Negatively Correlated Bandits*

Nicolas Klein† Sven Rady‡

This version: December 2, 2010

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 at the University of Munich as well as from the National Research Fund of Luxembourg is gratefully acknowledged.

†Department of Economics, University