - c_1$ on $[p^*, 1]$. To show that u_1 and the policy k_1 solve player 1's Bellman equation given player 2's strategy k_2, and hence that k_1 is a best response to k_2, it is enough to establish that $b_1 < c_1$ on $]0, p^*[$ and $b_1 > c_1$ on $]p^*, 1[$. On this last interval, $u_1 = s + b_1 - c_1$ and $u_1 > s$ (by monotonicity of $u_1 = u_1^*$) immediately imply $b_1 > c_1$. On $]0, p^*[$, we have $u_1 = s$ and $u_1' = 0$, hence $b_1 - c_1 = \frac{\lambda}{r}p(g - s) - (s - pg) = \frac{(r+\lambda)g - \lambda s}{r}\, p - s < 0$.

Next, suppose $k_2^{-1}(1) = [0, \hat{p}_2]$ with $\hat{p}_2 \geq p^m$. Then, player 1's payoff from the strategy $k_1^{-1}(1) = [p^m, 1]$ is given by

$$u_1(p) = \begin{cases} s + (1-p^m)\frac{\lambda}{r+\lambda}s\,\frac{u_0(1-p)}{u_0(1-p^m)} & \text{if} \quad p \leq p^m, \\ pg + (1-p)\frac{\lambda}{r+\lambda}s & \text{if} \quad p^m \leq p \leq \hat{p}_2, \\ pg + (1-\hat{p}_2)\frac{\lambda}{r+\lambda}s\,\frac{u_0(p)}{u_0(\hat{p}_2)} & \text{if} \quad p \geq \hat{p}_2. \end{cases}$$

We note that u_1 is of class C^1 except at $\hat{p}_2$, where its first derivative jumps downward; moreover, u_1 is increasing and satisfies $u_1 = s + \beta_1$ on $[0, p^m[$, $u_1 = s + \beta_1 + b_1 - c_1 = w_1$ on $[p^m, \hat{p}_2]$, and $u_1 = s + b_1 - c_1$ on $]\hat{p}_2, 1]$. On $]0, p^m[$, it is easily verified that $u_1 > w_1$, so Lemma A.1 implies $b_1 < c_1$. At p^m, we have $b_1 = 0 = c_1$. On $]p^m, \hat{p}_2[$, we have $b_1 = 0 > c_1$. On $]\hat{p}_2, 1[$, $u_1 = s + b_1 - c_1$ and $u_1 > s$ (by monotonicity of u_1) also imply $b_1 > c_1$. To complete the proof that k_1 is a best response to k_2, it suffices to note that there are no admissible strategy pairs $(\tilde{k}_1, k_2)$ for which

$\tilde{k}_1(\hat{p}_2) = 0$. In fact, any strategy $\tilde{k}_1$ with $\tilde{k}_1(\hat{p}_2) = 0$ would give rise to a transition in a class $(\Delta(\hat{p}_2-), 1, \Delta(\hat{p}_2+))$ with $\Delta(\hat{p}_2+) \in \{-1, 0\}$, and none of these is admissible.

Finally, suppose $k_2^{-1}(1) = [0, \hat{p}_2]$ with $p^* < \hat{p}_2 \le p^m$. Then player 1's payoff from playing $k_1^{-1}(1) = [\hat{p}_2, 1]$ is given by

$$u_1(p) = \begin{cases} s + \left[\hat{p}_2 g + (1-\hat{p}_2)\frac{\lambda}{r+\lambda}s - s\right] \frac{u_0(1-p)}{u_0(1-\hat{p}_2)} & \text{if } p \le \hat{p}_2, \\ pg + (1-\hat{p}_2)\frac{\lambda}{r+\lambda}s \, \frac{u_0(p)}{u_0(\hat{p}_2)} & \text{if } p \ge \hat{p}_2. \end{cases}$$

The function u_1 is of class C^1 except at $\hat{p}_2$, where its derivative jumps downward; moreover, it is increasing and satisfies $u_1 = s + \beta_1$ on $[0, \hat{p}_2[$, $u_1(\hat{p}_2) = w_1(\hat{p}_2)$ and $u_1 = s + b_1 - c_1$ on $]\hat{p}_2, 1]$. As $u_1 > w_1$ on $[0, \hat{p}_2[$, we have $b_1 < c_1$ on this interval by Lemma A.1. On $]\hat{p}_2, 1]$, we have $u_1 > s$, hence $b_1 > c_1$. At the belief $\hat{p}_2$ itself, the same argument as in the previous paragraph establishes that there are no admissible strategy pairs $(\tilde{k}_1, k_2)$ for which $\tilde{k}_1(\hat{p}_2) = 0$.

Analogous arguments apply to player 2. ▌

Proof of Proposition 4

It remains to prove uniqueness of the equilibrium for $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$. Of the transitions covered by Lemmas B.3–B.4, the following nine could occur: $(0,0)$—$(0,0)$—$(1,0)$ at p^*; $(1,0)$—$(1,0)$—$(0,0)$ and $(1,0)$—$(0,0)$—$(0,0)$ in $[1-p^*, p^*[$; $(0,1)$—$(0,0)$—$(0,0)$ at $1-p^*$; $(0,0)$—$(0,1)$—$(0,1)$ and $(0,0)$—$(0,0)$—$(0,1)$ in $]1-p^*, p^*]$; $(0,1)$—$(1,1)$—$(1,1)$ at p^m; $(1,1)$—$(1,1)$—$(1,0)$ at $1-p^m$; $(1,0)$—$(0,0)$—$(0,1)$ in $[1-p^*, p^*]$. In addition, the transitions $(1,0)$—$(0,1)$—$(0,1)$ and $(1,0)$—$(1,0)$—$(0,1)$ could potentially arise. Moving from left to right along the unit interval, we consider possible sequences of transitions leading from $(k_1(0), k_2(0)) = (0,1)$ to $(k_1(1), k_2(1)) = (1,0)$.

Players have two ways to transition out of $(k_1(0), k_2(0)) = (0,1)$: either into $(1,1)$ at p^m, or into $(0,0)$ at $1-p^*$. The former is incompatible with $(k_1(1), k_2(1)) = (1,0)$ as there is no possible transition out of $(1,1)$ to the right of p^m. So players have to transition from $(0,1)$ to $(0,0)$ at $1-p^*$.

The available transitions out of $(0,0)$ lead to $(0,1)$ or $(1,0)$. The only transition out of $(0,1)$ available to the right of $1-p^*$ would lead to $(1,1)$ at p^m, which we have already ruled out. Therefore, players must transition out of $(0,0)$ into $(1,0)$ at p^*.

To the right of p^*, the only available transitions out of $(1,0)$ lead into $(0,1)$, and the only available transition out of $(0,1)$ leads into $(1,1)$, which we have ruled out before. So there cannot be any further transition to the right of p^*. ▌

Proof of Proposition 5

For uniqueness when $\frac{g}{s} \ge 2$, we note that of the transitions covered in Lemmas B.3–B.4, the following nine might occur: $(0,0)$—$(0,0)$—$(1,0)$ at p^*; $(0,1)$—$(0,0)$—$(0,0)$ at $1-p^*$; $(0,1)$—$(1,1)$—$(1,1)$ at p^m; $(1,1)$—$(0,1)$—$(0,1)$ and $(1,1)$—$(1,1)$—$(0,1)$ in $]p^m, 1-p^m]$; $(1,1)$—$(1,1)$—$(1,0)$ at $1-p^m$; $(1,0)$—$(1,0)$—$(1,1)$ and $(1,0)$—$(1,1)$—$(1,1)$ in $[p^m, 1-p^m[$; $(1,0)$—$(1,1)$—$(0,1)$ in $[p^m, 1-p^m]$. In addition, the transitions $(1,0)$—$(0,1)$—$(0,1)$ and $(1,0)$—$(1,0)$—$(0,1)$ could potentially arise.

Players have two ways to transition out of $(k_1(0), k_2(0)) = (0, 1)$: either into $(0, 0)$ at $1 - p^*$, or into $(1, 1)$ at p^m. The former is incompatible with $(k_1(1), k_2(1)) = (1, 0)$ as there is no possible transition out of $(0, 0)$ to the right of $1 - p^*$. Therefore, players have to transition from $(0, 1)$ to $(1, 1)$ at p^m.

The available transitions out of $(1, 1)$ lead to $(0, 1)$ or $(1, 0)$. The only transition out of $(0, 1)$ available to the right of p^m would lead to $(0, 0)$ at $1 - p^*$, which we have already ruled out. So players must transition out of $(1, 1)$ into $(1, 0)$ at $1 - p^m$.

To the right of $1 - p^m$, the only available transitions out of $(1, 0)$ lead into $(0, 1)$, and the only available transition out of $(0, 1)$ leads into $(0, 0)$, which we have ruled out before. So there cannot be any further transition to the right of $1 - p^m$. ∎

Proof of Proposition 7

We fix one of the beliefs $\tilde{p}_{(\ell)}$ and introduce two auxiliary functions. Let $y\colon [\hat{p}_{(\ell-1)}, 1] \to [s, g]$ be the unique solution of the ODE $y(p) = s + b_1(p, y) - c_1(p)$ with initial value $y(\hat{p}_{(\ell-1)}) = w_1(\hat{p}_{(\ell-1)})$, and $z\colon [0, \hat{p}_{(\ell)}] \to [s, g]$ the unique solution of the ODE $z(p) = s + \beta_1(p, z)$ with terminal value $z(\hat{p}_{(\ell)}) = w_1(\hat{p}_{(\ell)})$. As $y(p) = pg + Cu_0(p)$ and $z(p) = s + Du_0(1 - p)$ for some positive constants C and D, both functions are strictly increasing and strictly convex. As $y(1) = g = w_1(1)$ and $z(0) = s > w_1(0)$, convexity implies $y < w_1$ on $]\hat{p}_{(\ell-1)}, 1[$ and $z > w_1$ on $]0, \hat{p}_{(\ell)}[$. Player 1's payoff function satisfies $u_1 = y$ on $[\hat{p}_{(\ell-1)}, \tilde{p}_{(\ell)}[$ and $u_1 = z$ on $]\tilde{p}_{(\ell)}, \hat{p}_{(\ell)}]$. This implies that $u_1(\tilde{p}_{(\ell)}-) = y(\tilde{p}_{(\ell)}) < w_1(\tilde{p}_{(\ell)}) < z(\tilde{p}_{(\ell)}) = u_1(\tilde{p}_{(\ell)}+)$, so u_1 has a jump discontinuity at $\tilde{p}_{(\ell)}$.

If the action profile played at $\tilde{p}_{(\ell)}$ is $(0, 1)$, then $u_1(\tilde{p}_{(\ell)}) = z(\tilde{p}_{(\ell)}) > w_1(\tilde{p}_{(\ell)})$, and player 1 has no incentive to deviate since the action profile $(1, 1)$ would give him the payoff $w_1(\tilde{p}_{(\ell)})$. If the action profile played at $\tilde{p}_{(\ell)}$ is $(1, 0)$, then $u_1(\tilde{p}_{(\ell)}) = y(\tilde{p}_{(\ell)}) > s$, and player 1 has no incentive to deviate since the action profile $(0, 0)$ would give him the payoff s. In either case, player 1 thus plays a best response.

Analogous arguments apply to player 2. This establishes that the strategy pairs described in the proposition constitute Markov perfect equilibria and that both players' payoffs jump at each of the beliefs $\tilde{p}_{(\ell)}$.

To see that there are no other equilibria, we note that of the transitions covered by Lemmas B.3–B.4 the following five could occur: $(0, 0)$—$(0, 0)$—$(1, 0)$ at p^*; $(0, 1)$—$(0, 0)$—$(0, 0)$ at $1 - p^*$; $(0, 1)$—$(1, 1)$—$(1, 1)$ at p^m; $(1, 1)$—$(1, 1)$—$(1, 0)$ at $1 - p^m$; $(0, 1)$—$(1, 1)$—$(1, 0)$ in $I = [\max\{1 - p^m, p^*\}, \min\{p^m, 1 - p^*\}]$. In addition, the transitions $(1, 0)$—$(0, 1)$—$(0, 1)$ and $(1, 0)$—$(1, 0)$—$(0, 1)$ could potentially arise.

Players have three ways to transition out of $(k_1(0), k_2(0)) = (0, 1)$: either into $(0, 0)$ at $1 - p^*$, or into $(1, 1)$ at p^m, or into $(1, 0)$ at some belief in I. The transition into $(0, 0)$ is incompatible with $(k_1(1), k_2(1)) = (1, 0)$ as there is no possible transition out of $(0, 0)$ to the right of $1 - p^*$. The transition into $(1, 1)$ is also incompatible with $(k_1(1), k_2(1)) = (1, 0)$ as there is no possible transition out of $(1, 1)$ to the right of p^m. Therefore, there must be a belief $\hat{p}_{\min} \in I$ such that the action profile $(0, 1)$ is played on $[0, \hat{p}_{\min}[$ and the transition $(0, 1)$—$(1, 1)$—$(1, 0)$ occurs at $\hat{p}_{\min}$.

By the same sequence of arguments started at $(k_1(1), k_2(1)) = (1, 0)$, there must also exist a belief $\hat{p}_{\max} \geq \hat{p}_{\min}$ in I such that the action profile $(1, 0)$ is played on $]\hat{p}_{\max}, 1]$ and the transition $(0, 1)$—$(1, 1)$—$(1, 0)$ occurs at $\hat{p}_{\max}$. If $\hat{p}_{\min} < \hat{p}_{\max}$, finally, any two "adjacent" transitions $(0, 1)$—$(1, 1)$—$(1, 0)$ must be separated by one transition $(1, 0)$—$(0, 1)$—$(0, 1)$ or $(1, 0)$—$(1, 0)$—$(0, 1)$. ∎

Proof of Proposition 9

When stakes are high, the expected delay will be maximal for $p_0 = 1 - \overline{p}$ because this yields the largest possible interval of beliefs on which equilibrium play differs from the efficient solution, and at the same time minimizes the expected time until uncertainty is resolved in the efficient solution. At the belief $1 - \overline{p}$, the planner will play the good risky arm for sure until it produces a breakthrough; the corresponding expected time to breakthrough is $\hat{t} = \frac{1}{\lambda}$. The expected equilibrium time to breakthrough is $\tilde{t} = (1 - \overline{p})\hat{t} + \overline{p}(\delta + \hat{t}) = \hat{t} + \overline{p}\delta$, where δ is the time needed to slide from $1 - \overline{p}$ to $1 - p^m$, conditional on risky arm 2 being good. Bayes' rule for the action profile $(1, 0)$ implies $\delta = \frac{1}{\lambda}\left[\ln\frac{1-\overline{p}}{\overline{p}} - \ln\frac{1-p^m}{p^m}\right]$. Writing $x = \frac{g}{s}$, we therefore have

$$\frac{\tilde{t} - \hat{t}}{\hat{t}} = \overline{p}\left[\ln\frac{1-\overline{p}}{\overline{p}} - \ln\frac{1-p^m}{p^m}\right] = \frac{r + \lambda}{(r + \lambda)x + \lambda}\ln\left(1 + \frac{\lambda}{r + \lambda}\frac{1}{x - 1}\right).$$

As this decreases in x, an upper bound on the relative delay for high stakes is obtained by setting $x = 2$, so that

$$\frac{\tilde{t} - \hat{t}}{\hat{t}} \leq \frac{r + \lambda}{2r + 3\lambda}\ln\left(1 + \frac{\lambda}{r + \lambda}\right) \leq \frac{\lambda}{2r + 3\lambda} < \frac{1}{3}$$

by the fact that $\ln(1 + y) \leq y$ for all $y \geq 0$.

Turning to intermediate stakes, we may assume that $p^m < 1 - \overline{p}$, for otherwise there exists an equilibrium that achieves the efficient outcome (see the discussion leading up to Proposition 10 in Section 4.5). We now calculate the delay that arises for $p_0 = 1 - \overline{p}$ in the equilibrium in cutoff strategies defined by $\hat{p} = p^m$, that is, the worst possible delay in the best possible equilibrium. Proceeding as above, we find

$$\frac{\tilde{t} - \hat{t}}{\hat{t}} = \overline{p}\left[\ln\frac{1-\overline{p}}{\overline{p}} - \ln\frac{p^m}{1-p^m}\right] = \frac{r + \lambda}{(r + \lambda)x + \lambda}\ln\left(x - \frac{r}{r + \lambda}\right).$$

Using the fact that $\ln z \leq z - 1$ for all $z > 0$, we obtain

$$\frac{\tilde{t} - \hat{t}}{\hat{t}} \leq \frac{r + \lambda}{(r + \lambda)x + \lambda}\left(x - \frac{2r + \lambda}{r + \lambda}\right).$$

As the right-hand side increases in x, and $x < 2$ for intermediate stakes, this yields the same upper bound as for high stakes. ∎

Proof of Proposition 10

For high stakes, the players' average equilibrium payoff function u is strictly below the planner's value function on $]1 - p^m, 1[$. To the right of $1 - \overline{p}$, the two functions differ only with respect to the constant that premultiplies the solution $u_0(p)$ of the homogenous ODE for the action profile $(1, 0)$; in particular, both functions and their difference are monotonic there. The minimum of the average payoff is therefore attained at some belief $\breve{p}$ strictly in between $1 - p^m$ and $1 - \bar{p}$, and the quotient $(u(\breve{p}) - s)/(u_{11} - s)$ is the minimum over all beliefs of our relative welfare measure.

For $p \geq 1 - p^m$, we have $u(p) = (s + pg)/2 + Cu_0(p)$ with some positive constant C, and so

$$u'(\breve{p}) = \frac{g}{2} - C\frac{\mu + \breve{p}}{\breve{p}(1 - \breve{p})}\, u_0(\breve{p}) = 0,$$

where $\mu = \frac{r}{\lambda}$. Solving for $Cu_0(\breve{p})$, we obtain

$$u(\breve{p}) = \frac{s + \breve{p}g}{2} + \frac{\breve{p}(1 - \breve{p})}{\mu + \breve{p}}\frac{g}{2},$$

which is easily seen to be increasing in $\breve{p}$. As $\breve{p} > 1 - p^m$, we thus have

$$u(\breve{p}) > \frac{s + (1 - p^m)g}{2} + \frac{p^m(1 - p^m)}{\mu + 1 - p^m}\frac{g}{2} = \frac{g}{2} + \frac{(g - s)s}{2[(\mu + 1)g - s]}$$

and, with $x = \frac{g}{s} \geq 2$ denoting the stakes involved,

$$\frac{u(\breve{p}) - s}{u_{11} - s} > \frac{x + \frac{x-1}{(\mu+1)x-1} - 2}{x + \frac{1}{\mu+1} - 2} = 1 - \frac{\mu}{(\mu + 1)^2(x - 1)^2 - \mu^2} \geq 1 - \frac{\mu}{(\mu + 1)^2 - \mu^2}\,.$$

The last term on the right-hand side decreases in μ and approaches the limit $\frac{1}{2}$ as $\mu \to \infty$.

For intermediate stakes, we only need to cover the case where $p^m < 1 - \overline{p}$, so that the efficient outcome cannot be achieved for initial beliefs below $1 - p^m$ or above p^m. By symmetry, it is enough to consider the latter scenario. Given a prior above p^m, the players' average payoff function u in the MPE in cutoff strategies defined by $\hat{p} = p^m$ is strictly below the planner's value function on $]p^m, 1[$. Arguing as above and exploiting the fact that $\breve{p} > p^m$, we now find

$$u(\breve{p}) > \frac{s + p^m g}{2} + \frac{p^m(1 - p^m)}{\mu + p^m}\frac{g}{2} = s + \frac{(g - s)s}{2(\mu g + s)}$$

and, writing $x = \frac{g}{s}$ again,

$$\frac{u(\breve{p}) - s}{u_{11} - s} > \frac{\frac{x-1}{\mu x+1}}{x + \frac{1}{\mu+1} - 2}\,.$$

As $p^m < 1 - \overline{p}$, we have $\mu x + 1 < (\mu + 1)(x - 1)x$, and so

$$\frac{u(\breve{p}) - s}{u_{11} - s} > \frac{1}{[(\mu + 1)(x - 2) + 1]\, x}\,.$$

The right-hand side exceeds $\frac{1}{2}$ since $\frac{2\mu+1}{\mu+1} < x < 2$, and hence $0 < (\mu + 1)(x - 2) + 1 < 1$, for intermediate stakes. ∎

Proof of Proposition 11

Low stakes. Let $\frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$, so that $p^* \geq \frac{1}{2}$. If players behave as described in the main text, their payoff functions coincide with the respective single-agent value functions, and either player is trivially playing a best response.

Intermediate stakes. Let $\frac{2r+\lambda}{r+\lambda} < \frac{g}{s} \leq 2$, so that $p^* < \frac{1}{2} \leq p^m$. It suffices to construct the players' payoff functions and verify the mutual best-response property on the set of beliefs where $p_1 \geq p_2$. At all such beliefs with $p_1 \leq p^*$, we trivially have $u_1 = u_2 = s$, and both players are clearly playing a best response there.

Starting from a prior belief in the interior of the triangle with corners $(p^*, 0)$, (p^*, p^*) and $(1, 0)$, the action profile $(1, 0)$ makes posterior beliefs move up a ray $p_2 = x\,(1 - p_1)$ with $x \leq \frac{p^*}{1-p^*}$ until either player 1 experiences a breakthrough or all experimentation stops at $p_1 = p^*$. So player 2 will never use his risky arm and earns a sure payoff of s; as $p_2 < p^*$, he is playing a best response. Player 1 achieves a payoff equal to the single-agent optimum, hence is playing a best response as well.

For $p \in \,]p^*, \frac{1}{2}]$, we have

$$u_1(p,p) = u_2(p,p) = p\left[g + \frac{\lambda}{r+\lambda}s\right] + C^* u_0(2p),$$

where

$$C^* u_0(2p^*) = s - p^*\left[g + \frac{\lambda}{r+\lambda}s\right] = \frac{\lambda}{r}\,p^*\left[g - \frac{2r+\lambda}{r+\lambda}s\right] > 0.$$

This implies that the above payoff is a strictly convex function of p. As p tends to p^* from above, moreover, this payoff reaches the level s with a slope of zero. As a consequence, it is increasing in p on $]p^*, \frac{1}{2}]$ and exceeds s there.

For $p_1 > \max\{p^*, p_2\}$, we write $p_2 = x\,(1 - p_1)$ with $x > \frac{p^*}{1-p^*}$. We recall the general form of the players' payoff functions from Appendix A.3 and determine $f^i_{10}(x)$ by value matching along the diagonal line segment where the action profile $(1, 1)$ is played. For given x, the corresponding point on this line segment is $(\frac{x}{x+1}, \frac{x}{x+1})$, and the players' common payoff at this point is

$$\frac{x}{x+1}\left[g + \frac{\lambda}{r+\lambda}s\right] + C^* u_0\!\left(\frac{2x}{x+1}\right).$$

Equating this with player 1's payoff from the action profile $(1, 0)$ at the belief $(\frac{x}{x+1}, \frac{x}{x+1})$,

$$\frac{x}{x+1}\,g + f^1_{10}(x) u_0\!\left(\frac{x}{x+1}\right),$$

yields

$$f^1_{10}(x) = \left\{\frac{x}{x+1}\frac{\lambda}{r+\lambda}\,s + C^* u_0\!\left(\frac{2x}{x+1}\right)\right\}\Big/\, u_0\!\left(\frac{x}{x+1}\right) = \frac{\lambda}{r+\lambda}\,s\,x^{\frac{r+\lambda}{\lambda}} + C^* 2^{-\frac{r}{\lambda}}\,(1-x)^{\frac{r+\lambda}{\lambda}}\ .$$

For player 2, we find

$$f^2_{10}(x) = \left(x\left[g - \frac{r}{r+\lambda}\,s\right] - s\right) x^{\frac{r}{\lambda}} + C^* 2^{-\frac{r}{\lambda}}\,(1-x)^{\frac{r+\lambda}{\lambda}}\ .$$

To verify that player 1 is playing a best response when $p_1 > \max\{p^*, p_2\}$, we note that $u_1 = s + b_1 - c_1$, so we only need to prove that $u_1 > s$. As $u_1 > s$ when $p_1 = p_2 > p^*$, it suffices to show that $p_1 g + f_{10}^1(x) u_0(p_1)$ is increasing in p_1 for $p_1 > \frac{x}{x+1}$. By convexity of u_0, it is enough to show that $g + f_{10}^1(x) u_0'(p_1) \geq 0$ at $p_1 = \frac{x}{x+1}$ or, equivalently,

$$\left\{ \frac{\lambda}{r+\lambda}\, s + C^* \, \frac{x+1}{x}\, u_0\!\left(\frac{2x}{x+1}\right) \right\} \left(\frac{r}{\lambda} + \frac{r+\lambda}{\lambda}\, x \right) \leq g.$$

As $2p^* < \frac{2x}{x+1} \leq 1$ and $u_0(1) = 0$, convexity of u_0 implies

$$u_0\!\left(\frac{2x}{x+1}\right) \leq \frac{1 - \frac{2x}{x+1}}{1 - 2p^*}\, u_0(2p^*).$$

Using the definition of C^* and the fact that $1 - 2p^* = \frac{r+\lambda}{rs}\, p^* \left(g - \frac{2r+\lambda}{r+\lambda} s \right)$, we thus find that it is enough to show that

$$\left(\frac{r}{r+\lambda}\, \frac{1}{x} + 1 \right) s \leq g.$$

The left-hand side of this inequality is obviously decreasing in x, and is easily seen to assume the value g at $x = \frac{p^*}{1-p^*} = \frac{r}{r+\lambda}\, \frac{s}{g-s}$. The inequality thus holds for all x in the relevant range. This completes the proof that player 1 is playing a best response at all beliefs such that $p_1 > \max\{p^*, p_2\}$.

To establish that player 2 is also playing a best response at these beliefs, we can invoke Lemmas A.2 and A.3. The former implies that b_2 tends to c_2 as we approach the diagonal $p_2 = p_1$ from below, while part (2) of the latter implies that $b_2 < c_2$ below the diagonal. In fact, $p_2 = p_1$ implies $p_2 \leq \frac{(r+\lambda p_1)s}{(r+\lambda)g - \lambda s}$ for intermediate stakes.

Finally, given player 2's strategy, admissibility rules out any strategy k_1 for player 1 such that $k_1(p,p) < 1$ for some $p \in \,]p^*, \frac{1}{2}]$. To see this, we note that any such strategy would imply $\dot{p_2} = \lambda p[p k_1(p,p) + p - 1] < 0$ in the point (p,p). For $p_2 < p_1$, however, $k_2(p_1, p_2) = 0$ implies $\dot{p_2} = \lambda p_2 p_1 k_1(p_1, p_2) \geq 0$. Starting in (p,p), therefore, there is no solution to the law of motion for beliefs unless $k_1(p,p) = 1$. The symmetric argument applies to player 2.

High stakes. Let $\frac{g}{s} > 2$, so that $p^m < \frac{1}{2}$. At all beliefs such that $p_2 \leq p_1 \leq \tilde{p}$ or $p_2 \leq \frac{\tilde{p}}{1-\tilde{p}}\, p_2 \leq \tilde{p}$, the players' actions and payoffs coincide with those in the MPE for intermediate stakes, so both players are playing a best response there.

For $p_1 > \tilde{p}$ and $p_2 > \frac{\tilde{p}}{1-\tilde{p}}\, p_1$, Lemma A.2 implies that b_2 tends to c_2 as we approach player 2's switching boundary, so Lemma A.3 implies that $b_2 < c_2$ whenever player 1 alone is playing risky, and $b_2 > c_2$ whenever both players play risky. By symmetry, this also means that $b_1 > c_1$ whenever both players play risky. The verification of the best-response property is thus complete if we can show that player 1 plays a best response when $p_1 > \tilde{p}$ and $\frac{\tilde{p}}{1-\tilde{p}}\, p_1 < p_2 < \frac{(r+\lambda p_1)s}{(r+\lambda)g - s}$. As $u_1 = s + b_1 - c_1$ at these beliefs, we only need to show that $u_1 > s$. This step is similar to the intermediate-stakes case and therefore omitted. ∎

Chapter 2: Strategic Learning in Teams[*]

Nicolas Klein[†]

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 with continuous value functions 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.

[*] 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, 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 SFB/TR 15 is gratefully acknowledged for 2009; for 2010, the author thanks the National Research Fund, Luxembourg, for its support of this project.

[†] 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 involving purely informational externalities is provided by the literature on strategic experimentation with bandits.[1]

As information is a public good, one's first intuition may be that, on account of free-riding, there will always be inefficiently little experimentation in equilibrium. Indeed, the previous literature on strategic experimentation with bandits shows that, with positively correlated two-armed bandits,[2] there never exists an efficient equilibrium; with negatively correlated bandits,[3] there exists an efficient equilibrium if, and only if, stakes are below a certain threshold. 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. Surprisingly, though, 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. The extension of players' action sets to include *how* they go about investigating a given hypothesis thus matters greatly for the results.

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

[1] See e.g. Chapter 1, of this dissertation, Bolton & Harris (1999, 2000), Keller, Rady, Cripps (2005), Keller & Rady (2010); for an overview of the bandit literature, consult Bergemann & Välimäki (2008).

[2] See the papers by Bolton & Harris (1999, 2000), Keller, Rady, Cripps (2005), Keller & Rady (2010).

[3] See Chapter 1.

[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 (2010) have then introduced negative correlation into the strategic model, see Chapter 1 of this dissertation.

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 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.

In the game with positively correlated two-armed bandits, Keller, Rady, Cripps (2005) find two dimensions of inefficiency in any equilibrium: 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, we find in Chapter 1 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 with continuous value functions for exactly the same parameter range as is the case in Chapter 1. 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 Chapter 1 one player will be stuck with the losing project.

There are two distinct effects at play that make players in the three-armed setup perform better than in the two-armed model. 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 approach. 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,[5] the *higher* planner's solution can be achieved in equilibrium with three arms, if the stakes are high enough. The reason for this is that with the stakes high enough, the safe option of doing essentially nothing becomes

[5] As already mentioned, the negatively correlated case with low stakes provides a notable exception, cf. Chapter 1.

so unattractive that it can be completely disregarded. But then, since there are no payoff rivalries or switching costs in my model, given an opponent behaves in the same fashion, a player is willing to go for the project that looks momentarily more promising, which is exactly what efficiency requires. In Chapter 1, by contrast, players will always choose the safe option if their assigned task looks sufficiently hopeless.

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 is 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,[6] and hence experimentation will 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 planner's and the single agent's solutions, as well as the equilibria I construct, all exhibit continuous value functions. While I do not provide a full characterization of the equilibrium set, I show in section 4 that learning will be complete in *any* equilibrium with continuous value functions, provided the stakes at play exceed a certain threshold.

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).[7] 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, 2010, Chapter 1). These previous papers analyzed variants of the two-armed positively correlated model, with the exception of Chapter 1, 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 &

[6] The encouragement effect was first identified in the Brownian motion model of Bolton & Harris (1999). 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 in Chapter 1, 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.

[7] Bandit models have been analyzed as early as the 1950s; see e.g. Bradt, Johnson, Karlin (1956).

Hörner (2010) 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 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.

Manso (2010) analyzes the case of a *single* worker, who can either shirk, or take risks 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. In a related fully dynamic continuous-time model, Chapter 3 also shows that incentives are optimally provided through continuation values after breakthroughs. I furthermore show there that there is no agency loss stemming from the delegation of the project to an agent. Chatterjee & Evans (2004) 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.

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) analyzes the effects of correlation between the arms of a two-armed bandit operated by a single decision maker.

[8] Bonatti & Hörner's (2010) 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).

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 Chapter 1, 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 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. 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. Thus, 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.[11] A Markov strategy for player i is any piecewise continuous function $(k_{i,A}, k_{i,B}) : [0,1] \to \{(a,b) \in [0,1]^2 : a + b \leq 1\}$, $p_t \mapsto (k_{i,A}, k_{i,B})(p_t)$, 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. 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.

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.

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

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 Chapter 1, 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, 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: $(1,0,1)$, $(0,0,1)$, $(-1,0,1)$, $(-1,0,0)$, $(-1,0,-1)$, $(-1,1,1)$, $(-1,-1,1)$, $(1,0,0)$, $(0,1,1)$, $(0,0,-1)$, $(-1,-1,0)$, $(1,0,-1)$.[12]

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

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

for each $i \in \{1, 2\}$, 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 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:

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

[12] The list of admissible changes in sign differs somewhat from the corresponding list in Chapter 1: $(1,0,1)$ and $(-1,0,-1)$ are ruled out in Chapter 1 by the requirement imposed there that the region on which a player uses his risky arm be the union of *non-degenerate* intervals. Furthermore, $(0,0,0)$ does not correspond to a *change in sign* according to my definition here; however, it can correspond to a *transition* in Chapter 1.

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

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

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; $B_A(p,u) := \frac{\lambda}{r}p[g - u(p) - (1-p)u'(p)]$ 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 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.

[13] By standard arguments, if a continuously differentiable function solves the Bellman equation, it is the value function; see also Chapter 1.

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

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 $p_1^* \equiv \frac{rs}{(r+\lambda)g-\lambda s}$ denotes the optimal single-agent cutoff in the standard two-armed problem with one safe and one risky arm A, and $1 - p_1^*$ is the corresponding threshold for the two-armed problem with one safe arm and one risky arm 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*

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

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

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.

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:

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

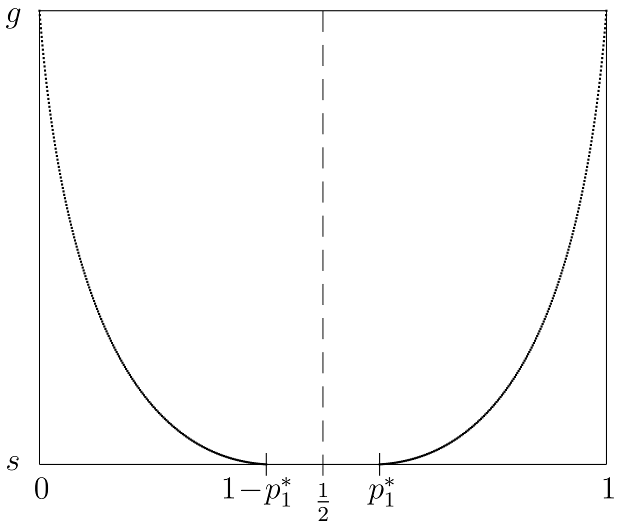

Figure 1: The single-agent value function for $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$.

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 $\frac{g}{s} = \frac{2r+\lambda}{r+\lambda}$.

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. As the planner does not care about the distribution of surplus,

[16] cf. also Presman (1990).

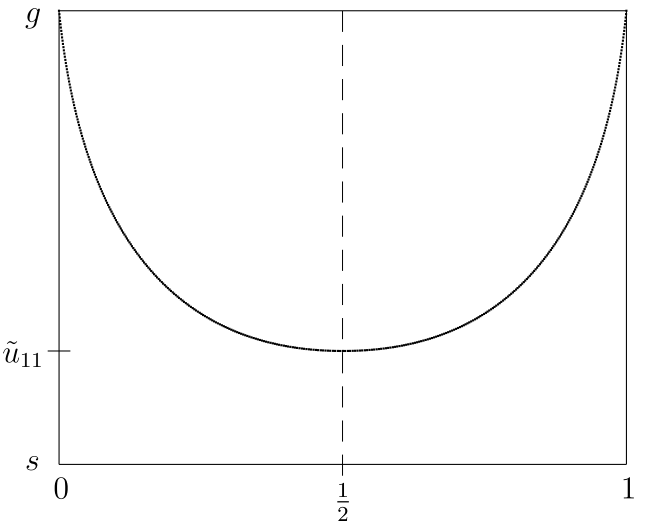

Figure 2: The single agent's value function for $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$.

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 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

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

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*

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

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 Chapter 1 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*

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

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

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, 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.

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 found out. 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, we find in Chapter 1 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 in which players' value functions are continuous. 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]

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

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

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

Lemma 4.1 (Lower Bound on Equilibrium Payoffs) *Let $u \in C^0$ be a player's 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 $(k_{1,A}, k_{1,B})(p) = (k_{2,A}, k_{2,B})(p) = (0, 0)$ 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 with continuous value functions.*

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.[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 there cannot be an open interval on which a player's value is less than s, it will be s almost everywhere on $[1 - p_2^*, p_2^*]$. 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 almost everywhere in $[1 - p_2^*, p_2^*]$, 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.*

[19] cf. Chapter 1

5 Strategic Problem

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

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

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.

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.

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

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

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

[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.

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, 2010). Introducing negative correlation into the strategic experimentation literature, in Chapter 1, we 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, in Chapter 1, the 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.

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 Chapter 1 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 Chapter 1, 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 Chapter 1. Indeed, with the stakes high enough, free-riding incentives can be overcome completely in non-cooperative equilibrium.

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, and exhibiting continuous value functions, for $\frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$, we are now in a position to strengthen our result on complete learning:

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

In Chapter 1, we 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 with continuous value functions 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, 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 $p^\dagger \equiv p_1^\dagger$ 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 with continuous value functions 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 with continuous value functions. If stakes are intermediate in size, all equilibria are inefficient, though they involve complete learning (provided there are no discontinuities in the value functions), 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 (2010), or to the assumption of a diffusion process, as in Bolton & Harris (1999, 2000), or even a general Lévy process in the vein of Cohen & Solan (2009), is an interesting question for future research.

Furthermore, while Chapter 1 explores 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 conditions under which there is excessive, or insufficient, specialization in equilibrium. 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

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

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 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 $\overline{u}_{11} \geq s$, 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

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 holds by continuity, because on any open interval between any two points of discontinuity in his opponent's strategy,[22], a player can always guarantee himself

[22] Note that, on account of my definition of strategies, there can be only finitely many such points of discontinuity.

a payoff of at least s by playing $(0,0)$.

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, by continuity, implies the 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. At $p = \frac{1}{2}$, admissibility uniquely pins down a player's response to the other player's action. 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

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

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. 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 $[1-p^\dagger,\frac{1}{2}[$. ∎

Chapter 3: The Importance of Being Honest[*]

Nicolas Klein[†]

Abstract

I analyze the case of a principal who wants to give an agent proper incentives to investigate a hypothesis which can be either true or false. The agent can shirk, thus never proving the hypothesis, or he can avail himself of a known technology to manipulate the data. If the hypothesis is false, a proper investigation never yields a success. I show that if, in the case the hypothesis is true, a proper investigation yields successes with a higher intensity than manipulation would, the option of faking a success creates no distortions. In the opposite case, honest investigation is impossible to implement.

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

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

[*] I thank Johannes Hörner and Sven Rady for their advice, patience, and encouragement, as well as Dirk Bergemann, Tri-Vi Dang, Federico Echenique, Lucas Maestri, Thomas Mariotti, Andy Skrzypacz, Juuso Välimäki, Tom Wiseman, and the seminar audience in Berlin for helpful comments and discussions. I am especially grateful to the Cowles Foundation for Research in Economics at Yale University for an extended stay during which the idea for this paper took shape. Financial support from the National Research Fund of Luxembourg and the Deutsche Forschungsgemeinschaft through SFB TR-15 is gratefully acknowledged.

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

1 Introduction

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

As is well known from the principal-agent literature, when an agent's actions cannot easily be monitored, his pay must be made contingent on his performance, so that he have proper incentives to exert effort. Thus, the scientist will get paid a substantial bonus, or will be afforded high peer recognition if, and only if, he proves his hypothesis. While this may well provide him with the necessary incentives to work, unfortunately, it might also give him incentives to fabricate, or manipulate, his data, in order to make it appear as though his hypothesis was proved. In a setting involving Bayesian learning on the agent's part, my model investigates how optimally to achieve the dual objective of providing the agent with the right incentives to work, while also making sure that he not be tempted to engage in manipulations and trickery, even if said manipulations were not verifiable in a court of law, or even completely unobservable.[1]

Or think of the owners of a firm, who might be interested in its long-term prospects. Depending on the incentive scheme they offer management, they may endogenously create an incentive for the latter to induce a short-term bubble in the price of the firm's stock.[2] Alternatively, one could interpret my model as a model of technology adoption: An agent is

[1] A case in point, where a scientist's untoward behavior was eventually discovered, might be provided by (in)famous South Korean stem cell researcher Hwang Woo-Suk. Mr. Hwang was considered one of the world's foremost authorities in the field of stem-cell research, and was even designated his country's first "top scientist" by the South Korean Government. He purported to have succeeded in creating patient-matched stem cells, which would have been a major breakthrough that had raised high hopes for new cures for hitherto hard-to-treat diseases, and that I am told had been the source of considerable pride in South Korea. Yet, a university panel found that "the laboratory data for 11 stem cell lines that were reported in the 2005 paper were all data made using two stem cell lines in total," forcing Mr. Hwang to resign in disgrace, and causing a major shock to people in South Korea and throughout the scientific community. I am indebted to Tri-Vi Dang for alerting me to this story; see e.g. the report by the Associated Press from December, 23, 2005.

[2] Bolton, Scheinkman, Xiong (2005) analyze a setting where even owners might have an interest in creating such short-term bubbles, and thus wittingly give incentives to this effect.

hired expressly to test some new production method, or some new way of doing business, yet his boss cannot monitor whether the successes he observes are really due to the new method, or whether the agent has surreptitiously availed himself of an old established method to produce the observed results.

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

I show that if even the investigation of a correct hypothesis yields breakthroughs at a lower frequency than manipulation, honesty is not implementable at all. If, however, investigating a correct hypothesis yields breakthroughs at a higher intensity than manipulation, I characterize the optimal incentive schemes making sure that the agent is always honest up to the first breakthrough at least.

While actually investigating the hypothesis, the agent increasingly grows pessimistic about the thesis being true as long as no breakthrough arrives. At the first breakthrough, though, all uncertainty is resolved, and the agent will know for sure that the hypothesis is true. Thus, depending on the incentive scheme, this learning aspect might give the agent additional incentives for investigating the hypothesis. The principal himself has no learning motive as he is only interested in the *first* breakthrough achieved on arm 1; however, when designing the incentive scheme, it will be one of his goals to make information valuable to the agent as a way of providing incentives.

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

To provide adequate incentives in the cheapest way possible, the principal will endeavor to give the lowest possible value to a dishonest agent, given the continuation value he has promised the on-path agent. While paying only for the $(m+1)$-st breakthrough ensures that off-path agents will not continue to cheat, they will nevertheless continue to update their beliefs after their first success, and might be tempted to switch to shirking once they have grown too pessimistic about the hypothesis, a possibility that, as is well known from the literature on strategic experimentation with bandits, gives them a positive option value. When designing his incentive scheme, it will be one of the principal's goals to reduce this additional option value. As a matter of fact, I will show that it is possible for the principal to structure incentives in such a way that cheating will be dominated even by shirking. Thus, in an optimal scheme, the agent only needs to be compensated for his outside option of shirking.

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

2 Related Literature

Holmström & Milgrom (1991) analyze a case where, not unlike in my model, the agent performs several tasks, some of which may be undesirable from the principal's point of view. The principal may be able to monitor certain activities more accurately than others. They show that in the limiting case with two activities where one activity cannot be monitored at all, incentives will only be given for the activity which can in fact be monitored; if the activities are substitutes (complements) in the agent's private cost function, incentives are more muted (steeper) than in the single task case. While their model could be extended to a dynamic model where the agent controls the drift rate of a Brownian Motion signal,[3] the learning motive I introduce fundamentally changes the basic trade-offs involved. Indeed, in my model, the optimal mechanism extensively leverages the fact that only an honest agent will have had a discontinuous jump in his beliefs.

Bergemann & Hege (1998, 2005), as well as Hörner & Samuelson (2009) examine a venture capitalist's provision of funds for an investment project of initially uncertain quality; the project is managed by an entrepreneur, who might divert the funds for his private ends. The investor cannot observe the entrepreneur's allocation of the funds, so that, off-path, the

[3] See Holmström & Milgrom (1987).

entrepreneur may accumulate some private information about the quality of the project. If the project is good, it yields a success with a probability that is proportional to the amount of funds invested in it; if it is bad, it never yields a success. While Bergemann & Hege (2005) and Hörner & Samuelson (2009) analyze the game without commitment, Bergemann & Hege (1998) investigate the problem under full commitment. These papers differ from my model chiefly in that there is no way for the entrepreneur to "fake" a success; any success that is publicly observed will have been achieved by honest means alone.

Gerardi & Maestri (2008) investigate the case of a principal who, in order to find out about the binary state of the world, has to employ an agent. The agent can decide to incur private costs to exert effort to acquire an informative binary signal, one realization of which is only possible in the good state. As for the principal, he can monitor neither the agent's effort choice nor the realization of the signal. The game ends as soon as the agent announces that he has had conclusive evidence in favor of the good state. They show that the agent needs to be left an information rent because of both the Moral Hazard and the Adverse Selection problems. In my model, by contrast, the game does not end after the first breakthrough; much to the contrary, I show that in my model, in order to give optimal incentives, it is absolutely vital that they be provided via the continuation game that follows the first breakthrough rather than via an immediate transfer.

One paper that is close in spirit to mine is Manso (2010), who analyzes a two-period model where an agent can either shirk, try to produce in some established manner with a known success probability, or experiment with a risky alternative. He shows that, in order to induce experimentation, the principal will optimally not pay for a success in the first period, and might even pay for early failure,[4] while a success in the second period is always rewarded. My continuous-time investigation confirms Manso's (2010) central intuition that it is better to give incentives through later rewards; furthermore, the richer action and signal spaces in my fully-fledged dynamic model yield additional insights into the structure of the optimal incentive scheme. Moreover, the dynamic structure allows me to analyze the principal's optimal stopping time.

De Marzo & Sannikov (2008) also incorporate private learning on the agent's part into their model, where current output depends both on the firm's inherent profitability and on the agent's effort, which is unobservable to the principal. Thus, off-path, the agent's private belief about the firm's productivity will differ from the public belief. Specifically, if the agent

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

withholds effort, this depresses the drift rate of the firm's Brownian motion cash flow. They show that the firm will optimally accumulate cash as fast as it can until it reaches some target level, after which it starts paying out dividends; the firm is liquidated as soon as it runs out of cash. De Marzo & Sannikov (2008) show that one optimal way of providing incentives is to give the agent an equity stake in the firm, which is rescindable in the case of liquidation, and that liquidation decisions are efficient, agency problems notwithstanding.

To capture the learning aspect of the agent's problem, I model it as a bandit problem.[5] Bandit problems have been used in economics to study the trade-off between experimentation and exploitation since Rothschild's (1974) discrete-time single-agent model. The single-agent two-armed exponential model, a variant of which I am using, has first been analyzed by Presman (1990). Strategic interaction among several agents has been analyzed in the models by Bolton & Harris (1999, 2000), Keller, Rady, Cripps (2005), Keller & Rady (2010), who all investigate the case of perfect positive correlation between players' two-armed bandit machines, as well as in Chapter 1, where the cases of perfect, as well as imperfect, negative correlation are investigated. Chapter 2 analyzes the case where bandits have three arms, with the two risky ones being perfectly negatively correlated. While the afore-mentioned papers all assumed that players' actions, as well as the outcomes of their actions, were perfectly publicly observable, Rosenberg, Solan, Vieille (2007), as well as Murto & Välimäki (2009), analyze the case where actions are observable, while outcomes are not. Bonatti & Hörner (2010) analyze the case where actions are not observable, while outcomes are. Bergemann & Välimäki (1996, 2000) consider strategic experimentation in buyer-seller interactions. My contribution to this literature is to introduce the question of optimal incentive provision into a fully-fledged dynamic bandit model.

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

3 The Model

There is one principal and one agent. The agent operates a bandit machine with three arms, i.e. one safe arm yielding the agent a private benefit flow of s, one that is known to yield breakthroughs according to $Po(\lambda_0)$ (arm 0), and arm 1, which either yields breakthroughs

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

according to $Po(\lambda_1)$ (if the time-invariant state of the world $\theta = 1$, which is the case with initial probability $p_0 \in]0,1[$) or never yields a breakthrough (if the state is $\theta = 0$). It is commonly known that $\lambda_1, \lambda_0 > 0$. The principal only observes if, and at what time, there has been a breakthrough; he does not observe on which arm the breakthrough has been achieved. The agent in addition observes on which arms the breakthroughs have occurred. The principal and the agent share a common discount rate r.

The principal, only being interested in the first breakthrough *achieved on arm 1*, chooses an end date $\check{T}(t) \in [t, \overline{T}]$ (where $\overline{T} \in]T, \infty[$ is arbitrary), in case the first breakthrough occurs at time t. Conditional on there having been no breakthrough, the game ends at time $T < \infty$. I the first half of this paper, I take T to be exogenously given, and assume the principal always wants to incentivize the agent to use arm 1 up until the first breakthrough at least. In the second half, the principal optimally chooses the end date T.[6] There, I shall assume that the first breakthrough achieved on arm 1 at time t gives the principal a payoff of $e^{-rt}\Pi$.

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

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

Formally, a strategy for the agent is a process $\mathbf{k} := \{(k_{0,t}, k_{1,t})\}_t$ which satisfies $(k_{0,t}, k_{1,t}) \in \{(a,b) \in \mathbb{R}_+ : a + b \le 1\}$ for all t, and is $\mathfrak{F}$-predictable, where $k_{i,t}$ $(i \in \{0,1\})$ denotes the fraction of the agent's resource that he devotes to arm i at instant t. The agent's strategy space, which I denote by $\mathcal{U}$, is given by all the processes $\mathbf{k}$ satisfying these requirements. I denote the set of abridged strategies $\mathbf{k}_T$ prescribing the agent's actions *before the first breakthrough* by $\mathcal{U}_T$.

A *wage scheme* offered by the principal is a non-negative, non-decreasing process $\{\mathcal{W}_t\}_{0 \le t \le \overline{T}}$ which is $\mathfrak{F}^N$-adapted, where $\mathcal{W}_t$ denotes the discounted time 0 value of the cumulated pay-

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

ments the principal has consciously made to the agent up to, and including, time t. I assume the agent is protected by limited liability; hence $\{\mathcal{W}_t\}_{0\leq t\leq \overline{T}}$ is non-negative and non-decreasing.[7] I furthermore assume that the principal has full commitment power, i.e. he commits to a wage scheme $\{\mathcal{W}_t\}_{0\leq t\leq \overline{T}}$, as well as a schedule of end dates $\{\check{T}(t)\}_{t\in[0,T]}$, at the outset of the game.

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

It is the principal's goal to induce the agent to use arm 1 at least up to the first breakthrough, and to do so in the most cost-efficient manner possible. Thus, I shall denote the set of *full-experimentation strategies* by $\mathcal{K} := \{\mathbf{k} \in \mathcal{U} : \forall\, t \in [0,T] : \, N_t = 0 \Rightarrow k_{1,t} = 1\}$. Clearly, as the principal wants to minimize wage payments subject to implementing a full-experimentation strategy, it is never a good idea for him to pay the agent in the absence of a breakthrough; moreover, since the principal is only interested in the first breakthrough, the notation can be simplified somewhat. Let $\{\mathcal{W}_t\}_{0\leq t\leq \overline{T}}$ be the principal's wage scheme, and t the time of the first breakthrough: In the rest of the paper, I shall write $h_t := e^{rt}\left(\mathcal{W}_t - \lim_{\tau\uparrow t}\mathcal{W}_\tau\right)$ for the instantaneous lump sum the principal pays the agent as a reward for his first breakthrough. By w_t I denote the expected continuation value of an agent who has achieved his first breakthrough on arm 1 at time t, given he will behave optimally in the future; formally,

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

i.e. the expectation conditions on the agent's knowledge that the first breakthrough has been achieved on arm 1 at time t. The corresponding expected continuation payoff of an off-path

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

agent, who achieves his first breakthrough on arm 0 at time t, I denote by ω_t; formally,

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

The state of the world is uncertain; clearly, whenever the agent uses arm 1, he gets new information about its quality; this *learning* is captured in the evolution of his (private) belief $\hat{p}_t$ that arm 1 is good. Formally, $\hat{p}_t := E\left[\theta|\mathfrak{F}_t, \{(k_{0,\tau},k_{1,\tau})\}_{0\leq\tau<t}\right]$. On the equilibrium path, the principal will correctly anticipate $\hat{p}_t$; formally, $p_t = \hat{p}_t$, where p_t is defined by $p_t := E\left[\hat{p}_t|\mathfrak{F}_t^N, \mathbf{k}\in\mathcal{K}\right]$.

The evolution of beliefs is easy to describe, since only a good arm 1 can ever yield a breakthrough. By Bayes' rule,

$$\hat{p}_t = \frac{p_0 e^{-\lambda_1\int_0^t k_{1,\tau}\,d\tau}}{p_0 e^{-\lambda_1\int_0^t k_{1,\tau}\,d\tau} + 1 - p_0},$$

and

$$\dot{\hat{p}}_t = -\lambda_1 k_{1,t}\hat{p}_t(1-\hat{p}_t)$$

prior to the first breakthrough. After the agent has achieved at least one breakthrough on arm 1, his belief will be $\hat{p}_t = 1$ forever thereafter.

As, in equilibrium, the agent will always operate arm 1 until the first breakthrough, it is clear that if on the equilibrium path $N_t \geq 1$, then $p_{t+\Delta} = 1$ for all $\Delta > 0$. If $N_t = 0$, Bayes' rule implies that

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

In the following, I shall write p_t whenever $p_t = \hat{p}_t$, even when analyzing the agent's optimization problem.

Now before the first breakthrough, given an arbitrary incentive scheme $\mathbf{g} := (h_t, w_t)_{0\leq t\leq T}$, the agent seeks to choose $\mathbf{k}_T \in \mathcal{U}_T$ so as to maximize

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

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

The following impossibility result is now immediate:

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

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

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

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

subject to $\mathbf{k}^*(\mathbf{g}) \cap \mathcal{K} \neq \emptyset$ and $p_t = \frac{p_0e^{-\lambda_1 t}}{p_0e^{-\lambda_1 t}+1-p_0}$.

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

4 Incentives After The Breakthrough

4.1 Introduction

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

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

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

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

4.2 Construction of An Optimal Continuation Scheme

My construction proceeds in several steps. First, the principal will only pay the agent for the m-th breakthrough after time t, where m is chosen large enough that any, even the most pessimistic of off-path agents will deem m breakthroughs more likely to occur on arm 1 than on arm 0. Then, for a given $\epsilon > 0$, I make sure that even the most pessimistic of off-path agents, whose belief $\hat{p}_t = p_t$, will not switch to playing safe with more than ϵ time left to go. This requires a certain minimum lump sum reward for the m-th breakthrough. Then, given this reward, the end date $\check{T}(t)$ is chosen appropriately so that the on-path agent exactly receives his promised continuation value of w_t in expectation.

Specifically, the agent is only paid a constant lump sum of $\overline{V}_0$ after his m-th breakthrough after time t, where m is chosen sufficiently high that even for the most pessimistic of all possible off-path agents, arm 1 dominate arm 0. As $\lambda_1 > \lambda_0$, such an m exists, as the following lemma shows:

Lemma 4.2 *There exists an integer m such that if the agent is only paid a lump sum reward $\overline{V}_0 > 0$ for the m-th breakthrough, arm 1 dominates arm 0 for any type of off-path agent whenever he still has m breakthroughs to go before collecting the lump sum reward.*

Proof: See appendix. ∎

Intuitively, the likelihood ratio of m breakthroughs being achieved on arm 1 vs. arm 0 on the time interval $(t, \check{T}(t)]$, $\hat{p}_t \left(\frac{\lambda_1}{\lambda_0}\right)^m e^{-(\lambda_1-\lambda_0)(\check{T}(t)-t)}$, is unbounded in m. The proof now shows that when m exceeds certain thresholds, it indeed never pays for the agent to use arm 0.

Now, given m, $\check{T}(t)$, and $\overline{V}_0$, I recursively define the auxiliary functions $V_i(.;\overline{V}_0) : [t,\check{T}(t)] \longrightarrow \mathbb{R}$ for $i = 1, \cdots, m$ according to

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

where $\mathcal{M}(\tilde{t})$ denotes the set of measurable functions $k_i : [\tilde{t},\check{T}(t)] \to [0,1]$, and I set $V_0(\tau;\overline{V}_0) := \overline{V}_0 + \frac{s}{r}\left(1 - e^{-r(\check{T}(t)-\tau)}\right)$. Thus, $V_i(\tilde{t};\overline{V}_0)$ denotes the agent's continuation value at time $\tilde{t}$ given the agent knows that $\theta = 1$ and that he has i breakthroughs to go before being able to collect the lump sum $\overline{V}_0$.

The following lemma notes that, once the agent knows that $\theta = 1$, a best response for him is given by a cutoff time t_i^* at which he switches to the safe arm given he has i breakthroughs to go. It also takes note of some useful properties of these functions V_i:

Lemma 4.3 *Let $\overline{V}_0 > \frac{s}{\lambda_1}$. A best response for the agent is given by a sequence of cutoff times $t_m^* \geq \cdots \geq t_2^* > t_1^* = \check{T}(t)$ (with all inequalities strict if $t_{m-1}^* > t$), such that he use arm 1 at all times $\tilde{t} \leq t_i^*$, and the safe arm at times $\tilde{t} > t_i^*$, when he still has i breakthroughs to go before collecting the lump sum $\overline{V}_0$. For $i = 2, \cdots, m$, there exists a constant C_i such that for $\overline{V}_0 > C_i$, the cutoff time t_i^* is strictly increasing in $\overline{V}_0$. The functions $V_i(.;\overline{V}_0)$ are of class C^1 and strictly decreasing; for $\tilde{t} < t_i^*$, $V_i(\tilde{t};\overline{V}_0)$ is strictly increasing in $\overline{V}_0$. Moreover, $\lim_{\overline{V}_0\to\infty} t_i^* = \check{T}(t)$, and $\lim_{\overline{V}_0\to\infty} V_i(\tilde{t};\overline{V}_0) = \infty$ for any $\tilde{t} \in [t,\check{T}(t))$. The functions V_i satisfy*

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

and $V_i(\tilde{t};\overline{V}_0) \leq V_{i-1}(\tilde{t};\overline{V}_0)$, with the inequality strict for $\tilde{t} < t_i^*$.

PROOF: See appendix. ▮

In our next lemma, I shall derive a sufficient condition on $V_{m-1}(\tilde{t};\overline{V}_0)$ for *any* type of off-path agent always to play risky up to some given time $\tilde{t} < \check{T}(t)$. This will then allow me, in a next step, to choose $\overline{V}_0$ in a way that makes sure that the sufficient condition holds at $\check{T}(t) - \epsilon$, in case $\check{T}(t) - t \geq \epsilon$.

Lemma 4.4 *If $V_{m-1}(\tilde{t};\overline{V}_0) \geq \frac{s(r+\lambda_1)}{p_{\underline{T}} r\lambda_1}$, any type of (off-path) agent who has not yet collected the lump sum $\overline{V}_0$ will use arm 1 a.e. on $(t,\tilde{t}]$.*

PROOF: See appendix. ▮

In what will follow, it will be crucial that my choice of $\overline{V}_0$ be independent of $\check{T}(t)$. To make this clear, I shall now use the transformation $\tilde{V}_i(\overline{\alpha};\overline{V_0})$, which is defined by

$$\tilde{V}_i(\overline{\alpha};\overline{V_0}) := \max_{\hat{\alpha}\in[0,\overline{\alpha}]} \int_0^{\hat{\alpha}} e^{-(r+\lambda_1)\tau}\lambda_1 \tilde{V}_{i-1}(\overline{\alpha}-\tau;\overline{V}_0)\,d\tau + \frac{s}{r}e^{-(r+\lambda_1)(\overline{\alpha}-\hat{\alpha})}\left(1-e^{-r\hat{\alpha}}\right),$$

with $\tilde{V}_0(\overline{\alpha};\overline{V_0}) = \overline{V}_0 + \frac{s}{r}\left(1-e^{-r\overline{\alpha}}\right)$. It is immediate to verify that $\tilde{V}_i(\overline{\alpha};\overline{V}_0) = V_i(\check{T}(t)-\overline{\alpha};\overline{V}_0)$ for *any* given $\check{T}(t)\in[t+\overline{\alpha},\overline{T}]$; in other terms, the value V_i at any point in time $\tilde{t}$ only depends on the remainder of time, $\overline{\alpha}=\check{T}(t)-\tilde{t}$.

Now, by Lemma 4.3, one can choose a reward $\overline{V}_0$ such that, for given $\epsilon>0$, $\tilde{V}_{m-1}(\epsilon;\overline{V}_0)\geq \frac{s(r+\lambda_1)}{p_{\overline{T}}r\lambda_1}$. By Lemma 4.4, this guarantees that *all* types of agents will play risky at the very least until there is only ϵ time left to go (or they have collected the prize).

Thus, through our choice of m, we have made sure that the agent will never use arm 0. Through our choice of $\overline{V}_0$, we can make sure that the agent will use arm 1 at least through time $\check{T}(t)-\epsilon$. After that, he may switch to the safe arm at a time that depends on his previous experience with arm 1. Since he will get to play the safe arm for a length of time of at most ϵ, the option value from being able to switch to the safe arm is bounded above by $\frac{s}{r}(1-e^{-r\epsilon})$.

As a last step, we now need to make sure that the on-path agent is indeed delivered an expected continuation value of w_t. In order to do so, I first define another auxiliary function $f:[t,\overline{T}]\times\mathbb{R}_+\longrightarrow\mathbb{R}$ by $f(\check{T}(t),\overline{V}_0)=V_m(t;\overline{V}_0;\check{T}(t))$, where, in a slight abuse of notation, I write $V_m(t;\overline{V}_0;\check{T}(t))$ for $V_m(t;\overline{V}_0)$ given the end date $\check{T}(t)$. Thus, $f(\check{T}(t),\overline{V}_0)$ maps the choice of the stopping time $\check{T}(t)$ into the on-path agent's time-t expected payoff, given the reward $\overline{V}_0$. The following lemma notes some properties of f:

Lemma 4.5 *$f(.,\overline{V}_0)$ is of class C^1 and strictly increasing with $f(t;\overline{V}_0)=0$.*

Proof: See appendix. ▌

Now, if $w_t\leq f(\overline{T},\overline{V}_0)$, Lemma 4.5 implies that we can choose $\check{T}(t)$ so that $w_t = f(\check{T}(t),\overline{V}_0)$. Otherwise, since $\lim_{\overline{V}_0\to\infty} f(\check{T}(t),\overline{V}_0)=\infty$ for any $\check{T}(t)\in(t,\overline{T}]$, we can choose a constant $\delta>0$ high enough so that $w_t\geq f(\overline{T},\overline{V}_0+\delta)$; then, by Lemma 4.5, we can find a $\check{T}(t)\in(t,\overline{T}]$ so that $w_t=f(\check{T}(t),\overline{V}_0+\delta)$.

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

$\xi \leq \check{T}(t) - t \leq \epsilon$, in which case his payoff is bounded above by $\frac{s}{r}(1 - e^{-r(\check{T}(t)-t)}) + \hat{p}_t w_t \leq \frac{s}{r}(1 - e^{-r\epsilon}) + \hat{p}_t w_t$.

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

5 Before the Breakthrough–Optimal Incentive Scheme

Whereas in the previous section, I have investigated how a principal would optimally deliver a given *continuation* value w_t, the purpose of this section is to understand to what extent the principal would optimally give incentives via continuation values w_t, as opposed to immediate rewards h_t, which are paid out right at the moment of the first breakthrough. I shall show in this section that, by Proposition 4.1, arm 0 can be made so unattractive that in any optimal scheme it is dominated by the safe arm. Thus, in order to induce the agent to use arm 1, he only needs to be compensated for his outside option of playing safe (Proposition 5.4), which pins dow the principal's wage costs (Corollary 5.5).

In order formally to analyze this question, we first have to consider the agent's best response to a given incentive scheme $(h_t, w_t)_{0 \leq t \leq T}$, in order to derive conditions for the agent to best reply by always using arm 1 until the first breakthrough. In a second step, we will then use these conditions as constraints in the principal's problem as he seeks to minimize his wage bill. While the literature on experimentation in bandits would typically use dynamic programming techniques, this would not be expedient here, as an agent's optimal strategy will depend not only on his current belief and the current incentives he is facing but also on the entire path of future incentives. To the extent we do not want to impose any *ex ante* monotonicity constraints on the incentive scheme, today's scheme need not be a perfect predictor for the future path of incentives; therefore, even a three-dimensional state variable (p_t, h_t, w_t) would be inadequate. Thus, I shall be using the Pontryagin approach of Optimal Control.

The Agent's Problem

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

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

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

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

Neglecting a constant factor, the Hamiltonian $\mathfrak{H}_t$ is now given by[8]

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

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

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

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

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

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

[8] In a slight abuse of notation, I now write ω_t as a function of x_t.

The Principal's Problem

Now, we turn to the principal's problem, who will take the agent's incentive constraint into account when designing his incentive scheme with a view toward implementing $k_{1,t} = 1$ for almost all $t \in [0, T]$; I shall refer to incentive schemes implementing this kind of full experimentation as *incentive compatible*. We note that $k_{1,t} = 1$ for all t implies $y_t = 1$ for all t. Thus, the principal's objective is to choose $(h_t, w_t)_{0 \le t \le T} \in [0, L]^2$ (for some L which I choose large enough) so as to minimize

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

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

Neglecting constant factors and using the fact that full experimentation implies that $x_t = x_0 + \lambda_1 t$, one can re-write the principal's objective in terms of the log-likelihood ratio as

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

This can be viewed as a variational problem with the co-state variables (μ_t, γ_t) from the agent's problem as the state variables. As we have seen, μ_t and γ_t evolve according to

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

and

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

In the following lemma, I make precise the intuition that if a plan is optimal, the agent's incentive constraint will bind for almost all t.

Lemma 5.1 *In any optimal plan, the agent's incentive constraint binds a.s.*

PROOF: Suppose $(h_t, w_t)_{0 \le t \le T}$ is an optimal plan. As the plan is optimal, it must be incentive compatible for a.a. t. This means that either $h_t > 0$ or $w_t > 0$ for a.a. t, for otherwise playing safe is a strictly dominant action for the agent. Now, suppose that, under $(h_t, w_t)_{0 \le t \le T}$, the incentive constraint was slack on a set of positive measure. This means that there exists an interval $[t_1, t_2]$, with $t_1 < t_2$, such that the incentive constraint is slack a.e. on $[t_1, t_2]$. Then there exists a collection $(\epsilon_t)_{t_1 \le t \le t_2}$ with $\epsilon_t > 0$ and a plan $(\tilde{h}_t, \tilde{w}_t)_{0 \le t \le T}$ satisfying $(\tilde{h}_t, \tilde{w}_t) = (h_t, w_t)$ if $t \in [0, t_1) \cup (t_2, T]$, and $\tilde{h}_t + \tilde{w}_t = h_t + w_t - \epsilon_t$ if $t \in [t_1, t_2]$ such that $(\tilde{h}_t, \tilde{w}_t)$ would be incentive compatible given the old state variable μ_t. It follows immediately from the explicit expression for the incentive constraint that $(\tilde{h}_t, \tilde{w}_t)$ is incentive compatible given the new state variable $\tilde{\mu}_t$ if $\tilde{\mu}_t \geq \mu_t$. Yet the appertaining law of motion immediately implies that $\tilde{\mu}_t = \mu_t$ if $t > t_2$, and $\tilde{\mu}_t = \mu_t + \lambda_1 \int_t^{t_2} e^{-r\tau - x_\tau} \epsilon_\tau \, d\tau \geq \mu_t$ otherwise, where we set $\epsilon_t = 0$ for $t < t_1$. Hence, $(\tilde{h}_t, \tilde{w}_t)$ is incentive compatible, and, as $[t_1, t_2]$ has positive measure, the principal is strictly better off under $(\tilde{h}_t, \tilde{w}_t)_{0 \leq t \leq T}$, contradicting the optimality of $(h_t, w_t)_{0 \leq t \leq T}$. ▮

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

Lemma 5.2 *The principal can without loss restrict himself to plans* $(h_t, w_t)_{0 \leq t \leq T}$ *with* $h_t = 0$ *for all* t.

Proof: Consider an incentive compatible plan $(\hat{h}_t, \hat{w}_t)_{0 \leq t \leq T}$ with $\hat{h}_t > 0$ for some t. Consider the alternative plan $(h_t, w_t)_{0 \leq t \leq T}$ with $h_t = 0$ and $w_t = \hat{w}_t + \hat{h}_t$. As is apparent from the explicit expression for $\dot{\mu}_t$, μ_t is unaffected by the change; applying Proposition 4.1 for the same ϵ in both cases, one shows that $(h_t, w_t)_{0 \leq t \leq T}$ is incentive compatible. Moreover, it gives the principal exactly the same payoff as the original plan $(\hat{h}_t, \hat{w}_t)_{0 \leq t \leq T}$. ▮

The intuition for this result is that when paying an immediate lump sum for the first breakthrough, the principal cannot discriminate between an agent who has achieved his first breakthrough on arm 0, and an equilibrium agent who will enjoy an informational advantage in the continuation game. Indeed, by Proposition 4.1, the principal can make sure that an increase in w_t translates into less of an increase in ω_t, whereas h_t is paid out indiscriminately to on-path and off-path agents alike. Hence, incentive provision can only be helped when incentives are given through the continuation game rather than through immediate lump sum payments.

Now, we are ready to characterize the optimal incentive scheme, which is essentially unique in the class of optimal schemes with $h_t = 0$ for a.a. t, as the following proposition shows. The characterization relies on the fact, which we have formalized in Lemma 5.1, that it never pays for the principal to give strict rather than weak incentives for the agent to do the right thing, because if he did, he could lower his expected wage bill while still providing adequate incentives. This means that the agent is indifferent between doing the right thing and using arm 1, on the one hand, and his next best outside option on the other hand. Yet, the wage scheme we have constructed in Section 4 makes sure that the agent's best outside option can never be arm 0. Indeed, playing arm 0 yields the agent approximately $p_t w_t$ after a breakthrough, which occurs with an instantaneous probability of $\lambda_0 dt$ if arm 0 is pulled over a time interval of infinitesimal length dt. Arm 1, by contrast, yields w_t in case of a breakthrough, which occurs with an instantaneous probability of $p_t \lambda_1 dt$; thus, as $\lambda_1 > \lambda_0$, arm 1 dominates arm 0. Any optimal incentive scheme now has the property that the agent is exactly indifferent between the safe arm and arm 1. To facilitate the statement of the next two propositions, it is judicious to define the function $\tilde{w}(t)$, the reward that the agent has to be paid to be kept indifferent between using arm 1 and the safe arm, given he will use arm 1 at all future times:

$$\tilde{w}(t) := \begin{cases} \frac{s}{\lambda_1 p_t} + \frac{s}{r}(1 - e^{-r(T-t)}) + \frac{1-p_t}{p_t} \frac{s}{r-\lambda_1} \left(1 - e^{-(r-\lambda_1)(T-t)}\right) & \text{if } \; r \neq \lambda_1 \\ \frac{s}{\lambda_1 p_t} + \frac{s}{r}(1 - e^{-r(T-t)}) + \frac{1-p_t}{p_t} s(T-t) & \text{if } \; r = \lambda_1. \end{cases}$$

Before I characterize the full set of optimal wage schemes, I first take note of the essentially unique optimal scheme with the feature that $h_t = 0$ for all t, as doing so facilitates our proof later on:

Proposition 5.3 *An optimal wage scheme is given by $h_t = 0$ and $w_t = \tilde{w}(t)$ for all $t \in [0, T]$.*

PROOF: By Lemma 5.2, we know that the principal can without loss restrict himself to wage schemes $(h_t, w_t)_{0 \leq t \leq T}$ with $h_t = 0$ for all t. If such a scheme is optimal, Lemma 5.1 implies that, for a.a. t, one of the following two constraints will bind almost surely:

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

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

Now, suppose that the constraint (4) is slack on a set of positive measure. This means that there exist times $t_1 < t_2$ such that (4) is slack a.s. on $[t_1, t_2]$. Lemma 5.1 now implies that constraint (5) will bind a.s. on $[t_1, t_2]$. Simple algebra now shows that for (4) to hold given that (5) binds, it has to be the case that $\omega_t \geq p_t w_t + \left(\frac{1}{\lambda_0} - \frac{1}{\lambda_1}\right) s$ a.e. on $[t_1, t_2]$. Yet, by Proposition 4.1, there exists an alternative scheme $(\tilde{h}_t, \tilde{w}_t, \tilde{\omega}_t)$ with $\tilde{h}_t := h_t = 0$ and $\tilde{w}_t := w_t$ for all t, yet $\tilde{\omega}_t < p_t w_t + \left(\frac{1}{\lambda_0} - \frac{1}{\lambda_1}\right) s$. Clearly, $(\tilde{h}_t, \tilde{w}_t, \tilde{\omega}_t) = (h_t, w_t, \tilde{\omega}_t)$ still satisfies (4) with slackness a.e. on $[t_1, t_2]$, since (4), as well as μ_t, are independent of $\tilde{\omega}_t$. Since $\tilde{\omega}_t < \omega_t$ a.e. on $[t_1, t_2]$, it follows that (5) is now also slack a.s. on $[t_1, t_2]$. Hence, by the same argument as in the proof of Lemma 5.1, there exists a sequence of $(\delta_t)_{t_1 \leq t \leq t_2}$ with $\delta_t > 0$ such that, for $\hat{w}_t := w_t - \delta_t$, $(h_t, \hat{w}_t, \tilde{\omega}_t)$ satisfy both constraints and imply lower wage costs for the principal on $[t_1, t_2]$, a set of positive measure, contradicting the optimality of (h_t, w_t, ω_t).

Thus, we have shown that if $h_t = 0$ for all t, and (h_t, w_t) is optimal, then (4) binds a.s. Conversely, this scheme is clearly optimal, since it is impossible further to reduce w_t on any time interval of positive measure without violating (4). Furthermore, as we have discussed, we have that $\mu_t = -\lambda_1 \int_t^T e^{-r\tau - x_\tau} w_\tau \, d\tau$, which we can plug into condition (4) to solve for w_t, thus completing the proof. $\blacksquare$

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

Proposition 5.4 *Any optimal wage scheme $(h_t, w_t)_{0\leq t\leq T}$ has the property that, prior to the first breakthrough, it keeps the agent indifferent between arm 1 and the safe arm almost surely.*

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

Indeed, suppose it is not the case that $h_t > 0$ a.s. on $[t_1, t_2]$. Then, there exists a time interval $[t_1', t_2'] \subseteq [t_1, t_2]$ such that $t_1' < t_2'$ and $h_t = 0$ a.s. on $[t_1', t_2']$. Since the agent a.s. strictly prefers arm 1 over the safe arm on this interval, it follows by Lemma 5.1 that the constraint (5) will bind a.s. on $[t_1', t_2']$, which, by the same argument as in the proof of Proposition 5.3, contradicts optimality.

Therefore, $h_t > 0$ a.s. on $[t_1, t_2]$. As the agent strictly prefers arm 1 over the safe arm, Lemma 5.1 implies that the incentive constraint

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

will bind for a.a. $t \in [t_1, t_2]$. Now, consider the alternative plan $(\hat{h}_t, \hat{w}_t)_{0\leq t\leq T}$ with $\hat{h}_t = 0$ and $\hat{w}_t = w_t + h_t$ for all $t \in [t_1, t_2]$, and $\hat{h}_t = h_t$ and $\hat{w}_t = w_t$ for all $t \in [0, t_1) \cup (t_2, T]$. Arguing as in the proof of Lemma 5.2, one shows that under $(\hat{h}_t, \hat{w}_t)_{0\leq t\leq T}$ the incentive constraint (3) is a.s. slack on $[t_1, t_2]$, and gives the principal exactly the same payoff as the original plan $(h_t, w_t)_{0\leq t\leq T}$. Therefore, by Lemma 5.1, the principal can strictly improve over $(\hat{h}_t, \hat{w}_t)_{0\leq t\leq T}$, and hence over $(h_t, w_t)_{0\leq t\leq T}$. Thus, we have shown that if (h_t, w_t) is optimal, the incentive constraint for the safe arm binds a.s. Conversely, these schemes are clearly optimal since the principal cannot further lower $h_t + w_t$ without violating the incentive constraint (3). $\blacksquare$

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

Corollary 5.5 *If $(h_t, w_t)_{0\leq t\leq T}$ is optimal, then $h_t + w_t = \tilde{w}(t)$ t-a.s.*

6 The Optimal Stopping Time

In this section, the principal can optimally choose the end date T, which had been exogenous thus far. As the first-best benchmark, I use the solution which is given by the hypothetical situation in which the principal operates the bandit himself, and he decides when to stop using arm 1, which he pulls at a flow cost of s, conditional on not having obtained a success thus far. Thus, he chooses T so as to maximize

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

subject to $\dot{p}_t = -\lambda_1 p_t(1-p_t)$ for all $t \in (0,T)$. Clearly, the integrand is positive if, and only if, $p_t\lambda_1\Pi \geq s$, i.e. as long as $p_t \geq \frac{s}{\lambda_1 \Pi} =: p^m$. As the principal is only interested in the first breakthrough, information has no value for him, meaning that, very much in contrast to the classical bandit literature, he is not willing to forgo current payoffs in order to *learn* something about the state of the world. In other words, he will behave myopically, i.e. as though the future was of no consequence to him, and stops playing risky at his myopic cutoff belief p^m, which is reached at time $T^{FB} = \frac{1}{\lambda_1} \ln \left(\frac{p_0}{1-p_0} \frac{1-p^m}{p^m} \right)$.

Now, when the principal delegates the investigation to an agent, his goal is to choose T so as to maximize

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

subject to $\dot{p}_t = -\lambda_1 p_t(1-p_t)$ for all $t \in (0,T)$.

Thus, all that changes with respect to the first best problem (6) is that the opportunity cost flow s is now replaced by the wage costs $h_t + w_t$, which only have to be paid out in case of a success, which happens with an instantaneous probability of $p_t\lambda_1 dt$. After plugging in the optimal wage costs, which we have computed in Proposition 5.4, one finds that the first-order derivative of the objective with respect to T is given by

$$\underbrace{e^{-(r+\lambda_1)T} \left(\lambda_1 \Pi - \frac{s}{p_T} \right)}_{\text{Marginal effect}} \underbrace{-e^{-rT} \frac{s}{p_T} \left(1 - e^{-\lambda_1 T} \right)}_{\text{Intra-marginal effect}} . \tag{8}$$

The marginal effect captures the benefit the principal could collect by extending experimentation for an additional instant at time T. Yet, the choice of an end date T also entails an intra-marginal effect at times $t < T$. Indeed, for him to use arm 1 at time t, the agent has to be compensated for the opportunity cost of the potentially forgone rewards for having his first breakthrough at some future date, an effect that is the stronger "the more future there is," i.e. the more distant the end date T is. Hence, by marginally increasing T, the principal also marginally raises his wage liabilities at times $t < T$. This creates a distortion, so that the following proposition comes as no surprise:

Proposition 6.1 *Let $p_0 > p^m$. The principal stops the game at the time $T^* \in (0, T^{FB})$ when $p_{T^*} = p^m e^{\lambda_1 T^*}$, i.e.*

$$T^* = \frac{1}{\lambda_1} \ln \left(\frac{-p^m p_0 + \sqrt{(p^m p_0)^2 + 4p^m p_0(1-p_0)}}{2p^m(1-p_0)} \right).$$

Proof: The formula for p_{T^*} is gotten by setting the expression (8) to 0, and verifying that the second-order condition holds. Now, T^* is the unique positive root of $\frac{p_0 e^{-\lambda_1 T^*}}{p_0 e^{-\lambda_1 T^*}+1-p_0} = p^m e^{\lambda_1 T^*}$. ∎

The size of the distortion can be measured by the ratio $\frac{p_{T^*}-p^m}{p^m}$, which is increasing both in the stakes at play, as measured by the ratio $\frac{1}{p^m} = \frac{\lambda_1 \Pi}{s}$, as well as in players' optimism, as measured by p_0. This is hardly surprising, since when delegating the project to an agent, the principal can only appropriate part of any increase in the overall pie. The following corollary summarizes these comparative statics:

Corollary 6.2 *The wedge $\frac{p_{T^*}-p^m}{p^m}$ is increasing in the stakes at play $\frac{\lambda_1 \Pi}{s}$, as well as in players' optimism p_0.*

Proof: For $\frac{1}{p^m}$, this immediately follows from the explicit expression for the wedge

$$\frac{p_{T^*} - p^m}{p^m} = \frac{p_0 - 2 + \sqrt{p_0^2 + 4\frac{p_0}{p^m}(1-p_0)}}{2(1-p_0)}.$$

For p_0, one shows that the sign of the first-order derivative of this expression is equal to the sign of

$$p_0 \left[1 - \sqrt{1 + \frac{4}{p^m p_0}(1-p_0)} \right] + (1-p_0)\frac{2}{p^m},$$

which is strictly positive if, and only if,

$$p_0 + (1-p_0)\frac{2}{p^m} > \sqrt{p_0^2 + \frac{4}{p^m} p_0 (1-p_0)}$$

$$\iff \left(\frac{2}{p^m}(1-p_0) \right)^2 > 0.$$

∎

Yet, also recall from the preceding sections that given the optimal incentive scheme we have computed there, the principal only needs to compensate the agent for his outside option of using the safe arm, which is exactly what arm 1 has to compensate the principal for in the first-best problem. Put differently, the presence of a cheating action, arm 0, does *not* give rise to any distortions; the only distortions that arise are due to the fact that high future rewards to some extent cannibalize today's rewards. Yet, in most applications, the principal's access will not be restricted to a single agent; rather, he will be able to hire several agents sequentially if he so chooses. Now, in the limit, if the principal can hire agents for a mere infinitesimal instant dt, he can completely shut down the intra-marginal effect we have discussed above.[9] Indeed, we can see from the formula for $\tilde{w}$ that the reward an agent who is only hired for an instant of length dt would have to be promised for a breakthrough is given by $\frac{s}{\lambda_1 p_t}(1+\lambda_1 dt)+o(dt^2)$. Hence, it pays for the principal to go on with the project as long as $p_t\lambda_1\left(\Pi - \frac{s}{\lambda_1 p_t}(1+\lambda_1 dt)\right)dt + o(dt^3) = p_t\lambda_1\left(\Pi - \frac{s}{\lambda_1 p_t}\right)dt + o(dt^2) > 0$, i.e. he stops at the first-best efficient stopping time, a result I summarize in the following proposition:

Proposition 6.3 *If the principal has access to a sequence of different agents, he stops the delegated project at the time T^{FB} when $p_{T^{FB}} = p^m$.*

Thus, while delegating the project to an agent forces the principal to devise quite a complicated incentive scheme, it only induces him to stop the exploration inefficiently early if he does not have access to a sequence of many different agents. In summary, if $\lambda_0 \geq \lambda_1$, it is impossible to have the agent use arm 1; if $\lambda_0 < \lambda_1$, with a sequence of many agents, it is even possible to give incentives in a manner that renders the principal willing to implement the efficient amount of exploration.

7 Conclusion

The present paper introduces the question of optimal incentive design into a dynamic single-agent model of experimentation on bandits. I have shown that even though the principal

[9] Intuitively, one might think that hiring *one* particularly myopic agent might remedy the problem as well. However, while it is true that the impact future rewards have on today's incentives, and hence the intra-marginal effect of an extended end time T, becomes arbitrarily small as the players become very impatient, the same holds true for the marginal benefit of extending play for an instant after a given time $T > 0$, so that in sum the distortion is independent of the players' discount rate. If one were to relax the assumption that the players share the same discount rate, the problem could conceivably be addressed by the principal's hiring an agent who is much more impatient than himself. However, doing so would impact our analysis in the previous sections, in that it would introduce an additional cost to the provision of incentives through the continuation value w_t, as opposed to the immediate rewards h_t.

only cares about the first breakthrough, he only rewards later ones.

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

Appendix

Proof of Lemma 4.2

Fix an arbitrary $\check{T}(t) \in (t, \overline{T}]$, $\tilde{t} \in (t, \check{T}(t)]$, $\hat{p}_{\tilde{t}} \in [p_{\tilde{t}}, p_0]$, and $\overline{V}_0 > 0$. Consider the restricted problem in which the agent can only choose between arms 0 and 1. Then, the agent's reward is given by

$$\int_{\tilde{t}}^{\check{T}(t)} e^{-r(\tau_m - \tilde{t})} \left(\overline{V}_0 + \frac{s}{r}\left(1 - e^{-r(\check{T}(t) - \tau_m)}\right)\right) dF,$$

where F is the distribution over τ_m, the time of the m-th breakthrough after time $\tilde{t}$. As the integrand is decreasing in τ_m, all that remains to be shown is that $F^*(\tau; \hat{p}_{\tilde{t}})$, the probability of m breakthroughs up to time $\tau \in (\tilde{t}, \check{T}(t)]$ when the agent always pulls arm 1, is first-order stochastically dominated by the probability of m breakthroughs up to time τ for *any* alternative strategy, which I shall denote by $\tilde{F}(\tau; \hat{p}_{\tilde{t}})$. Fix an arbitrary $\tau \in (\tilde{t}, \check{T}(t)]$. Now,

$$F^*(\tau; \hat{p}_{\tilde{t}}) = \hat{p}_{\tilde{t}} \frac{\lambda_1^m}{m!} (\tau - \tilde{t})^m e^{-\lambda_1(\tau - \tilde{t})}.$$

Whatever the alternative strategy under consideration may be, $\tilde{F}$ can be written as

$$\tilde{F}(\tau; \hat{p}_{\tilde{t}}) = \int_0^1 F_\alpha(\tau; \hat{p}_{\tilde{t}})\, \mu(d\alpha),$$

with

$$F_\alpha(\tau; \hat{p}_{\tilde{t}}) = \hat{p}_{\tilde{t}} \frac{[\alpha\lambda_1 + (1-\alpha)\lambda_0]^m}{m!} \left(\tau - \tilde{t}\right)^m e^{-(\alpha\lambda_1 + (1-\alpha)\lambda_0)(\tau - \tilde{t})} \\ + (1 - \hat{p}_{\tilde{t}}) \frac{[(1-\alpha)\lambda_0]^m}{m!} \left(\tau - \tilde{t}\right)^m e^{-(1-\alpha)\lambda_0(\tau - \tilde{t})}$$

for some probability measure μ on $\alpha \in [0, 1]$. The weight α can be interpreted as the fraction of the time interval $[\tilde{t}, \tau]$ devoted to arm 1; of course, since the agent's strategy allows him to condition his action on the entire previous history, α will generally be stochastic. Therefore, the strategy of the proof is to find an m such that for any $\tilde{t} \in (t, \overline{T})$, $\tau \in (\tilde{t}, \overline{T}]$ and $\hat{p}_{\tilde{t}} \in [p_{\overline{T}}, p_0]$, it is the case that

$$F^*(\tau; \hat{p}_{\tilde{t}}) > F_\alpha(\tau; \hat{p}_{\tilde{t}}) \tag{.1}$$

uniformly for all $\alpha \in (0, 1)$.

Computations show that

$$\frac{\partial F_\alpha}{\partial \alpha} = \frac{(\tau - \tilde{t})^m}{m!} \Big\{ \hat{p}_{\tilde{t}} e^{-(\alpha\lambda_1 + (1-\alpha)\lambda_0)(\tau - \tilde{t})} (\lambda_1 - \lambda_0) \left(\alpha\lambda_1 + (1-\alpha)\lambda_0\right)^{m-1} \left[m - (\alpha\lambda_1 + (1-\alpha)\lambda_0)(\tau - \tilde{t})\right] \\ - (1 - \hat{p}_{\tilde{t}}) e^{-(1-\alpha)\lambda_0(\tau - \tilde{t})} \lambda_0 \left((1-\alpha)\lambda_0\right)^{m-1} \left[m - (1-\alpha)\lambda_0(\tau - \tilde{t})\right] \Big\}.$$

Further computations show that $\frac{\partial F_\alpha}{\partial \alpha} > 0$ if and only if

$$\xi(\alpha) > 1 - \lambda_0 \frac{\tau - \tilde{t}}{m}$$

with

$$\xi(\alpha) := \frac{\hat{p}_{\tilde{t}}}{1-\hat{p}_{\tilde{t}}} e^{-\alpha\lambda_1(\tau-\tilde{t})} \left(\frac{\lambda_1}{\lambda_0} - 1\right) \left(\frac{\alpha\lambda_1 + (1-\alpha)\lambda_0}{(1-\alpha)\lambda_0}\right)^{m-1} \left[1 - (\alpha\lambda_1 + (1-\alpha)\lambda_0)\frac{\tau-\tilde{t}}{m}\right] - \alpha\lambda_0 \frac{\tau-\tilde{t}}{m}.$$

Now, we choose $m \geq 2$ high enough that

$$(m-1)\left(\frac{\lambda_1}{\lambda_0} - \frac{\lambda_1^2}{\lambda_0}\frac{\overline{T}}{m}\right) - \frac{\lambda_1^2}{\lambda_0}\overline{T}\left(1 + \frac{1}{m}\right) > \frac{1-p_{\overline{T}}}{p_{\overline{T}}} \frac{\lambda_0^2}{\lambda_1 - \lambda_0} e^{\lambda_1 \overline{T}} \frac{\overline{T}}{m}. \tag{.2}$$

As the left-hand side of (.2) is increasing in m, and diverging for $m \to +\infty$, such an $m \geq 2$ exists. Algebra shows that this ensures that $\lim_{\alpha\uparrow 1} \xi(\alpha) = \infty$ and $\xi'(\alpha) > 0$ for all $\alpha \in (0,1)$. Now, if $\hat{p}_{\tilde{t}}\lambda_1 \geq \lambda_0$, it is the case that $\xi(0) \geq 1 - \lambda_0 \frac{\tau-\tilde{t}}{m}$, and hence $\frac{\partial F_\alpha}{\partial \alpha} > 0$ for all $\alpha \in (0,1)$, so that $F^*(\tau;\hat{p}_{\tilde{t}}) > F_\alpha(\tau;\hat{p}_{\tilde{t}})$ for all $\alpha \in [0,1)$. In case $\hat{p}_{\tilde{t}}\lambda_1 < \lambda_0$, we have that $\xi(0) < 1 - \lambda_0 \frac{\tau-\tilde{t}}{m}$; now, there exists a unique $\alpha^* \in (0,1)$ such that $\frac{\partial F_\alpha}{\partial \alpha} < 0$ for all $\alpha \in (0,\alpha^*)$, $\frac{\partial F_\alpha}{\partial \alpha} > 0$ for all $\alpha \in (\alpha^*,1)$, and we can conclude that $F_\alpha(\tau;\hat{p}_{\tilde{t}})$ is maximized either by $\alpha = 1$ or $\alpha = 0$. Choosing m such that

$$p_{\overline{T}} \left(\frac{\lambda_1}{\lambda_0}\right)^m > e^{(\lambda_1-\lambda_0)\overline{T}} \tag{.3}$$

ensures that the maximum is indeed attained at $\alpha = 1$.

In summary, there clearly exists an $m \in \mathbb{N} \cap [2,\infty)$ satisfying both (.2) and (.3). Choosing m in this manner ensures that

$$F^*(\tau;\hat{p}_{\tilde{t}}) > F_\alpha(\tau;\hat{p}_{\tilde{t}})$$

for all $\alpha \in [0,1)$. Now, for such an m, it is clearly the case that $F^*(\tau;\hat{p}_{\tilde{t}}) > \tilde{F}(\tau;\hat{p}_{\tilde{t}})$ for any $\tau > \tilde{t}$ whenever $\mu \neq \delta_1$, where δ_1 denotes the Dirac measure associated with the strategy of always of always pulling arm 1, whatever befall.

It remains to be shown that the preference ordering does not change if the agent also has access to the safe arm. In this case, his goal is to maximize

$$\int_{\tilde{t}}^{\check{T}(t)} \left\{(1-k_\tau)e^{-r(\tau-\tilde{t})}s + \int_{\tilde{t}}^{\check{T}(t)} e^{-r(\tau_m-\tilde{t})}\left(\overline{V}_0 + \frac{s}{r}(1-e^{-r(\check{T}(t)-\tau_m)})\right) d\tilde{F}_{\{k_\tau\}}(\tau_m)\right\} d\nu\left(\{k_\tau\}_{\tilde{t}\leq\tau\leq\check{T}(t)}\right)$$

for some probability measures $\tilde{F}$ and ν, with the process $\{k_\tau\}$ satisfying $0 \leq k_\tau \leq 1$ for all $\tau \in [\tilde{t}, \check{T}(t)]$.

First, we fix some arbitrary probability measure ν and an arbitrary $\tau \in (\tilde{t}, \check{T}(t)]$ such that $Prob_\nu\left(\int_{\tilde{t}}^{\tau} k_\sigma\, d\sigma > 0\right) > 0$. [If τ is such that $Prob_\nu\left(\int_{\tilde{t}}^{\tau} k_\sigma\, d\sigma > 0\right) = 0$, the objective is invariant in α.] Now, consider an arbitrary $\{k_\sigma\}_{\sigma=\tilde{t}}^{\tau}$ with $\int_{\tilde{t}}^{\tau} k_\sigma\, d\sigma > 0$. Arguing as above, we can now write

$$\tilde{F}(\tau;\hat{p}_{\tilde{t}}) = \int_0^1 F_\alpha(\tau;\hat{p}_{\tilde{t}})\,\mu(d\alpha)$$

for

$$F_\alpha(\tau;\hat{p}_{\tilde{t}}) = \hat{p}_{\tilde{t}} \frac{[\alpha\lambda_1 + (1-\alpha)\lambda_0]^m}{m!}\left(\int_{\tilde{t}}^{\tau} k_\sigma\, d\sigma\right)^m e^{-(\alpha\lambda_1+(1-\alpha)\lambda_0)\int_{\tilde{t}}^{\tau} k_\sigma\, d\sigma}$$
$$+ (1-\hat{p}_{\tilde{t}})\frac{[(1-\alpha)\lambda_0]^m}{m!}\left(\int_{\tilde{t}}^{\tau} k_\sigma\, d\sigma\right)^m e^{-(1-\alpha)\lambda_0\int_{\tilde{t}}^{\tau} k_\sigma\, d\sigma}$$

and some probability measure μ. Since all that changes with respect to our calculations above is for $\tau - \tilde{t}$ to be replaced by $\int_{\tilde{t}}^{\tau} k_\sigma \, d\sigma \leq \tau - \tilde{t}$, and our previous τ was arbitrary, the previous calculations continue to apply. In particular, any $m \geq 2$ satisfying conditions (.2) and (.3) ensures that F^* be first-order stochastically dominated by any $\tilde{F}$, as long as $\mu \neq \delta_1$. As $e^{-r(\tau_m - \tilde{t})} \left(\overline{V}_0 + \frac{s}{r}(1 - e^{-r(\check{T}(t) - \tau_m)} \right)$ is decreasing in τ_m, we can conclude that $\alpha = 1$ is (strictly) optimal for all $\{k_\sigma\}_{\sigma=\tilde{t}}^{\tau}$ (with $\int_{\tilde{t}}^{\tau} k_\sigma \, d\sigma > 0$). ∎

Proof of Lemma 4.3

To analyze the agent's best responses, I shall make use of Bellman's Principle of Optimality. Standard arguments imply that the player's payoff function from playing a best response is once continuously differentiable, and satisfies the Bellman equation. A first-order Taylor expansion gives that

$$V_i(\tilde{t}; \overline{V}_0) = \left[s + k_{1,\tilde{t}} \left(\lambda_1 V_{i-1}(\tilde{t}; \overline{V}_0) - s \right) \right] dt + (1 - r dt)(1 - k_{1,\tilde{t}} \lambda_1 dt) \left(V_i(\tilde{t}; \overline{V}_0) + \dot{V}_i(\tilde{t}; \overline{V}_0) dt \right)$$

$$\iff r V_i(\tilde{t}; \overline{V}_0) = s + \dot{V}_i(\tilde{t}; \overline{V}_0) + k_{1,\tilde{t}} \left[\lambda_1 \left(V_{i-1}(\tilde{t}; \overline{V}_0) - V_i(\tilde{t}; \overline{V}_0) \right) - s \right]. \tag{.4}$$

Hence, by Bellman's principle, $k_{1,\tilde{t}} = 1$ is optimal if, and only if,

$$V_{i-1}(\tilde{t}; \overline{V}_0) - V_i(\tilde{t}; \overline{V}_0) \geq \frac{s}{\lambda_1}, \tag{.5}$$

and it is uniquely optimal if, and only if, this inequality is strict.

For $i = 1$, setting $k_{1,\tau} = 1$ for all $\tau \in [\tilde{t}, \check{T}(t)]$ implies

$$V_1(\tilde{t}; \overline{V}_0) = \frac{\lambda_1}{\lambda_1 + r} \left(1 - e^{-(r+\lambda_1)(\check{T}(t) - \tilde{t})} \right) \left(\overline{V}_0 + \frac{s}{r} \right) - \frac{s}{r} e^{-r(\check{T}(t) - \tilde{t})} \left(1 - e^{-\lambda_1(\check{T}(t) - \tilde{t})} \right).$$

$\dot{V}_1$ exists, and, because $\overline{V}_0 > \frac{s}{\lambda_1}$, satisfies

$$\dot{V}_1(\tilde{t}; \overline{V}_0) = -\lambda_1 e^{-(r+\lambda_1)(\check{T}(t) - \tilde{t})} \overline{V}_0 - s e^{-r(\check{T}(t) - \tilde{t})} \left(1 - e^{-\lambda_1(\check{T}(t) - \tilde{t})} \right) \leq -s e^{-r(\check{T}(t) - \tilde{t})} < 0.$$

By simple algebra, one finds that

$$V_0(\tilde{t}; \overline{V}_0) - V_1(\tilde{t}; \overline{V}_0) = \left(\frac{r}{r + \lambda_1} + \frac{\lambda_1}{r + \lambda_1} e^{-(r+\lambda_1)(\check{T}(t) - \tilde{t})} \right) \overline{V}_0 + \frac{s}{r + \lambda_1} \left(1 - e^{-(r+\lambda_1)(\check{T}(t) - \tilde{t})} \right),$$

which one shows strictly to exceed $\frac{s}{\lambda_1}$ for all $\tilde{t} \in (t, \check{T}(t)]$ if $\overline{V}_0 > \frac{s}{\lambda_1}$. We conclude that for $i = 1$, a cutoff strategy with $t_1^* = \check{T}(t)$ is optimal, and that V_1 is of class C^1 and strictly decreasing with $\dot{V}_1(\tilde{t}; \overline{V}_0) \leq -s e^{-r(\check{T}(t) - \tilde{t})}$ for all $\tilde{t}$.

Now let $i > 1$. As my induction hypothesis, I posit that V_{i-1} is of the following structure:

$$V_{i-1}(\tilde{t}; \overline{V}_0) = \begin{cases} \int_{\tilde{t}}^{t_{i-1}^*} e^{-(r+\lambda_1)(\tau - \tilde{t})} \lambda_1 V_{i-2}(\tau; \overline{V}_0) \, d\tau + e^{-(r+\lambda_1)(t_{i-1}^* - \tilde{t})} \frac{s}{r} \left(1 - e^{-r(\check{T}(t) - t_{i-1}^*)} \right) & \text{if } \tilde{t} \leq t_{i-1}^* \\ \frac{s}{r} \left(1 - e^{-r(\check{T}(t) - \tilde{t})} \right) & \text{if } \tilde{t} > t_{i-1}^* \end{cases}$$

for some $t_{i-1}^* \leq \check{T}(t)$. It is furthermore assumed that V_{i-1} is C^1, and that $\dot{V}_{i-1}(\tilde{t}; \overline{V}_0) \leq -s e^{-r(\check{T}(t) - \tilde{t})}$ for all $\tilde{t} \in (t, \check{T}(t))$.

Now, if $V_{i-1}(t;\overline{V}_0) < \frac{s}{\lambda_1} + \frac{s}{r}\left(1 - e^{-r(\check{T}(t)-t)}\right)$, I set $t_i^* = t$. Otherwise, I define t_i^* as the lowest t^* satisfying $V_{i-1}(t^*;\overline{V}_0) = \frac{s}{\lambda_1} + \frac{s}{r}\left(1 - e^{-r(\check{T}(t)-t^*)}\right)$. Since $\dot{V}_{i-1}(\tilde{t};\overline{V}_0) \leq -se^{-r(\check{T}(t)-\tilde{t})}$ for all $\tilde{t} \in (t, \check{T}(t))$, V_{i-1} is continuous, and $V_{i-1}(\check{T}(t);\overline{V}_0) = 0$, it is the case that t_i^* exists, and $t_i^* < \check{T}(t)$.

Fix an arbitrary $\tilde{t} \in (t, \check{T}(t))$. If $V_{i-1}(\tilde{t};\overline{V}_0) \leq \frac{s}{\lambda_1} + \frac{s}{r}\left(1 - e^{-r(\check{T}(t)-\tilde{t})}\right)$, i.e. $\tilde{t} \geq t_i^*$, condition (.5) implies that setting $k_{1,\tau} = 0$ for all $\tau \in [\tilde{t}, \check{T}(t)]$ is a best response (since $\dot{V}_{i-1}(\tau;\overline{V}_0) \leq -se^{-r(\check{T}(t)-\tau)}$), and $V_i(\tilde{t};\overline{V}_0) = \frac{s}{r}\left(1 - e^{-r(\check{T}(t)-\tilde{t})}\right)$. This establishes that $k_{1,\tilde{t}} = 0$ is a best response for all $\tilde{t} \geq t_i^*$.

Now, let us assume that $V_{i-1}(\tilde{t};\overline{V}_0) > \frac{s}{\lambda_1} + \frac{s}{r}\left(1 - e^{-r(\check{T}(t)-\tilde{t})}\right)$. Yet, suppose that $V_{i-1}(\tilde{t};\overline{V}_0) - V_i(\tilde{t};\overline{V}_0) \leq \frac{s}{\lambda_1}$. This implies that $V_i(\tilde{t};\overline{V}_0) > \frac{s}{r}\left(1 - e^{-r(\check{T}(t)-\tilde{t})}\right)$; hence, it follows that there exists some $\hat{t} \in (\tilde{t}, \check{T}(t))$ when $k_{1,\hat{t}} = 1$ is a best response. Let $\hat{t}_1$ be the lowest such $\hat{t}$.[10] First, suppose that $\hat{t}_1 > \tilde{t}$. Then, for all $t^\dagger \in (\tilde{t}, \hat{t}_1)$, $V_i(t^\dagger;\overline{V}_0) > \frac{s}{r}\left(1 - e^{-r(\check{T}(t)-t^\dagger)}\right)$, and hence, by (.4), $\dot{V}_i(t^\dagger;\overline{V}_0) > -se^{-r(\check{T}(t)-t^\dagger)} \geq \dot{V}_{i-1}(t^\dagger;\overline{V}_0)$ for all $t^\dagger \in [\tilde{t}, \hat{t}_1]$—contradicting $V_{i-1}(\hat{t}_1;\overline{V}_0) - V_i(\hat{t}_1;\overline{V}_0) \geq \frac{s}{\lambda_1}$. Now, suppose $\hat{t}_1 = \tilde{t}$, and $V_{i-1}(\tilde{t};\overline{V}_0) - V_i(\tilde{t};\overline{V}_0) = \frac{s}{\lambda_1}$. In this case, $\lim_{\tau\downarrow\tilde{t}} \dot{V}_i(\tau;\overline{V}_0) > -se^{-r(\check{T}(t)-\tilde{t})} \geq \dot{V}_{i-1}(\tilde{t};\overline{V}_0)$. Hence, $k_1 = 0$ is a strict best response to the immediate right of $\tilde{t}$, and hence, by our previous step, at all $t^\ddagger > \tilde{t}$. By continuity of V_i, this contradicts $V_i(\tilde{t};\overline{V}_0) > \frac{s}{r}\left(1 - e^{-r(\check{T}(t)-\tilde{t})}\right)$. This establishes that $k_{1,\tilde{t}} = 1$ is a best response for all $\tilde{t} \leq t_i^*$.

Thus, I have shown that

$$V_i(\tilde{t};\overline{V}_0) = \begin{cases} \int_{\tilde{t}}^{t_i^*} e^{-(r+\lambda_1)(\tau-\tilde{t})}\lambda_1 V_{i-1}(\tau;\overline{V}_0)\,d\tau + e^{-(r+\lambda_1)(t_i^*-\tilde{t})}\frac{s}{r}\left(1 - e^{-r(\check{T}(t)-t_i^*)}\right) & \text{if } \tilde{t} \leq t_i^* \\ \frac{s}{r}\left(1 - e^{-r(\check{T}(t)-\tilde{t})}\right) & \text{if } \tilde{t} > t_i^* \end{cases}.$$

This implies that $\dot{V}_i(\tilde{t};\overline{V}_0) = -se^{-r(\check{T}(t)-\tilde{t})}$ if $\tilde{t} > t_i^*$, and

$$\begin{aligned}\dot{V}_i(\tilde{t};\overline{V}_0) = -\lambda_1 e^{-(r+\lambda_1)(t_i^*-\tilde{t})}V_{i-1}(t_i^*;\overline{V}_0) + \frac{r+\lambda_1}{r}e^{-(r+\lambda_1)(t_i^*-\tilde{t})}\left(1 - e^{-r(\check{T}(t)-t_i^*)}\right)s \\ + \lambda_1 \int_{\tilde{t}}^{t_i^*} e^{-(r+\lambda_1)(\tau-\tilde{t})}\dot{V}_{i-1}(\tau;\overline{V}_0)\,d\tau\end{aligned}$$

for $\tilde{t} < t_i^*$. Hence, using $V_{i-1}(t_i^*;\overline{V}_0) = \frac{s}{\lambda_1} + \frac{s}{r}\left(1 - e^{-r(\check{T}(t)-t_i^*)}\right)$, one shows that $\lim_{\tilde{t}\uparrow t_i^*} \dot{V}_i(\tilde{t};\overline{V}_0) = -se^{-r(\check{T}(t)-t_i^*)} = \lim_{\tilde{t}\downarrow t_i^*} \dot{V}_i(\tilde{t};\overline{V}_0)$, implying that V_i is of class C^1.

It remains to prove that $\dot{V}_i(\tilde{t};\overline{V}_0) \leq -se^{-r(\check{T}(t)-\tilde{t})}$ for $\tilde{t} < t_i^*$. Yet, this is easily shown to follow from the fact that, by induction hypothesis, $\dot{V}_{i-1}(\tilde{t};\overline{V}_0) \leq -se^{-r(\check{T}(t)-\tilde{t})}$, and hence

$$\lambda_1 \int_{\tilde{t}}^{t_i^*} e^{-(r+\lambda_1)(\tau-\tilde{t})}\dot{V}_{i-1}(\tau;\overline{V}_0)\,d\tau \leq -se^{-r(\check{T}(t)-\tilde{t})}\left(1 - e^{-\lambda_1(t_i^*-\tilde{t})}\right),$$

which completes the induction step.

Now, consider some $i \in \{1, \cdots, m-1\}$. Having established that the agent's best response is given by a cutoff strategy, I shall now show that $t_{i+1}^* \leq t_i^*$. Consider an arbitrary time $\tilde{t} \geq t_i^*$, and

[10] Since I am looking at weak best responses here, $\hat{t}_1$ exists as a proper minimum. If $V_{i-1}(\tilde{t};\overline{V}_0) - V_i(\tilde{t};\overline{V}_0) = \frac{s}{\lambda_1}$, $\hat{t}_1 = \tilde{t}$.

suppose the agent still has $i+1$ breakthroughs to go. By stopping at an arbitrary time $t^* \in (\tilde{t}, \check{T}(t)]$, the agent can collect

$$\int_{\tilde{t}}^{t^*} \lambda_1 \frac{s}{r} e^{-(r+\lambda_1)(\tau-\tilde{t})} \left(1-e^{-r(\check{T}(t)-\tau)}\right) d\tau + \frac{s}{r} e^{-(r+\lambda_1)(t^*-\tilde{t})} \left(1-e^{-r(\check{T}(t)-t^*)}\right)$$

$$= \frac{s}{r}\left[\frac{\lambda_1}{\lambda_1+r}\left(1-e^{-(r+\lambda_1)(t^*-\tilde{t})}\right) - e^{-r(\check{T}(t)-\tilde{t})}\left(1-e^{-\lambda_1(t^*-\tilde{t})}\right)\right] + \frac{s}{r} e^{-(r+\lambda_1)(t^*-\tilde{t})}\left(1-e^{-r(\check{T}(t)-t^*)}\right).$$

By stopping immediately at time $\tilde{t}$, he can collect $\frac{s}{r}\left(1-e^{-r(\check{T}(t)-t)}\right)$. Thus, since

$$1-e^{-r(\check{T}(t)-t)} > \frac{\lambda_1}{\lambda_1+r}\left(1-e^{-(r+\lambda_1)(t^*-\tilde{t})}\right) - e^{-r(\check{T}(t)-\tilde{t})}\left(1-e^{-\lambda_1(t^*-\tilde{t})}\right) + e^{-(r+\lambda_1)(t^*-\tilde{t})}\left(1-e^{-r(\check{T}(t)-\tilde{t})}\right)$$

$$\iff 1 > \frac{\lambda_1}{r+\lambda_1} + \frac{r}{r+\lambda_1} e^{-(r+\lambda_1)(t^*-\tilde{t})},$$

the agent strictly prefers to stop immediately at $\tilde{t}$. For $\tilde{t} = t_i^*$ in particular, we can conclude that $t_{i+1}^* \leq t_i^*$; if $t_i^* > t$, we have that $t_{i+1}^* < t_i^*$.

Clearly V_1 is strictly increasing in $\overline{V}_0$ for all $\tilde{t} < t_1^* = \check{T}(t)$, with $\lim_{\overline{V}_0\to\infty} V_1(\tilde{t}; \overline{V}_0) = \infty$. A simple induction argument now establishes that V_i is strictly increasing in $\overline{V}_0$, with $\lim_{\overline{V}_0\to\infty} V_i(\tilde{t}; \overline{V}_0) = \infty$ for all $i = 1, \cdots, m$ whenever $\tilde{t} < t_i^*$.

Suppose $t_{i+1}^* > t$. Then, t_{i+1}^* is defined by $V_i(t_{i+1}^*; \overline{V}_0) = \frac{s}{\lambda_1} + \frac{s}{r}\left(1-e^{-r(\check{T}(t)-t_{i+1}^*)}\right)$. By our previous step, we furthermore know that $t_i^* > t_{i+1}^*$, implying V_i is strictly increasing in $\overline{V}_0$ at t_{i+1}^*. Hence, the cutoff t_{i+1}^* is strictly increasing in $\overline{V}_0$.

Now, suppose that $t_{i+1}^* = t$. Then, $V_i(t; \overline{V}_0) \leq \frac{s}{\lambda_1} + \frac{s}{r}\left(1-e^{-r(\check{T}(t)-t)}\right)$. Let $j := \min\{\iota \in \{1, \cdots, m\} : t_\iota^* = t\}$. Since $t_1^* = \check{T}(t) > t$, we have that $j \geq 2$. Now, $V_{j-1}(t; \overline{V}_0)$ is strictly increasing in $\overline{V}_0$ with $\lim_{\overline{V}_0\to\infty} V_{j-1}(t; \overline{V}_0) = \infty$. Hence, there exists a constant C_{j-1} such that for $\overline{V}_0 > C_{j-1}$, we have that $V_{j-1}(t; \overline{V}_0) > \frac{s}{\lambda_1} + \frac{s}{r}\left(1-e^{-r(\check{T}(t)-t)}\right)$, and hence $t_j^* > t$. Iterated application of this argument yields the existence of a constant C_i such that $\overline{V}_0 > C_i$ implies that $t_{i+1}^* > t$. Hence, by our previous step, t_{i+1}^* is strictly increasing in $\overline{V}_0$ for $\overline{V}_0 > C_i$.

Now, consider arbitrary $\tilde{t} \in [t, \check{T}(t))$ and $i \in \{1, \cdots, m\}$. Let σ be defined by $\sigma := \max\left\{\iota \in \{1, \cdots, m\} : t_\iota^* > \tilde{t}\right\}$. As $\tilde{t} < \check{T}(t) = t_1^*$, $\sigma \geq 1$. As $\tilde{t} < t_\sigma^*$, $V_\sigma(\tilde{t}; \overline{V}_0)$ is strictly increasing in $\overline{V}_0$ with $\lim_{\overline{V}_0\to\infty} V_\sigma(\tilde{t}; \overline{V}_0) = \infty$. Hence, there exists a constant $\tilde{C}_\sigma$ such that $\overline{V}_0 > \tilde{C}_\sigma$ implies $V_\sigma(\tilde{t}; \overline{V}_0) > \frac{s}{\lambda_1} + \frac{s}{r}\left(1-e^{-r(\check{T}(t)-\tilde{t})}\right)$, and hence $t_{\sigma+1}^* > \tilde{t}$. Iterated application of this argument yields the existence of a constant $\tilde{C}_{i-1}$ such that $\overline{V}_0 > \tilde{C}_{i-1}$ implies $t_i^* > \tilde{t}$. As $\tilde{t} \in [t, \check{T}(t))$ was arbitrary, we conclude that $\lim_{\overline{V}_0\to\infty} t_i^* = \check{T}(t)$, and that $\lim_{\overline{V}_0\to\infty} V_i(\tilde{t}; \overline{V}_0) = \infty$ for any $\tilde{t} \in [t, \check{T}(t))$.

For $\tilde{t} \geq t_i^*$, we have that $V_i(\tilde{t}; \overline{V}_0) = \frac{s}{r}\left(1-e^{-r(\check{T}(t)-\tilde{t})}\right) \leq V_{i-1}(\tilde{t}; \overline{V}_0)$. It remains to be shown that for $\tilde{t} < t_i^*$, $V_i(\tilde{t}; \overline{V}_0) < V_{i-1}(\tilde{t}; \overline{V}_0)$. Since V_{i-1} is strictly decreasing, we have that

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

$$\leq \frac{\lambda_1}{\lambda_1+r} V_{i-1}(\tilde{t}; \overline{V}_0)\left(1-e^{-(r+\lambda_1)(t_i^*-\tilde{t})}\right) + e^{-(r+\lambda_1)(t_i^*-\tilde{t})} \frac{s}{r}\left(1-e^{-r(\check{T}(t)-t_i^*)}\right).$$

Now, suppose that $V_i(\tilde{t};\overline{V}_0) \geq V_{i-1}(\tilde{t};\overline{V}_0)$. Then, the above inequality implies that

$$\left(\frac{r}{r+\lambda_1}+\frac{\lambda_1}{r+\lambda_1}e^{-(r+\lambda_1)(t_i^*-\tilde{t})}\right)V_i(\tilde{t};\overline{V}_0) \leq e^{-(r+\lambda_1)(t_i^*-\tilde{t})}\frac{s}{r}\left(1-e^{-r(\check{T}(t)-t_i^*)}\right).$$

Yet, as $V_i(\tilde{t};\overline{V}_0) \geq \frac{s}{r}\left(1-e^{-r(\check{T}(t)-\tilde{t})}\right) > \frac{s}{r}\left(1-e^{-r(\check{T}(t)-t_i^*)}\right)$, this implies

$$\frac{r}{r+\lambda_1}+\frac{\lambda_1}{r+\lambda_1}e^{-(r+\lambda_1)(t_i^*-\tilde{t})} < e^{-(r+\lambda_1)(t_i^*-\tilde{t})},$$

a contradiction. ▮

Proof of Lemma 4.4

Let t and $\hat{p}_t$ be given. The agent now chooses $\{k_{1,\tau}\}_{t<\tau\leq\check{T}(t)}$ (with $k_{1,\tau}\in[0,1]$ for all τ) so as to maximize

$$\int_t^{\check{T}(t)}\left\{re^{-r(\tau-t)-\lambda_1\int_t^\tau \hat{p}_\tau k_{1,\tau}\,d\tau}\left[(1-k_{1,t})s+k_{1,t}\lambda_1\hat{p}_t V_{m-1}(\tau;\overline{V}_0)\right]\right\}d\tau.$$

subject to $\dot{\hat{p}}_\tau = -\lambda_1 k_{1,t}\hat{p}_\tau(1-\hat{p}_\tau)$.

As in Section 5, it is again convenient to work with the log-likelihood ratio $x_\tau := \ln\left(\frac{1-\hat{p}_\tau}{\hat{p}_\tau}\right)$. Since $\dot{x}_\tau = \lambda_1 k_{1,\tau}$ (co-state variable μ_τ), the Hamiltonian for this problem is now given by

$$\tilde{\mathfrak{H}}_\tau = e^{-r(\tau-t)}\left[k_{1,\tau}\lambda_1 e^{-x_\tau}V_{m-1}(\tau;\overline{V}_0)+(1-k_{1,\tau})(1+e^{-x_\tau})s\right]+\mu_\tau\lambda_1 k_{1,\tau}.$$

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

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

Now, Pontryagin's conditions are given by $\mu_{\check{T}(t)} = 0$,

$$\dot{\mu}_\tau = e^{-r(\tau-t)-x_\tau}\left[k_{1,\tau}\lambda_1 V_{m-1}(\tau;\overline{V}_0)+(1-k_{1,\tau})s\right].$$

Furthermore, the agent prefers arm 1 over the safe arm at time τ if, and only if,

$$e^{-r(\tau-t)}\left[e^{-x_\tau}\lambda_1 V_{m-1}(\tau;\overline{V}_0)-(1+e^{-x_\tau})s\right] \geq -\mu_\tau\lambda_1. \tag{.6}$$

Since, by Lemma 4.3, V_{m-1} is strictly decreasing, $-\mu_\tau$ is bounded by

$$\begin{aligned}-\mu_\tau &\leq e^{-r(\tau-t)-x_\tau}\lambda_1\max\left\{\frac{s}{\lambda_1},V_{m-1}(\tau;\overline{V}_0)\right\}\int_\tau^{\check{T}(t)}e^{-(r+\lambda_1)(\sigma-\tau)}\,d\sigma\\ &< e^{-r(\tau-t)-x_\tau}\lambda_1\max\left\{\frac{s}{\lambda_1},V_{m-1}(\tau;\overline{V}_0)\right\}\int_\tau^{\infty}e^{-(r+\lambda_1)(\sigma-\tau)}\,d\sigma\end{aligned}$$

$$= \frac{\lambda_1}{r+\lambda_1} e^{-r(\tau-t)-x_\tau} \max\left\{\frac{s}{\lambda_1}, V_{m-1}(\tau;\overline{V}_0)\right\}.$$

Thus, assuming $V_{m-1}(\tau;\overline{V}_0) \geq \frac{s}{\lambda_1}$, the following condition is sufficient for (.6):

$$V_{m-1}(\tau;\overline{V}_0) \geq \frac{r+\lambda_1}{r\lambda_1}(1+e^{x_\tau})s = \frac{r+\lambda_1}{r\lambda_1}\frac{s}{\hat{p}_\tau}. \tag{.7}$$

We note that $V_{m-1}(\tau;\overline{V}_0) \geq \frac{s}{\lambda_1}$ is implied by (.7). Now, suppose $V_{m-1}(\tilde{t};\overline{V}_0) \geq \frac{s(r+\lambda_1)}{p_{\overline{T}} r\lambda_1}$ for some $\tilde{t} \in (t, \check{T}(t))$. Since, by Lemma 4.3, V_{m-1} is strictly decreasing, equation (.7) holds for all $\tau \leq \tilde{t}$. As equation (.7) also implies that $\tilde{t} < t^*_{m-1}$, and $t^*_{m-1} < t^*_i$ $(i = 1, \cdots, m-2)$ by Lemma 4.3, the claim follows. ∎

Proof of Lemma 4.5

By Lemma 4.3, we have that

$$f(\check{T}(t),\overline{V}_0) = \begin{cases} \int_t^{t^*_m} e^{-(r+\lambda_1)(\tau-t)}\lambda_1 V_{m-1}(\tau;\overline{V}_0)\,d\tau + e^{-(r+\lambda_1)(t^*_m-\tilde{t})}\frac{s}{r}\left(1-e^{-r(\check{T}(t)-t^*_m)}\right) & \text{if} \quad t < t^*_m \\ \frac{s}{r}\left(1-e^{-r(\check{T}(t)-\tilde{t})}\right) & \text{if} \quad t = t^*_m \end{cases}.$$

Differentiability of $f(.,\overline{V}_0)$ is thus immediate. When computing derivatives, we distinguish three different cases: First, suppose $t^*_m = t$. Then, it immediately follows from the explicit expression for f that

$$\frac{df}{d\check{T}(t)} = se^{-r(\check{T}(t)-t)}.$$

If $\check{T}(t) > t^*_m > t$, the Envelope Theorem implies that

$$\frac{df}{d\check{T}(t)} = \frac{\partial f}{\partial t^*_m}\frac{dt^*_m}{d\check{T}(t)} + \frac{\partial f}{\partial \check{T}(t)} = \frac{\partial f}{\partial \check{T}(t)} = se^{-r(\check{T}(t)-t)-\lambda_1(t^*_m-t)}.$$

If $t^*_m = \check{T}(t)$, i.e. $m = 1$, we get from the explicit expression for $f(\check{T}(t),\overline{V}_0) = V_1(t;\overline{V}_0;\check{T}(t))$ that

$$\frac{df}{d\check{T}(t)} = \lambda_1 e^{-(r+\lambda_1)(\check{T}(t)-t)}\left(\overline{V}_0 - \frac{s}{\lambda_1}\right) + se^{-r(\check{T}(t)-t)}.$$

$f(t;\overline{V}_0) = 0$ immediately follows from the fact that $V_m(\check{T}(t);\overline{V}_0) = 0$ for any $\check{T}(t) \in [t, \overline{T}]$. ∎

The Agent's Optimization Problem

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

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

with the state variables evolving according to $\dot{x}_t = \lambda_1 k_{1,t}$ (co-state μ_t) and $\dot{y}_t = -\lambda_0 k_{0,t} y_t$ (co-state γ_t), x_0 given, $y_0 = 1$, and x_T and y_T free, and $(k_{0,t}, k_{1,t}) \in \{(a,b) \in \mathbb{R}^2_+ : k_{0,t} + k_{1,t} \leq 1\} = \check{\mathcal{U}}$. Clearly, $\check{\mathcal{U}}$ is closed, bounded and convex , the set of admissible policies is non-empty, and the state variables are bounded. Using in addition the linearity of the objective and the laws of motion in the choice variables, one shows that the conditions of Filippov-Cesari's Existence Theorem (Thm. 8 in Seierstad & Sydsæter, 1987, p. 132) are satisfied.

To show sufficiency of Pontryagin's conditions, I invoke Arrow's Sufficiency Theorem (Thm. 5 in Seierstad & Sydsæter, 1987, p. 107). To do so, I define the new state variable $z_t := y_t e^{-x_t}$ (co-state ζ_t). It is straightforward to verify that $\dot{z}_t = -z_t\left(\lambda_0 k_{0,t} + \lambda_1 k_{1,t}\right)$. Now, $\mathfrak{H}_t$ can be written as

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

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

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

$$\mathfrak{H}_t = e^{-rt}(y_t + z_t)s,$$

i.e. $\mathfrak{H}_t$ is linear.

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

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

i.e. $\mathfrak{H}_t$ is linear again.

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

$$\mathfrak{H}_t = e^{-rt}(y_t + z_t)\lambda_0(h_t + \omega_t(z_t)) - \lambda_0(\gamma_t y_t + \zeta_t z_t).$$

Now, after plugging in $\omega_t(z_t) = \frac{z_t}{y_t + z_t} w_t + \hat{\delta}_t$,[11] we find that $\mathfrak{H}_t$ is again linear:

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

which thus completes the proof. ∎

[11] Here, I set $\omega_t =: w_t(p_t + \hat{\delta}_t)$. In Proposition 4.1, we have seen that for any $\epsilon > 0$, there exists a continuation scheme such that $p_t w_t \leq \omega_t \leq p_t w_t + \frac{s}{r}(1 - e^{-r\epsilon})$. Equivalently, we can think of the principal choosing $\hat{\delta}_t \in (0, L]$ such that $\omega_t(p_t) = w_t(p_t + \hat{\delta}_t)$. If the proof of Proposition 4.1 goes through with $m = 1$, $\hat{\delta}_t = 0$ can also be chosen.

Chapter 4: Expert Experimentation*

Nicolas Klein[†] Tymofiy Mylovanov[‡]

Abstract

We analyze a dynamic game between an expert and a decision maker. In each period, the decision maker has the option of buying cheap-talk advice from the expert, who is merely interested in his continued employment. The expert's quality is uncertain and is initially unknown to either of the parties. We characterize properties of the optimal decision rules, describing the effects of the expert's strategic reporting behavior. If the price of the expert's advice is low, the optimal decision rule calls for *more* employment as compared to the first best, and does so after *bad* news about the expert's quality. If the decision maker can commit to this rule, the expert reports his information truthfully. Yet even if the decision maker lacks commitment power, there exists an equilibrium in which the expert is always truthful when the price of advice is low. Yet, without commitment the decision maker may be better off in an alternative equilibrium in which the expert is allowed to accumulate some private information about his type.

Keywords: Reputational cheap talk, experimentation, optimal contract.

JEL Classification Numbers: C73, D83.

[*] We are grateful to Dirk Bergemann, Matthias Fahn, Lucas Maestri, Stephen Morris, Sven Rady, Phil Reny, as well as seminar audiences at Bonn, Moscow, and Paris for useful comments and discussions. The first author is particularly grateful to the Department of Economics at Yale University and the Cowles Foundation for Research in Economics for an extended stay, during which this paper took shape. Financial support from the National Research Fund, Luxembourg, and the Deutsche Forschungsgemeinschaft through SFB TR 15 is gratefully acknowledged.

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

[‡] Department of Economics, Penn State University; email: txm41@psu.edu.

1 Introduction

This paper studies a model of a decision maker (she) experimenting with employing an expert (he) whose quality is uncertain and is unknown to either of them. The expert is strategic and is only interested in his continued employment. The core issue in the model is the following: If the expert were not strategic, the decision maker would like to employ the expert if and only if the expert continued to prove his competence. Yet, if this decision rule were in fact followed, the expert might have incentives to suppress a priori unlikely information to maximize his chances of appearing competent; this would slow down learning about the quality of the expert. In this paper, we are interested in the effect the expert's concern for reputation has on the structure of the optimal decision rules and, in particular, on the amount of experimentation by the decision maker.

To motivate the question, think of a legislator, for instance, who wants to find out whether the people she represents would prefer a vote in favor of, or against, a particular bill under consideration. Knowing the ideological bent of her district, she would have some initial belief on which position her constituents would tend to favor. Still, she might want to hire a pollster, the quality of whose analyses she will only gradually learn to ascertain over time. While, initially, she may trust her own instincts more than the pollster's judgment, she may still find it worthwhile experimenting with a particular polling firm, as information on how good they are may come in handy in subsequent decision problems.[1]

We investigate how this trade-off between experimentation and exploitation is affected by the presence of incentive constraints, or, equivalently, by a risky bandit arm which now decides strategically whether to release the information it produces. Indeed, concerned about his reputation, the pollster, unsure of how well-suited his methodology is to the district in question, may prefer to base his recommendation on the district's ideological tilt, which is common knowledge, rather than on his actual polling, thus incidentally preventing the legislator from learning about the accuracy of his polling methods. It is the goal of this paper to analyze the effects of the dynamic interplay of the experimentation motive on the one hand, and the presence of agency costs arising from the need to provide incentives for

[1] More broadly, the model applies to many instances when a decider who faces a sequence of multiple decision problems can either follow the information she already has, or else can elect to seek the help of outside advisors. Thus, a firm trying to gauge the demand for a new product may either trust in its own prior belief, or seek the advice of outside consultants. People filing their income tax returns may either follow their own information when trying to figure out if a particular set of expenses is deductible, or hire a tax consultant, whose help may be very valuable in years to come should he turn out to be competent. Or a college student having set aside an afternoon to study for a test may do so by herself, or spend the time with a tutor; even if she may not initially trust the tutor's advice, he might over time turn out to be quite competent, thus helping her improve her performance on future tests.

truth-telling, on the other hand.

While one's first intuition may be that the introduction of additional agency costs associated with pulling the risky arm would render the decision maker more reluctant to pull her risky arm, our analysis shows that this need not be so. Indeed, if agency costs are relatively low, the reverse happens. Now, there is optimally more experimentation with agency costs than in the first best without agency costs, and the distortion only occurs after *bad* news, while after good news, the decision maker continues to behave as in the first best (Proposition 4.7).

By the Revelation Principle, the expert will always report his information truthfully in the optimal decision rule. To ensure truth-telling, the decision maker has two conceptually distinct tools at her disposal: She could penalize the expert after he has correctly announced the state that was initially more likely in the first place, or she could reward him even if his minority opinion turns out to be incorrect. With agency costs low, future option values are quite high, and thus, incentive costs are minimized by encouraging the expert to venture a minority opinion by the decision maker's continuing to employ him with some probability even if his information turned out to have been mistaken. This tool crucially depends, however, on the decision maker's enjoying full commitment power; indeed, without commitment, she will only have the former tool at her disposal.

Regarding the concrete implementation of incentives in an optimal decision rule, we can show that the threat of firing dominates mere temporary suspension (Lemma 4.5). Furthermore, firing right away dominates firing later (Lemma 4.6); this is obvious in the case where the expert has revealed himself to be of the bad type. We show that it also holds true if firing has to occur in the context of punishment for the expert's sticking to the prevailing majority opinion.

Our model belongs to the class of 'bad reputation' models: the expert's concern about appearing to be competent creates incentives for the expert to distort his reports. Ottaviani and Sørensen (2006a, 2006b) explore the effect of this type of reputation concerns on the reporting strategy of the expert in *a single-decision environment* and show that, under weak conditions, truthful reporting is not feasible; these conditions are satisfied by the static version of our model.[2]

By contrast, we show that all agency problems vanish, and the first-best policy becomes incentive compatible, as the reputation concerns grow sufficiently high, which we formalize as the total number of periods (Proposition 4.2). While this result, which is independent of the

[2] Similar results are obtained in Morris (2001) and Morgan & Stocken (2003) who study cheap talk models in which (bad) reputation concerns are about the preferences of the expert rather than his competence.

decision maker's commitment possibilities, may seem reminiscent of familiar folk theorems, the fundamental mechanism underlying it is completely different and is primarily driven by the fact that the expert is conditioning the expectation of his future continuation payoff on different events on and off the equilibrium path. This continuation payoff effect is absent in single-decision environments.

Furthermore, even in the presence of agency problems, truthful reporting of the expert's information can be supported in equilibrium. Of course, if the decision maker can commit to a decision rule, this is implied by the Revelation Principle. Yet even without commitment there exists an equilibrium in which the expert reports his information truthfully if the price of advice is low (Proposition 5.1). This equilibrium is non-Markovian and conditions the expert's retainment both on the belief about his competence as well as on the payoff-irrelevant features of the reporting history.

However, truthtelling should not be an end in itself for the decision maker. Without commitment power, the decision maker can be better off in an equilibrium in which the expert babbles in the first periods and accumulates private information about his type (Proposition 5.5). The reason is that implementing truthtelling in early periods requires firing a competent expert with non-negligible probability; this could be excessively costly for the decision maker.

The rest of this paper is structured as follows: Section 2 reviews some related literature; Section 3 presents the model, and introduces necessary notation; in Section 4, we analyze the game under full commitment, whereas Section 5 analyzes the case of a decision maker who cannot commit to a specific decision rule at the outset of the game; Section 6 concludes. Most proofs are provided in the appendix.

2 Related Literature

Our decision maker's problem is akin to that of an economic agent operating a two-armed bandit machine with one safe arm, whose expected payoff is known, and one risky arm, whose expected payoff is initially unknown but can potentially be learnt through use over time. Bandit models have been used in economics, at least since Rothschild's (1974) discrete-time model, to analyze the trade-off between experimentation and exploitation, providing insights into agents' incentives to forgo current payoffs in exchange for information which can potentially be parlayed into better decisions, and thus higher payoffs, come tomorrow.[3] The same trade-off is central to our paper, with the additional twist that our "risky arm" is acting strategically, in the sense that the expert will choose his messages with a view to

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

maximizing his expected duration of employment.

The single-agent two-armed bandit decision problem has been analyzed as early as the 1950s (cf. e.g. Bradt, Johnson, Karlin, 1956). The analysis of the strategic interplay between several agents operating replica bandits (cf. Bolton & Harris, 1999, 2000, Keller, Rady, Cripps, 2005, Keller & Rady, 2010), negatively correlated bandits (cf. Chapter 1), or three-armed bandits (cf. Chapter 2), is much more recent, by contrast.

"Strategic risky arms", in the sense that the party about whom information is gathered is also taking an action, have been studied by Bergemann & Välimäki (1996, 2000) and Bar-Isaac (2003). Bergemann & Välimäki (1996) investigate the case of a *single* buyer desiring to purchase a good from any one of multiple firms, which produce at different, and initially unknown, qualities. They show that, with firms pricing the good strategically, experimentation is at efficient levels in any Markov perfect equilibrium. In Bergemann & Välimäki (2000), by contrast, they show that with multiple buyers and two firms, the firms are subsidizing experimentation to the point that it reaches inefficiently high levels in equilibrium. In Bar-Isaac (2003), a seller of a good of initially unknown quality faces a sequence of buyer pairs, who, in each period the seller (strategically) decides to sell, Bertrand-compete for the good. After the transaction, the quality of the good is publicly observed, and beliefs about the seller's quality are updated. In contrast to these papers, we investigate the role of long-term contracts an experimenter could potentially offer the "risky arm" to induce behavior conducive to the experimenter's interests. Moreover, in the previously mentioned papers, the risky arm is deciding on a costly action—i.e. players are engaged in a signaling, rather than a cheap-talk, game.

In a simple two-period Moral Hazard model, Manso (2010) shows that in order to induce the agent to eschew an undesirable action, the principal will be very tolerant of, and may even reward, early failure. Chapter 3 also shows that incentives are best provided through later continuation values; moreover, he shows that if the optimal incentive scheme is run there is no deadweight loss from agency. In our present setting, by contrast, the provision of incentives through continuation values will lead to a distortion of the first-best rule.

Our model is also related to the literature on dynamic agency under gradual learning, which has e.g. been investigated by Bergemann & Hege (1998, 2005) and Hörner & Samuelson (2009), who examine a venture capitalist's optimal provision of funds to an entrepreneur, who might divert the funds toward his private ends. As is the case in our model, the accumulation of private information by an agent subsequent to a deviation from the equilibrium path provides an interesting dynamic aspect to the strategic interaction. However, our model is probably closest to Gerardi & Maestri (2008) who analyze the case of an expert who, in order to receive an informative signal about the decision-relevant state of nature, has to incur some private effort costs. After choosing whether or not to expend effort on acquiring the signal, the expert then sends the decision maker a cheap-talk message about the state of the world. Our model crucially differs from all these papers, though, in that we are investigating a problem of *adverse selection*, or *hidden information*, rather than *moral hazard*: Our expert privately observes a signal without having to incur any effort costs; it is rather his innate quality, which determines the precision of his signal, which is uncertain at first.[4] Over the course of the interaction, the parties gradually update their beliefs about the expert's type; as he wants to continue to be thought of highly in order to be re-employed, our expert will have incentives strategically to manipulate the decision maker's learning process. To the best of our knowledge, our paper is the first to investigate the dynamic interplay between a decision maker's reliance on cheap-talk advice and her desire to screen out advisors of poor quality.

The problem of testing, and discriminating between, experts of different, and uncertain, quality has recently received considerable attention in the literature. Many contributions stress the impossibility of fully discriminating between an expert who knows the stochastic process about which he is called on to make predictions, and false experts who just make wild guesses, see e.g. Olszewski & Sandroni (2008). Stewart (2009) relaxes the strong requirement generally made in this literature that good experts must *never* be mistaken for bad ones, constructing a test that can distinguish experts' types with high probability, while Olszewski & Peski (2009) construct a Bayesian test which allows the principal under certain circumstances to overcome any potential incentive problems stemming from the informational asymmetry concerning the expert's quality.

As our expert wants to be employed as long as possible, he cares about his reputation, i.e. the principal's belief that he is of the good type. Building on the seminal contribution by Crawford & Sobel (1982), several papers have researched the implications of experts' reputational concerns in cheap-talk games: Morris (2001), e.g., studies the effects of social norms of political correctness on an advisor's candor; in Ottaviani & Sørensen (2006), an expert's perceived competence directly enters his utility function, whereas Fong (2009) studies optimal evaluation rules for surgeons.

[4] While initially and on the equilibrium path, there is incomplete but symmetric information as to the expert's quality, private information may arise endogenously off the equilibrium path. Moreover, there is obviously asymmetric information concerning the realized signals.

3 Model Setup

There is a decision maker and an expert who live for $N + 1$ periods, $t = 0, 1, \ldots, N$.[5] In each period, the decision maker has to choose a policy from the set $\{0, 1\}$. The optimal policy is uncertain and is described by the state $\omega_t \in \{0, 1\}$. The value of ω_t is distributed independently across periods and is equal to 0 with a probability of $p_0 \in (0, 1/2)$, which is common knowledge. The decision maker obtains a period payoff of 1 whenever the policy matches the state and 0 otherwise; there is no discounting.

In each period, the decision maker can either consult an expert or she can simply rely on her prior to make the decision. Our goal is to study the effect of the expert's strategic reporting behavior on an optimal decision rule in an environment where there is a tradeoff between an immediate cost of employing the expert and the future benefit of learning about the expert's competence. To this end, we assume some exogenous cost $c > 0$ that is incurred by the decision maker whenever she employs the expert.

If the decision maker decides to consult the expert, the expert observes a non-verifiable signal, $\tilde{s} \in \{0, 1\}$, about the realized state. There are two types of expert: With probability $\alpha_0 \in (0, 1)$ the expert is competent and his signal reveals the optimal policy. With complementary probability, the expert is incompetent and his signals are uninformative. We denote the type of the expert by $\theta \in \{0, 1\}$ where $\theta = 1$ denotes the competent expert.

This assumption implies that the decision maker will learn with certainty that the expert is incompetent whenever he is expected to tell the truth and makes a mistake. Similar full-revelation assumptions are commonly made for reasons of tractability, e.g. in the strategic experimentation literature, where it is often assumed that one instance of good (or bad) news resolves all uncertainty about the value of the risky arms.[6] It is also concordant with the assumptions often made in the literature on repeated games with reputation and perfect monitoring, where a single deviation from the behavior expected from a Stackelberg leader reveals that the player is strategic.[7]

The type of the expert is uncertain and not known to either of the players. The expected probability that the expert's signal is correct at time $t = 0$ is $\beta_0 = \alpha_0 + (1 - \alpha_0)/2$. The belief of the decision maker that the expert is competent given the information available in

[5] Alteratively, we could assume infinitely many periods with discounting. Yet, end-of-game effects are not much of a concern in our model, because the uncertainty about the expert's quality will gradually resolve over time, so that players' preferences will eventually become aligned. The same would be true in a model with infinitely many periods, so that we would expect similar qualitative results.

[6] See e.g. Keller, Rady, Cripps (2005) or Chapter 1 of this dissertation.

[7] See Mailath & Samuelson (2006) for a review of the (non-cheap talk) literature on repeated games with reputations.

period t is denoted by α_t and is called the *reputation* of the expert. We use β_t to denote the probability that the expert's signal is correct in period t.

We introduce *reputational concerns* on the expert's part into our dynamic setting by assuming it is the expert's objective to be employed for as many periods as possible. Hence, we assume that the expert gets a payoff of 1 per period whenever he is employed and 0 otherwise. Again, there is no discounting.

To rule out uninteresting cases, we assume that it is commonly believed that $\beta_0 < 1-p_0$; i.e. the decision maker obtains a higher payoff if she follows her prior beliefs than if she follows the signals of the expert with reputation α_0. Thus, employing the expert in the early periods of the game entails a cost of c per period while providing the decision maker with an opportunity to learn about his type and to screen out a bad expert. Our assumption also implies that an expert with a reputation of α_0 will believe that state 1 is more likely to occur regardless of his signal. This can create an incentive for the expert to lie about his signal, thus impeding learning about his type.

The timing of the interaction in each period is as follows. First, the decision maker decides whether to hire the expert. If he is employed, the expert then observes a signal and sends a subsequent cheap-talk report to the decision maker, after which the decision maker chooses a policy. Then, at the end of the period, the actual state of the world is publicly observed, and payoffs are realized.

We will consider two variants of the model. In the game with full commitment, the decision maker can commit to her strategy at the outset of the game in period 0. In the game without commitment, the decision maker has no commitment power and her choices must be sequentially rational in each period. Our solution concept is perfect Bayesian equilibrium.

Whether the expert is employed or not, the signal reported by the expert (if any) and the realized state are common knowledge in each period. Hence, the public history at *the beginning* of period $t \geq 1$, denoted by h_t^p, is a sequence of t triples $\nu_\tau^p = (\varphi_\tau, \hat{s}_\tau, \omega_\tau)$, $\tau = 0, \dots, t-1$, where we set $\varphi_\tau = 1$ if the expert is consulted in period τ and $\varphi_\tau = 0$ otherwise, and where $\hat{s}_\tau$ denotes the expert's message in period τ.[8] We extend our definition of a public history to $t = 0$ by setting $h_0^p = \emptyset$. In addition, we will use the notation $h_{t+1}^p = (\nu_t^p, h_t^p)$ and $h_{t+1}^p = (\varphi_t, \hat{s}_t, \omega_t, h_t^p)$. Also, let $h_{t+\tau}^p = h_t^p h_\tau^p$ denote a public history of length $t+\tau$ in which the first t periods are described by history h_t and the last τ periods are described by history h_τ. We set $h_t^p = h_t^p \emptyset$.

Because he privately observes his signals, the expert's information partition is finer.

[8] We set $\hat{s}_\tau = \emptyset$ if the expert is not consulted in period τ.

Therefore, the private history of the expert at the beginning of period $t \geq 1$, denoted by h_t^a, is a sequence of t four-tuples $\nu_\tau^a = (\varphi_\tau, \tilde{s}_\tau, \hat{s}_\tau, \omega_\tau)$.[9] We define $h_0^a = \emptyset$. As in the case of public histories, we will use the notation $h_{t+1}^a = (\nu_t^a, h_t^p)$, $h_{t+1}^a = (\varphi_t, \tilde{s}_t, \hat{s}_t, \omega_t, h_t^a)$, and $h_{t+\tau}^a = h_\tau^a h_t^a$.

Let $\mathcal{H}^p$ and $\mathcal{H}^a$ respectively denote the set of all public and private histories. In addition, we define the set of private histories in which the expert reports his signals truthfully in all previous periods:

$$\tilde{\mathcal{H}}^a := \{h_t^a \in \mathcal{H}^a | \hat{s}_\tau = \tilde{s}_\tau, \tau \leq t\}.$$

$\mathcal{H}^p$ is partitioned by $\mathcal{H}_0^p$, the subset of histories in which there is a report inconsistent with the state, and $\mathcal{H}_1^p$, the subset of histories in which all reports coincide with the realized states.

In our environment, the decision maker faces two objectives. The first objective is the choice of an optimal policy in each period given the available information. The second objective is choosing optimally when to consult the expert given the history of his performance.

Achieving the first objective is straightforward and will not be the focus of our analysis: If the expert is employed, the decision maker will follow her recommendation if and only if it is sufficiently informative. Otherwise, the decision maker will comply with her prior beliefs and choose policy 1. If the expert is not consulted, the decision maker will follow her prior belief.

The focus of our analysis is the second objective of the decision maker. Therefore, our definition of the decision maker's rule only considers whether the expert will be employed in a given period. Specifically, a (behavioral) *decision rule* is a function mapping the public history in period t into the probability of consulting the expert in this period:

$$\rho : \mathcal{H}^p \to [0, 1].$$

Every decision rule induces a decision problem for the expert, in which the expert maximizes his expected payoff. A (behavioral) strategy of the expert is a function that maps the private history in period t and the observed signal into a distribution of reports to the decision maker:

$$\hat{s} : \mathcal{H}^a \times \{0, 1\} \to [0, 1], \quad (h_t^a, \tilde{s}_t) \mapsto \hat{s}(h_t^a, \tilde{s}_t),$$

where we set $\hat{s}(h_t^a, \tilde{s}_t)$ to be the probability of report 1. A strategy is *truthful* if

$$\hat{s}(h_t^a, \tilde{s}_t) = \begin{cases} 1, & \text{if } \tilde{s}_t = 1; \\ 0, & \text{otherwise.} \end{cases}$$

[9] We set $\tilde{s}_\tau = \hat{s}_\tau = \emptyset$ if the expert is not consulted in period τ.

We denote the truthful strategy by $\hat{s}^*$. Let $\mathcal{S}$ be the set of all strategies $\hat{s}$.

Given decision rule ρ, let $u^\rho(h_t^a)$ denote the expert's expected payoff in period t after the private history h_t^a, conditional on being employed in this period. Similarly, let $u^\rho(\hat{s}, \tilde{s}, h_t^a)$ denote the expert's expected payoff given ρ in period t after private history h_t^a conditional on being employed in this period and after observing signal $\tilde{s}$ and using a strategy $\hat{s}$. A decision rule ρ is *incentive compatible after the signal* s if it is optimal for the expert to report his signals truthfully given that he did not lie in any of the preceding periods:

$$u^\rho(\hat{s}^*, s, h_t^a) \geq u^\rho(\hat{s}', s, h_t^a), \text{ for all } h_t^a \in \tilde{\mathcal{H}}^a, \hat{s}' \in \mathcal{S}, s \in \{0,1\} \tag{1}$$

Let ρ be an incentive compatible decision rule. Then, on the equilibrium path after any history in $\mathcal{H}_\emptyset^p$ the public posterior belief is that the expert is certainly incompetent, $\alpha_t^p = 0$. By contrast, after a history in $\mathcal{H}_1^p$, $t \geq 1$, the public posterior belief about the competence of the expert increases in the number of periods

$$\alpha_t^p = \frac{\alpha_0}{\alpha_0 + (1-\alpha_0)\frac{1}{2^t}} = \frac{\alpha_{t-1}}{\alpha_{t-1} + (1-\alpha_{t-1})\frac{1}{2}} \text{ for any history } h_t^p \in \mathcal{H}_1^p.$$

If $\alpha_t \geq 1 - 2p_0$ or, equivalently,

$$t \geq K := \log_2 \frac{1-2p_0}{2p_0}\frac{1-\alpha_0}{\alpha_0}.$$

the expert's signals become sufficiently informative for the decision maker to match her action to the signal. Let K_0 be the largest integer that is smaller than K. As $\beta_0 < 1 - p_0$, the value of $K_0 \geq 0$.

4 Optimal Decision Rule

Throughout this section, we shall assume that the decision maker can commit to whatever decision rule she would like to at the outset of the game.

Let $v^\rho(h_t^p)$ denote the expected payoff of the decision maker at the beginning of period t given history h_t^p and decision rule ρ, from which we subtract the payoff she can obtain if she always relies on her own information, $(N+1-t)(1-p_0)$, by way of normalization. A decision rule is *optimal* if it maximizes the ex-ante expected payoff of the decision maker among all incentive compatible decision rules:

$$\begin{aligned}(P_0): \quad & \max v^\rho(\emptyset) \\ & \text{s.t. } (1).\end{aligned}$$

In the game with full commitment, the Revelation Principle implies that we can without loss of generality restrict our attention to incentive compatible rules. Any optimal decision rule is an equilibrium outcome in the game with full commitment; an optimal decision rule exists because the set of incentive compatible decision rules is compact in the weak topology and v^{ρ} is continuous on this set.

We denote by $V(\alpha)$ the ex-ante expected payoff of the decision maker in an optimal decision rule where we make explicit that it depends on the prior reputation of the expert. The value of $V(\alpha)$ is non-negative for any $\alpha \in [0,1]$ because the decision maker can always ensure the payoff of 0 by not employing the expert and following her prior. Furthermore, $V(\alpha)$ is positive and increasing in α if the expert is employed in the optimal decision rule.

Consider a hypothetical environment in which the expert's signals are observed by the decision maker. In this environment, the (*first-best*) optimal decision rule is either never to buy the signal, and thus to realize a payoff of 0, or to continue buying the signal as long as it has been correct and to stop doing so after a first incorrect forecast. In the latter case, this rule yields a payoff of

$$\begin{aligned} v^{FB}(\alpha_0, N+1) &= \alpha_0 \left[(N-K)p_0 - (N+1)c\right] \\ &+ (1-\alpha_0)\left\{\left(\tfrac{1}{2}\right)^K \left[1 + \tfrac{1}{2} + \cdots + \left(\tfrac{1}{2}\right)^{N-K-1}\right]\left(\tfrac{1}{2} - (1-p_0)\right) - \left[1 + \tfrac{1}{2} + \cdots + \left(\tfrac{1}{2}\right)^N\right] c\right\} \end{aligned}$$

In order to avoid the uninteresting case in which buying the signal would not be optimal even in a first-best world, we impose the following

Assumption 4.1

$$v^{FB}(\alpha_0, N+1) > 0.$$

The agency problem in our model arises because the first best decision rule might not be incentive compatible if the expert's signals are not observable. Let, for instance, $N = 2$ and imagine that the expert observes $\tilde{s}_0 = 0$ in period 0. Then, the expert believes that the state $\omega_0 = 1$ is more likely because his signals are not sufficiently informative to outweigh his prior beliefs.[10] Thus, the probability of employment in the second period is maximized by reporting $\hat{s}_0 = 1$. As a result, the expert's best response to the first best decision rule would entail "babbling" in period 0, i.e., for instance, a report of 1 irrespective of the observed signal. Thus, there is no learning about the expert's type, and the decision maker would find herself compelled to revert to her prior in either period, thus realizing her outside option of 0.

[10] This is so because $p_0\beta_0 < (1-p_0)(1-\beta_0)$.

Nevertheless, the first best decision rule becomes incentive compatible if the duration of the interaction between the expert and the decision maker is sufficiently large. This result is based on the observation that, as N increases, the rate of growth of the payoff from telling the truth in any given period t exceeds the rate of growth of the payoff from lying in that period.

Although there is an obvious analogy, the proof is not a folk theorem type of argument. First of all, there is no discounting in our environment and the number of periods is finite. More importantly, the reason behind the different rates of growth of the payoffs from lying and from telling the truth is that the two parties are conditioning on different events. If the expert lies and his report turns out to be correct, he privately learns that he is incompetent. By contrast, if he reports his signal truthfully and it is correct, then the expert believes that he is more likely to be competent. While deciding about his report, the expert's expectation about how likely he is to *continue* to be correct in the future if retained is, therefore, conditioned on two distinct events. This conditioning is the reason why the rates of growth of the payoffs from lying and telling the truth differ.

Proposition 4.2 (Vanishing Concerns for Reputation) *For any given c, p_0, α_0, there exists an integer N_0 such that the first best decision rule is incentive compatible if and only if $N \geq N_0$.*

Proof: A formal version of the argument expounded above proves that for any t there exists an integer $N'(t)$ such that for all $N \geq N'(t)$ there is no profitable (possibly, multi-period) deviation from truth-telling that starts in period t. It is left to show, then, that there exists an N_0 such that $N(t) \leq N_0$ for all t or, in other words, that as we increase N the incentive constraints are not violated in the newly added periods. This, however, holds true because, if the expert is employed toward the end of the relationship under the first best rule, then his reputation is necessarily high, the expert considers his signals very informative, and truth-telling is his strict best response. A complete proof is provided in the appendix. ∎

If $N < N_0$, the first best decision rule is not incentive compatible. The decision maker will want to fire the expert if he reveals himself to be incompetent, which can potentially create incentives for the expert to report a signal of 1 more frequently than he observes it. Thus, incentive compatibility requires adjusting the decision rule to make reporting 0 relatively more attractive. *A priori*, there are two options: To encourage reports that go against the grain of received prior wisdom, one may want to promise the expert that he will be retained (with some probability) after such a report even if it does not match the state. An alternative is to discourage sticking with the majority opinion by promising to fire the expert (again, with some probability) after such a report even if the report turns out to be correct in the end. Whether the decision maker is better off relying on the carrot or on the stick will depend on the parameters; in the latter case, she will under-employ the expert as compared to the first-best benchmark, whereas, in the former case, the expert will be over-employed, an option, which, as we shall see in Section 5, crucially relies on the assumption that the decision maker has full commitment power.

In order to characterize certain properties of the optimal decision rule, we proceed in several steps, progressively restricting the class of decision rules the decision maker will want to consider. Our first lemma shows that there is no gain for the decision maker in conditioning on the different histories that may arise after the expert has made a mistake. The intuition for this is that, as long as the players are on the equilibrium path, both will know that the expert is of the low type for sure after he has made a mistake; conditioning the rule in the manner proposed could thus only be leveraged by a privately informed expert, who, by lying, had erroneously convinced the decision maker that he was no good. Affording off-equilibrium experts who lied such opportunities can only encourage lying, and therefore cannot be in the decision maker's best interests.

Lemma 4.3 *Within the class of incentive compatible decision rules, the decision maker can without loss restrict herself to rules that do not condition on the continuation histories that occur after the expert makes an incorrect report.*

Proof: Let ρ be an incentive compatible decision rule and $h^p \in \mathcal{H}_\emptyset$ be a public history after which the decision rule retains the expert with positive probability that depends on the continuation history. Then, consider a decision rule ρ' that (i) coincides with ρ everywhere except h^p and (ii) after history h^p retains the expert for the same expected number of periods as ρ regardless of the continuation history, where the expectation is taken with respect to a probability distribution induced by truthful reporting of the incompetent expert.

By construction, ρ' is incentive compatible at all private histories that induce h^p and their successors. Furthermore, the expert's payoff along the equilibrium path is the same in both rules at all other private histories. Furthermore, off the equilibrium path, the expert's payoff is affected only for the private histories that are predecessors and successors of private histories that result in public history h^p and for which at the node corresponding to h^p the expert believes that he is competent with a positive probability.

Let h^a be such a history. In ρ', the expert's expected payoff at h^a is constant in his private belief and coincides with that of the incompetent expert. By contrast, in ρ, the expert's payoff at h^a is non-decreasing in his private belief and is not lower than that of the incompetent expert; the expert can guarantee himself the expected payoff of an incompetent expert by using the latter's strategy conditioning, if necessary, his reports on a coin flip rather than his signals. Hence, the expert's payoff at history h^a in ρ' is not greater than in ρ. Furthermore, by construction of ρ', truth-telling is a best response in this rule for any successor history of h^a regardless of the continuation history. It follows then that ρ' is incentive compatible.

By construction, both decision rules induce the same expected payoff for the decision maker.

Iterated application of the preceding construction proves the claim. ▮

Our next observation is that we can focus on the decision rules that do not condition on the realization of the state in periods in which the expert is not employed. Let $h^0_\emptyset$ and $h^1_\emptyset$ denote one period histories in which the expert is not employed and the state is 0 and 1 respectively.

Lemma 4.4 *Within the class of incentive compatible decision rules, the decision maker can without loss restrict herself to rules that satisfy*

$$\hat{\rho}_i(h^p_t h^0_\emptyset h^p_\tau) = \hat{\rho}_i(h^p_t h^1_\emptyset h^p_\tau) \text{ for any } h^p_{t'}, h^p_\tau \in \mathcal{H}^p, \tag{2}$$

where $t + \tau + 1 \leq N$.

Proof: Consider any incentive compatible decision rule ρ. We will construct, by induction, another incentive compatible decision rule $\hat{\rho}$.

Step 1. Set $\hat{\rho}_0 = \rho$.

Step 2. For any $i = 1, \cdots, N$ let $t' = N - i$ and define $\hat{\rho}_i$ as follows.

1. $\hat{\rho}_i(h^p_t) = \hat{\rho}_{i-1}(h^p_t)$ for all histories of length less than or equal to t' ($t \leq t'$) and for histories of any length $t > t'$ in which the expert is employed in period $t'+1$ ($\phi_{t'+1} = 1$);
2. $\hat{\rho}_i(h^p_{t'} h^0_\emptyset h^p_\tau) = \hat{\rho}_i(h^p_{t'} h^1_\emptyset h^p_\tau) := p_0 \hat{\rho}_{i-1}(h^p_{t'} h^0_\emptyset h^p_\tau) + (1-p_0)\hat{\rho}_{i-1}(h^p_{t'} h^1_\emptyset h^p_\tau)$ for any $h^p_{t'}, h^p_\tau \in \mathcal{H}^p$, where $\tau \leq N - t' - 1$.

Define $\hat{\rho} = \hat{\rho}_N$. Decision rule $\hat{\rho}$ is incentive compatible. To see this, note that $\hat{\rho}_0$ is trivially incentive compatible as it coincides with ρ. Furthermore, if $\hat{\rho}_{i-1}$ is incentive compatible, then $\hat{\rho}_i$ is also incentive compatible: This follows from the following two observations: (a)

The expert who finds it optimal to report truthfully after history $h^p_{t'}h^0_\emptyset h^p_\tau$ and after history $h^p_{t'}h^1_\emptyset h^p_\tau$ in $\hat{\rho}_{i-1}$ will also find it optimal to report truthfully after either of these histories in $\hat{\rho}_i$, as beliefs are the same after both histories. (b) Moreover, the expert's expected payoff in period t' after histories considered in step 2.2 of construction is the same in both rules.

By construction, $\hat{\rho}$ satisfies (2) and generates the same expected payoff for the decision maker as ρ. ∎

Let $\tilde{C}$ denote the set of incentive compatible decision rules that satisfy the conditions of Lemma 4.3 and Lemma 4.4. Our next lemma shows that when the decision maker wants to penalize the expert even though his forecast had been correct, it is best to fire the expert for good with a certain probability rather than to suspend him for a limited amount of time. Its proof formalizes the intuition that the decision maker can always exactly reproduce the incentives provided by the threat of temporary suspension with the threat of complete termination with the appropriate probability. Indeed, termination is a much sharper and fearsome tool; moreover, since we allow for stochastic decision rules, the decision maker can exactly fine-tune the dosage of this threat as she needs to in any given situation. While our lemma only proves that the decision maker can always do at least as well by threatening to fire the expert as by temporarily suspending him, she may often be able to do strictly better with the sharper tool, as it can enforce compliance where the mere threat of temporary suspension may fail.

Lemma 4.5 (No Temporary Suspensions) *Within $\tilde{C}$, the decision maker can without loss neglect such rules $\hat{\rho}$ for which there exist a time period $\tilde{t}$ and a history $\tilde{h}^p_{\tilde{t}}$ such that $\hat{\rho}(\tilde{h}^p_{\tilde{t}}) < 1$ and $\hat{\rho}(0, \emptyset, \omega_t, \tilde{h}^p_{\tilde{t}}) > 0$.*

Proof: The idea of the proof is to construct a decision rule that instead of suspending the expert in period $\tilde{t}$ and employing him in the next period employs the expert in period $\tilde{t}$ and suspends him in the next period. As the actual time of employment of the expert is payoff irrelevant for either of the parties, the constructed decision rule is incentive compatible and generates the same expected payoff for the decision maker. The full proof is provided in the appendix. ∎

Let $\tilde{C}'$ be the subset of the decision rules in $\tilde{C}$ that satisfy the conditions of Lemma 4.5. Our next lemma uses a somewhat related argument: The decision maker can never gain by waiting to exact a penalty that is due. For one thing, the expert will with a positive probability have revealed himself to be of poor quality, and hence have gotten fired, before the "penalty period"; therefore, if the firing is to take place later, it must occur with a higher probability to give the same incentives. Conversely, conditional on the expert's reaching the "penalty period", his expected ability is higher than it is today, and hence firing him with the *same* given probability would already be costlier for the decision maker. As our previous lemma, Lemma 4.6 merely shows that it is always possible to reproduce the incentives imposed by a decision rule that fires the expert later by one which fires him earlier.

Lemma 4.6 (No Delay in Firing) *Within $\tilde{C}'$, the decision maker can without loss neglect such rules $\hat{\rho}$ for which there exists a history $\tilde{h}^p_{\tilde{t}} \in \mathcal{H}^p_1$ such that $\hat{\rho}(\tilde{h}^p_{\tilde{t}}) > 0$ and such that $\hat{\rho}(h^p_{\tilde{t}+1}) < 1$ for* all *successor histories $h^p_{\tilde{t}+1}$ and $\hat{\rho}(h^p_{\tilde{t}+1}) > 0$ for some successor history $h^p_{\tilde{t}+1}$.*

Proof: The proof again proceeds by construction. The idea behind the construction is to decrease the probability of retainment in period $\tilde{t}$ and simultaneously to increase the probability of retainment in the subsequent period in a manner that will uniformly increase the expert's continuation payoffs in the next period on the equilibrium path. We prove that this construction will not disturb incentive compatibility of the decision rule and will not decrease the expected payoff of the decision maker. Details are provided in the appendix. ▮

While we are providing the full proof in the appendix, we here illustrate the construction underlying our proof with an example. Let $N = 1$ and assume that $\hat{\rho}(\emptyset) = 1$, $\hat{\rho}(h^p_1) = \gamma \in (0, 1)$ for all $h^p_1 \in \mathcal{H}^p_1$, and $\hat{\rho}(h^p_1) = 0$ for all $h^p_1 \in \mathcal{H}^p_\emptyset$. In the new decision rule ρ', we set $\rho'(h^p_1) = 1$ if $h^p_1 \in \mathcal{H}^p_1$ and $\rho'(h^p_1) = 1 - \gamma$ otherwise. Finally, we let $\rho'(\emptyset) = \gamma$.

It is clear that ρ' is incentive compatible as the continuation payoffs of the expert after each history in the first period are increased by $1 - \gamma$. Furthermore, in the first period the decision maker's payoff is decreased by $1 - \gamma$ times the expected value of the expert, whilst in the second period the decision maker's payoff is increased by $1 - \gamma$ times the expected value of the expert after the first period history. The former is necessarily not bigger than the latter as the decision maker can condition her later action on additional payoff relevant information. Hence, the decision maker is weakly better off under ρ'.

Several additional issues arise when we generalize this argument. First, the expression for the amount by which we should increase the probability of employment in the consequent period is more complex and depends on the continuation payoff of the expert as well as the probability of employment in the current period. Second, the degree by which the payoffs of the expert are affected will depend on his private belief; this requires some additional arguments to prove incentive compatibility of the constructed rule. Finally, extending the argument to show that the decision maker is better off with the new rule is not immediate because the amount of increase in the probability of the expert's employment in latter periods is not necessarily constant across histories.

The following proposition shows that whenever employment costs are low, the risky arm will be over-used in the optimal decision rule, as compared to the first-best benchmark; in order to give proper incentives for truthtelling, the expert is kept on for some time even after a mistake.

Proposition 4.7 (Optimal decision rule, low costs) *There exists* $\underline{c} > 0$ *such that for all* $c < \underline{c}$ *in the optimal decision rule*

1. *the expert is always retained after a correct report;*
2. *incentives to report truthfully are provided by a promise to retain the expert for a certain amount of time even if his report turns out to be incorrect.*

PROOF: See appendix. ▮

5 No Commitment

A decision maker's option of continuing to employ the expert after he has mistakenly announced the less likely state crucially depends on the assumption that she can credibly commit to doing so *ex ante*. Indeed, after the expert has made a mistake in an equilibrium where he is supposed always truthfully to reveal his signal, he is commonly known to be of poor quality, and it is not sequentially rational for the decision maker to keep him on. By the same token, a decision maker would not fire the expert after he correctly forecast the state, albeit the more likely one, unless she is bound to do so. Indeed, after the expert has correctly forecast the state, even though it be the more likely one, his reputation increases, so that his services become more, rather than less, valuable to the decision maker.

This brief discussion shows that many of the tools the decision maker may be using in the optimal decision rule may no longer be available to her in the absence of long-term commitment possibilities. However, she could still try and enforce truth-telling by threatening to fire the expert with some probability after he has correctly forecast the more likely state, a threat made incentive compatible for her by the expert's off-equilibrium threat to send but uninformative messages if he continues to be employed and the decision maker has randomized with an off-equilibrium probability.[11] Of course, this construction relies on the assumption that the decision maker's randomizing probability is perfectly observable to the expert.

[11] This is incentive compatible for the expert, because the decision maker does not condition his continued employment on the reports.

One result from the commitment case readily extends, however. If, in response to the decision maker's pursuing her first-best policy, the expert maximizes his reputation by being truthful, then it is the decision maker's (sequentially rational!) best response to pursue her first-best policy; therefore, if the number of sequential decision problems is large enough, the first-best policy becomes incentive compatible, by the same argument as in the problem with full commitment. We summarize this finding in the following remark:

Remark For any given c, p_0, α_0, there exists an integer N_0 such that the first-best decision rule is incentive compatible if and only if $N \geq N_0$.

PROOF: Identical to Proposition 4.2, and therefore omitted. ∎

5.1 Fully Revealing Equilibria

Even without commitment, it is still possible to construct equilibria in which the expert always truthfully reveals his information.

Proposition 5.1 *Suppose assumption 4.1 holds. Then, for any N, α_0 and p_0, there exists a $\hat{c} > 0$ such that for all $c < \hat{c}$, there exists a fully revealing equilibrium in the game without commitment.*

PROOF: Let the principal pursue her first-best policy after t^{FB}, and let her always fire the expert after he has made a mistake, while always keeping him on after he has announced state 0 correctly. After he has forecast state 1 correctly, he is kept on with a probability $\rho_{1,t}^{s}$ such that, before time t^{FB}, the incentive constraint for a signal of 0 exactly bind. From the explicit expression for the incentive constraint, it is immediate that $\rho_{1,t}^{s} > \frac{p_0}{1-p_0}$. Hence, the probability of reaching period t^{FB}, $\tilde{\mu}_{t^{FB}}$, is bounded below by $\alpha_0 \left(2p_0\right)^{t^{FB}} > 0$. Thus, the principal's payoff from this policy is bounded below by

$$\alpha_0 \left(2p_0\right)^{t^{FB}} v^{FB}(\alpha_{t^{FB}}, N - t^{FB}) - t^{FB}c.$$

As the first term is positive by Assumption 4.1, as well as increasing in c, the verification of the mutual best response property, as indicated in the text, completes the proof. ∎

However, without commitment, the Revelation Principle no longer applies, which means that we do not know *ex ante* that full revelation will necessarily be optimal. Indeed, below, we shall construct an equilibrium which entails the accumulation of endogenous private information on the agent's part; it can be shown that for any fully revealing equilibrium, there exists a threshold $\underline{c} > 0$ such that the equilibrium with private information dominates it for all $c < \underline{c}$. In the following proposition, we show that whenever the first-best solution is not incentive compatible at time 0, the decision maker's payoff in any fully revealing equilibrium is bounded away from $v^{FB}(\alpha_0, N+1)$.

Proposition 5.2 *Let $t^{FB} \geq 1$, and let V, and ρ, be the decision maker's payoff, and strategy, in a fully revealing equilibrium. Then, there exists a time $t < t^{FB}$, a constant μ_t and a history $h_t \in \mathcal{H}_1$ such that*

$$V \leq v^{FB}(\alpha_0, N+1) - \mu_t(1-\rho(h_t))v^{FB}(\alpha_{t+1}, N-t) < v^{FB}(\alpha_0, N+1).$$

Proof: Since $t^{FB} \geq 1$, there exists some $t < N$ and some history $h_t \in \mathcal{H}_1$ with $\rho(h_t) < 1$ and $\mu_t > 0$, where μ_t denotes the probability of the expert's still being employed in period t, and history h_t realizing, in this equilibrium. The upper bound is then given by the hypothetical situation where the first best became implementable after period t. By assumption 4.1, $v^{FB}(\alpha_{t+1}, N-t) > 0$; hence V is bounded away from $v^{FB}(\alpha_0, N+1)$. ∎

As an illustration, we give the following corollary pertaining to a particularly intuitive class of fully revealing equilibria:

Corollary 5.3 *Suppose there exists a fully revealing equilibrium which has the feature that after t^{FB}, the first-best decision rule is played. Then, there exists a constant $\mu_{t^{FB}-1} > 0$ such that the principal's payoff in this equilibrium is bounded above by*

$$\begin{aligned} &v^{FB}(\alpha_0, N+1) - \hat{\mu}_{t^{FB}-1}(1-p_0)\beta_{t^{FB}-1}v^{FB}(\alpha_{t^{FB}}, N-t^{FB}) \\ &\quad \times \left[1 - \frac{p_0\beta_{t^{FB}-1}}{(1-p_0)(1-\beta_{t^{FB}-1})} \frac{1+\beta_{t^{FB}-1}+\cdots+\beta_{t^{FB}-1}^{N-t^{FB}}}{1+(1-p_0)+\cdots+(1-p_0)^{N-t^{FB}}}\right] < v^{FB}(\alpha_0, N+1). \end{aligned}$$

Proof: Let $\hat{\mu}_{t^{FB}-1}$ denote the probability of the expert's still being employed in period $t^{FB}-1$. Then, the upper bound on the principal's payoff is derived by computing the loss in payoffs as compared to the first-best from the binding incentive constraint in period $t^{FB}-1$, given the first-best decision rule is played starting from the next period onward. By assumption 4.1, $v^{FB}(\alpha_{t^{FB}}, N-t^{FB}) > 0$; moreover, $\hat{\mu}_{t^{FB}-1} > 0$, since otherwise the decision maker would be better off not employing the expert at all in period 0. ∎

5.2 Equilibrium With Endogenous Private Information

A quite natural way for the decision maker to handle the expert's incentive problem would be for her to grant him an initial "grace period", during which he was allowed to send uninformative signals each period, and to gain confidence in his abilities, finding his mark in his new job. Once this probationary period, which lasts for t^{FB} periods, ends, though, he is expected to be right every time, i.e. the first-best policy will be implemented. The expert will then report his signals truthfully if, and only if, his signals have all been correct during the first t^{FB} periods; otherwise, he will best respond by continuing to babble, i.e. to announce state 1 no matter what his signal may have been.

Of course, for the decision maker to be willing to extend such initial leniency, it has to be the case that the overall length of this probationary period, during which she incurs a per-period loss of c, is sufficiently small relative to the expected gains in the later stages of the relationship. Specifically, the following condition must hold:

$$\begin{aligned} & v^{FB}(\alpha_0, N+1) \geq \\ & (1-\alpha_0)c\Big\{t^{FB} - 1 - \tfrac{1}{2} - \cdots - \frac{1}{2^N} + \Big(1 - \big(\tfrac{1}{2}\big)^{t^{FB}}\Big)\Big[(1-p_0) + (1-p_0)^2 + \cdots + (1-p_0)^{N-t^{FB}}\Big] \\ & \qquad + \big(\tfrac{1}{2}\big)^{t^{FB}}\Big[\tfrac{1}{2} + \big(\tfrac{1}{2}\big)^2 + \cdots + \big(\tfrac{1}{2}\big)^{N-t^{FB}}\Big]\Big\}. \end{aligned} \tag{3}$$

We summarize this equilibrium in the following proposition:

Proposition 5.4 *Let (3) hold. Then, there exists an equilibrium in which no information is transmitted, and the expert is never fired during the first t^{FB} periods; thereafter, the expert truthfully reveals his signals if, and only if, his first t^{FB} signals were correct. Moreover, he will only be fired as soon as he has made an incorrect forecast after the first t^{FB} periods.*

PROOF: The expert's equilibrium strategy is for him always to report state 1, unless $t \geq t^{FB}$, and he has been employed, and has received correct signals, in all previous periods, in which case he truthfully reveals his signal. The decision maker's equilibrium strategy calls for employing the expert in $t = 0$, and for retaining him if he reports 1 during the first t^{FB} periods. In all other cases, she retains the expert if, and only if, his current prediction turns out to be correct. Optimality of the expert's behavior immediately follows from the definition of t^{FB}. Optimality of the principal's behavior follows from (3). ■

In this equilibrium, experts are costly but harmless, in the sense that it is always optimal for the decision maker to follow the expert's advice. Indeed, suppose the expert has privately learned that he is of the bad type; he then maximizes his expected employment duration by always reporting state 1. For a decision maker who knew that she cannot rely on a high-quality expert's advice, it is also optimal always to stick with her prior, and to implement policy 1. Hence, as compared to the first-best solution, babbling only imposes an employment cost on the decision maker; there is no cost in terms of inefficient decision making. Thus, for low c, the decision maker's payoff in this babbling equilibrium is arbitrarily close to the first best, something we explore in more detail in the following subsection.

5.3 Welfare Comparison

We shall now show that if c is below a certain threshold it pays for the principal to allow the agent to accumulate private information on the equilibrium path, as doing so dominates any given fully revealing equilibrium. Specifically, we shall show that the equilibrium we have constructed in the previous subsection dominates our fully revealing equilibria for low enough c. We summarize this result in the following proposition:

Proposition 5.5 *Let $t^{FB} \geq 1$. Then, there exists a threshold $\underline{c} > 0$ such that for all $c \in]0, \underline{c}[$ the principal is strictly better off in the equilibrium with babbling (see Proposition 5.4) than in any fully revealing equilibrium.*

Proof: By construction, the equilibrium exhibited in Proposition 5.4 yields the principal a value of V^p, with

$$\begin{aligned} V^p &= v^{FB} - (1-\alpha_0)c \\ &\times \{t^{FB} - 1 - \tfrac{1}{2} - \cdots - \frac{1}{2^N} + \left(1 - \left(\tfrac{1}{2}\right)^{t^{FB}}\right)\left[(1-p_0) + (1-p_0)^2 + \cdots + (1-p_0)^{N-t^{FB}}\right] \\ &\qquad + \left(\tfrac{1}{2}\right)^{t^{FB}}\left[\tfrac{1}{2} + \left(\tfrac{1}{2}\right)^2 + \cdots + \left(\tfrac{1}{2}\right)^{N-t^{FB}}\right]\}, \end{aligned}$$

while, by Proposition 5.2, the payoff in any fully revealing equilibrium is bounded above by

$$v^{FB}(\alpha_0, N+1) - \mu_t(1-\rho(h_t))v^{FB}(\alpha_{t+1}, N-t).$$

Clearly, any fully revealing equilibrium for some given value of $\underline{c}$ remains an equilibrium for all $c < \underline{c}$; moreover, $v^{FB}(\alpha_{t+1}, N-t)$ is decreasing in c. Furthermore, as the agent's incentives are independent of c, any new fully revealing equilibria that appear as we lower c are dominated by the original equilibrium. Thus, as c decreases, the equilibrium from Proposition 5.4 achieves a payoff arbitrarily close to the first-best, whereas the payoff in any fully revealing equilibrium remains bounded away from it. ∎

Thus, whereas with full commitment the Revelation Principle tells us that there is no loss for the principal to restrict herself to fully revealing decision rules, we have shown that this is no longer true without commitment. Indeed, while in the game with commitment, incentives may be given through rewards for minority opinion, this tool is no longer available without commitment. Thus, absent commitment, it may become too costly for the principal to extract the agent's information; he may in fact be better off forgoing learning for a while, thus endogenously allowing the expert to build up private information, which will then, in turn, facilitate the transmission of information.

6 Conclusion

We have investigated the dynamic interaction between a decision maker and an expert of unknown quality who privately observes a potentially decision-relevant signal. As he only cares about his reputation insofar as it translates into a longer expected duration of employment, the expert may have an incentive strategically to manipulate his cheap-talk relay of the signal to the decision maker. We have shown that when the number of periods is large enough, the expert's reputational concerns vanish, and the first-best becomes implementable. We have shown that in the game with commitment, agency costs may lead to a situation where the risky arm is being over-used as compared to the first-best benchmark. In the game without commitment, the decision maker may be better off letting the expert accumulate some private information during an initial grace period than he would be in any fully revealing equilibrium.

While we have assumed a simple binary type space for the expert, one could of course imagine richer type spaces. While analytical results will likely be rather difficult to get in that setting, as the number of relevant incentive constraints increases, we should expect the qualitative results of our analysis to continue to hold. Moreover, it seems interesting to introduce multiple, possibly even a continuum of, experts into our setup. One could imagine that the introduction of competition amongst experts would make them more inclined to gamble on the initially less likely state, thus easing incentive constraints, as there will be pressure favorably to surprise the decision maker with an unlikely success, especially in the early stages of the game. We intend to explore these questions in future work.

Appendix

Proof of Proposition 4.2

Suppose the principal pursues the first-best policy of immediately firing the expert if, and only if, the expert has made a mistake. Then, the agent is willing to reveal a signal indicating the less likely state 0 truthfully at any time t, if at all times $0 \leq t \leq N$, the following incentive constraint holds:

$$p_0 \left[\alpha_t (N-t) + \frac{1-\alpha_t}{2} \left(1 + \tfrac{1}{2} + \cdots + \frac{1}{2^{N-t-1}} \right) \right] \\ \geq (1-p_0) \frac{1-\alpha_t}{2} \left[1 + (1-p_0) + \cdots + (1-p_0)^{N-t-1} \right]. \quad \text{(A.1)}$$

To understand the right-hand side of the incentive constraint, the reader should note that if, upon lying, the expert finds out *ex post* that his message was in fact correct, he then privately learns that he is of the low type and will maximize his continuation payoff by reporting the *a priori* more likely state in all subsequent periods.

It is now immediate to verify that, as $N \to \infty$, the left-hand side diverges to $+\infty$, whereas the right-hand side converges to $\frac{1-p_0}{p_0} \frac{1-\alpha_t}{2} < \infty$. Let N_0 be the smallest value of N for which this constraint is satisfied for all $t \leq K$. By our assumption that $\beta_0 < 1 - p_0$, we have $N_0 \geq 2$.

It is left to check that the constraint is also satisfied for all $t > K$. It is direct to verify that the constraint holds for any N if $\alpha_t = 1 - 2p_0$. Furthermore, the left hand side of the constraint is increasing in α_t while the right hand side is decreasing in α_t. Therefore, the constraint is satisfied for all $\alpha_t \geq 1 - 2p_0$, which is equivalent to $t \geq K$.

As is straightforward to verify, the left-hand side of the incentive constraint conditional on a signal indicating the more likely state 1, is $\frac{1-p_0}{p_0} > 1$ times the left-hand side of the above constraint, whereas the right-hand side is $\frac{p_0}{1-p_0}$ times the above right-hand side. Therefore, this constraint also holds for all $N \geq N_0$. ∎

Proof of Lemma 4.5

Consider an incentive compatible decision $\hat{\rho}$ that satisfies the conditions of the lemma. We show that there exists an incentive compatible decision rule ρ' in which

$$\begin{array}{rcl} \rho'(\tilde{h}^p_{\tilde{t}}) & > & \hat{\rho}(\tilde{h}^p_{\tilde{t}}), \\ \rho'(0, \emptyset, \omega, \tilde{h}^p_{\tilde{t}}) & = & 0 \end{array}$$

and this rule yields the same payoffs as $\hat{\rho}$ to both parties. Iterated application of this construction proves the claim.

To make the argument more transparent, we offer separate constructions for deterministic and stochastic decision rules, even though the former construction is a special case of the latter.

Let $h^p_{\emptyset}$ and h^p_1 be any one period public history with $\phi = 0$ and $\phi = 1$ respectively. The argument for a deterministic decision rule is straightforward: We have $\hat{\rho}(\tilde{h}^p_{\tilde{t}}) = 0$ and $\hat{\rho}(\tilde{h}^p_{\tilde{t}} h^p_{\emptyset}) = 1$ and construct ρ' as follows:

(i) $\rho'(\tilde{h}^p_{\tilde{t}}) := 1$ (compare with $\hat{\rho}(\tilde{h}^p_{\tilde{t}}) = 0$);

(ii) $\rho'(\tilde{h}^p_{\tilde{t}} h^p_1) := 0$ (observe that $\tilde{h}^p_{\tilde{t}} h^p_1$ is reached with 0 probability under $\hat{\rho}$);

(iii) $\rho'(\tilde{h}^p_{\tilde{t}} h^p_\emptyset) := 0$ (compare with $\hat{\rho}(\tilde{h}^p_{\tilde{t}} h^p_\emptyset) = 1$);

(iv) $\rho'(\tilde{h}^p_{\tilde{t}} h^p_1 h^p_\emptyset h^p_\tau) := \hat{\rho}(\tilde{h}^p_{\tilde{t}} h^p_\emptyset h^p_1 h^p_\tau)$ for any $h^p_1, h^p_\emptyset$ and h^p_τ with $\tau \geq 0$,

For the remaining histories, we set $\rho' = \hat{\rho}$.

By construction of ρ', the expected payoff of the decision maker is the same in both rules.

Furthermore, the expert's payoff is affected only at the histories that are successors of $\tilde{h}^p_{\tilde{t}}$. The incentive compatibility after any private history on the equilibrium path that is a successor of $\tilde{h}^p_{\tilde{t}} h^p_1 h^p_\emptyset$ follows from the incentive compatibility at the corresponding successor history of $\tilde{h}^p_{\tilde{t}} h^p_\emptyset h^p_1$ in $\hat{\rho}$. The incentive compatibility in period $\tilde{t}+1$ after histories $\tilde{h}^p_{\tilde{t}} h^p_1$ and $\tilde{h}^p_{\tilde{t}} h^p_\emptyset$ is not relevant as the expert is not employed in this period. The incentive compatibility of ρ' in period $\tilde{t}$ after history $\tilde{h}^p_{\tilde{t}}$ follows from the incentive compatibility in period $\tilde{t}+1$ after history $\tilde{h}^p_{\tilde{t}} h^p_\emptyset$ under $\hat{\rho}$. For other successor histories of $\tilde{h}^p_{\tilde{t}}$, both decision rules coincide and hence incentive compatibility is immediate.

For stochastic decision rules, the argument is a bit more tedious as we make sure that the new rule generates the same distribution of payoff relevant events. By Lemma 4.4, it holds that $\hat{\rho}(0, \emptyset, 1, \tilde{h}^p_{\tilde{t}}) = \hat{\rho}(0, \emptyset, 0, \tilde{h}^p_{\tilde{t}}) =: r$. We construct ρ' as follows:

(i) $\rho'(\tilde{h}^p_{\tilde{t}}) := \hat{\rho}(\tilde{h}^p_{\tilde{t}}) + (1 - \hat{\rho}(\tilde{h}^p_{\tilde{t}}))r$;

(ii) $\rho'(\tilde{h}^p_{\tilde{t}} h^p_1) := \gamma \hat{\rho}(\tilde{h}^p_{\tilde{t}} h^p_1)$;

(iii) $\rho'(\tilde{h}^p_{\tilde{t}} h^p_\emptyset) := 0$;

(iv) $\rho'(\tilde{h}^p_{\tilde{t}} h^p_1 h^p_\emptyset h^p_\tau) := \gamma \hat{\rho}(\tilde{h}^p_{\tilde{t}} h^p_1 h^p_\emptyset h^p_\tau) + (1-\gamma)\hat{\rho}(\tilde{h}^p_{\tilde{t}} h^p_\emptyset h^p_1 h^p_\tau)$ for any $h^p_1, h^p_\emptyset$ and h^p_τ with $\tau \geq 0$,

where

$$\gamma \;=\; \frac{\hat{\rho}(\tilde{h}^p_{\tilde{t}})}{\hat{\rho}(\tilde{h}^p_{\tilde{t}}) + (1 - \hat{\rho}(\tilde{h}^p_{\tilde{t}}))r}.$$

For the remaining histories, we set $\rho' = \hat{\rho}$.

By construction of ρ', the expected payoffs of the decision maker is the same in both rules. Verifying incentive compatibility of $\hat{\rho}$ is direct and we omit it. ■

Proof of Lemma 4.6

Assume that $t \leq K_0$. Define H' to be the set of all private histories in $\tilde{\mathcal{H}}^a$ whose corresponding public history is a one period successor history of $\tilde{h}^p_{\tilde{t}}$. With some abuse of notation, let $h^p(h^a_t)$ denote the public history induced by a private history h^a_t.

Set h^* to be a private history for which $u^{\hat{\rho}}(h^a_t)(1/\hat{\rho}(h^p(h^a_t))-1)$ is minimized among all $h^a_t \in H'$ and define $\delta := u^{\hat{\rho}}(h^*)(1/\hat{\rho}(h^p(h^*)) - 1)$. In addition, denote by $h' \in \tilde{\mathcal{H}}^a$ the private history that induces $\tilde{h}^p_{\tilde{t}}$ and let u^* be the expert's expected payoff under $\hat{\rho}$ after h' conditional on being employed in period $\tilde{t}$.

We construct ρ' as follows: First, define $\rho'(\tilde{h}^p_{\tilde{t}}) = \hat{\rho}(\tilde{h}^p_{\tilde{t}})\frac{u^*}{u^*+\delta}$. Next, pick any public history h that is a one period successor of $\tilde{h}^p_{\tilde{t}}$. If $\hat{\rho}(h) > 0$, we set $\rho'(h) = \hat{\rho}(h)(u^{\hat{\rho}}(h^a_t) + \delta)/u^{\hat{\rho}}(h^a_t)$, where h^a_t is the private history in $\tilde{\mathcal{H}}^a$ that corresponds to h. (Observe that $\rho'(h) \leq 1$ by definition of δ.) If $\hat{\rho}(h) = 0$, we define ρ' at h and any successor of h such that (a) ρ' does not condition on the continuation history after h and (b) generates for the expert the expected payoff $u(h^a_t) = \delta$, where h^a_t is any private history that induces h. (This construction is feasible by definition of δ.) Finally, we keep $\rho' = \hat{\rho}$ at any other public history.

By construction, the expert's payoffs on the equilibrium path are either unaffected or, at history $\tilde{h}^p_{\tilde{t}}$ and conditional on retainment are increased uniformly by δ for all continuation histories. Now, consider any out of equilibrium path private history that corresponds to the public history $\tilde{h}^p_{\tilde{t}}$. Then, the expert knows that he is incompetent.[12] Observe that the continuation payoffs of the expert for all two period successor private histories are not affected. Furthermore, all payoffs for one period successor histories that result in a public history in $\mathcal{H}^p_{\emptyset}$ are increased by δ. Furthermore, since a competent expert can always ensure the same expected payoff as a bad expert,[13] the payoffs of the incompetent expert are smaller than, or equal to, the payoffs of the expert who is competent with a positive probability, after any history, and the incompetent expert's payoffs after all one period successor private histories do not increase by more than δ. Thus, ρ' is incentive compatible.

It remains to be shown that the decision rule ρ is in the class of decision rules that do not satisfy the conditions of the lemma. This follows, however, from the observation that $\rho'(\tilde{h}^p(h^*)) = 1$ by the definition of δ.

Finally, from an *ex ante* perspective, the expected number of periods of employment of the expert is the same under both rules, and hence the payoff of the decision maker is weakly higher under ρ' as she is better informed about the expert's type in future periods, and her payoff in future periods is bounded below by $-c$, the payoff in period $\tilde{t}$.

The argument for $\tilde{t} > K_0$ is different and simpler. The new decision rule can be constructed by increasing the probability of employment in period $\tilde{t}+1$ after correct reports and decreasing the probability of employment in $\tilde{t}$ to keep the expected payoff of the expert on the equilibrium path at the node corresponding to $\tilde{h}^p_{\tilde{t}}$ constant in both rules. The new rules delivers a strictly higher payoff for the decision maker; incentive compatibility of the new rule follows from $\tilde{t} > K_0$.

Iterated application of the preceding construction proves the claim. ∎

Proof of Proposition 4.7

Proof is by contradiction. Suppose ρ is an optimal decision rule and suppose there exists a history $\hat{h} \in \mathcal{H}_1$ such that $\rho(\hat{h}) < 1$. Let $\hat{h}_{t^*}$ be the longest such history. If there are more such histories

[12] By Lemma 4.3, we need not consider histories after which the expert privately thinks he might be of the good type, while, according to the public belief, he is certain to be of the bad type.

[13] The competent expert can always play the strategy of the incompetent expert by using the outcome a fair coin toss to generate the signal of the incompetent expert.

of equal length t^*, we choose any one of them. We shall now recursively construct an alternative decision rule ρ', which is incentive compatible and performs better than ρ, thus contradicting the optimality of ρ.

For all histories h of length greater than t^*, we set $\rho'(h) \equiv \rho(h)$; hence ρ' is incentive compatible after all these histories, as the optimality of ρ implies its incentive compatibility. Moreover, we set $\rho'_1(\hat{h}_{t^*}) \equiv 1$; in order to preserve incentive compatibility after an immediate predecessor history h_{t^*-1}, the employment probabilities may have to be adjusted after other successor histories of $\hat{h}_{t^*-1}$ that are of length t^*. Define $\Delta := [1-\rho(\hat{h}_{t^*})]u^{\rho}(\hat{h}_{t^*})$, where $u^{\rho}(\hat{h}_{t^*})$ denotes the expected duration of the expert's employment under ρ conditional on his being re-employed after $\hat{h}_{t^*}$. It follows from the explicit expression of the incentive constraints after any given predecessor history $\hat{h}_{t^*-1}$ that an upper bound on the *ex interim* cost, as evaluated at period t^*-1, from readjusting the employment probabilities to restore incentive compatibility after history $\hat{h}_{t^*-1}$ is given by $Pr(\hat{h}_{t^*}|\hat{h}_{t^*-1}, \rho)\frac{1-p_0}{p_0}\Delta c$. Call δ_{t^*} the overall increase in the expert's expected employment conditional on his being re-employed after history $\hat{h}_{t^*-1}$.

Now, consider $\hat{h}_{t^*-2}$, any given immediate predecessor history of $\hat{h}_{t^*-1}$. Again, the explicit expression for the incentive constraint after history $\hat{h}_{t^*-2}$ shows us that the cost of restoring incentive compatibility in period t^*-2 is bounded above by $Pr(\hat{h}_{t^*-1}|\hat{h}_{t^*-2}, \rho)\frac{1-p_0}{p_0}\delta_{t^*}c \leq Pr(\hat{h}_{t^*}|\hat{h}_{t^*-2}, \rho)\left(\frac{1-p_0}{p_0}\right)^2 \Delta c$. By iterating these steps, one shows that the overall cost of making ρ'_1 incentive compatible is bounded above by $Pr(\hat{h}_{t^*}|\rho)\left(\frac{1-p_0}{p_0}\right)^{t^*+1}(t^*+1)c\Delta$.

If $t^* > K$, the benefit of moving from ρ to ρ'_1 is bounded below by $Pr(\hat{h}_{t^*}|\rho'_1)(1-\rho(\hat{h}_{t^*}))\left(\beta_{t^*+1}-(1-p_0)-c\right)$; if $t^* \leq K$, it is bounded below by $Pr(\hat{h}_{t^*}|\rho'_1)(1-\rho(\hat{h}_{t^*}))[\alpha_{t^*}(N-K)(p_0-c)-(K-t^*)c]$, which, by assumption 4.1, is strictly positive. By construction, we have that $Pr(\hat{h}_{t^*}|\rho) = Pr(\hat{h}_{t^*}|\rho'_1)$. Thus, for ρ'_1 to dominate ρ, it is sufficient that

$$\alpha_{t^*}(N-K)(p_0-c) - c(K-t^*) \geq \left(\frac{1-p_0}{p_0}\right)^{t^*+1}(t^*+1)c.$$

Since the left-hand side is bounded away from zero by assumption 4.1, this holds for c below a certain threshold.

If there exists a history $\tilde{h} \in \mathcal{H}_1$ such that $\rho'_1(\tilde{h}) < 1$, we repeat the above procedure for ρ'_1. As the number of periods and histories is finite, there exists some finite n such that $h \in \mathcal{H}_1 \Rightarrow \rho'_n(h) = 1$. Setting $\rho' \equiv \rho'_n$ completes the proof. ▮

References

The Associated Press (December 23, 2005): "Korean Scientist Resigns Over Fake Stem Cell Research," available online e.g. at
http://www.somaliaonline.com/cgi-bin/ubb/ultimatebb.cgi?ubb=get_topic;f=6;t=005923;p=0.

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.

Bar-Isaac, H. (2003): "Reputation and Survival: Learning in a Dynamic Signalling Model," *Review of Economic Studies*, 70, 231–251.

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

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

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

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

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

Bergin, J. (1992): "A Model of Strategic Behavior in Repeated Games," *Journal of Mathematical Economics*, 21, 113–153.

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.

Bolton, P., J. Scheinkman and W. Xiong (2005): “Pay for Short-term Performance: Executive Compensation in Speculative Markets,” *Journal of Corporation Law*, 30(4), 721–748.

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

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.

Crawford, V. and J. Sobel (1982): “Strategic Information Transmission,” *Econometrica*, 50, 1431–1451.

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

Dewatripont, M. and J. Tirole (1999): “Advocates,” *Journal of Political Economy*, 107, 1–39.

Fong, K. (2009): “Evaluating Skilled Experts: Optimal Scoring Rules for Surgeons,” working paper, Stanford University.

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

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

Harrington, J.E. Jr. (1995): “Experimentation and Learning in a Differentiated-Products Duopoly,” *Journal of Economic Theory*, 66, 275–288.

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

Hörner, J. and L. Samuelson (2009): “Incentives for Experimenting Agents,” Cowles

Foundation Discussion Paper No. 1726.

Keller, G. and S. Rady (2003): "Price Dispersion and Learning in a Dynamic Differentiated-Goods Duopoly," *RAND Journal of Economics*, 34, 138–165.

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 T. Mylovanov (2010): "Expert Experimentation," working paper, University of Munich and Pennsylvania State University.

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

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

Klein, N. (2010b): "The Importance of Being Honest," working paper, University of Munich.

Mailath, G. and L. Samuelson (2006): *Repeated Games and Reputations.* Oxford University Press.

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

Morris, S. (2001): "Political Correctness," *Journal of Political Economy*, 109, 231–265.

Morgan, J. and P. C. Stocken (2003): "An Analysis of Stock Recommendations," *RAND Journal of Economics*, 34(1), 183–203.

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

Olszewski, W. and M. Peski (2009): "The Principal-Agent Approach to Testing Experts," working paper, Northwestern University and University of Texas at Austin.

Olszewski, W. and A. Sandroni (2008): "Manipulability of Future-Independent Tests," *Econometrica*, 76, 1437–1466.

Ottaviani, M. and P. N. Sørensen (2006a): "Professional Advice," *Journal of Economic Theory*, 126, 120–142.

Ottaviani, M. and P. N. Sørensen (2006b): “Reputational Cheap Talk,” *RAND Journal of Economics*, 37(1), 155–175.

Pastorino, E. (2005): “Essays on Careers in Firms,” Ph.D. Dissertation, University of Pennsylvania.

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

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

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

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

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

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

Simon, L.K. and M.B. Stinchcombe (1989): “Extensive Form Games in Continuous Time: Pure Strategies,” *Econometrica*, 57, 1171–1214.

Stewart, C. (2009): “Nonmanipulable Bayesian Testing,” University of Toronto, working paper 360.

Eidesstattliche Versicherung

Ich versichere hiermit eidesstattlich, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe verfasst habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sowie mir gegebene Anregungen sind als solche kenntlich gemacht. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.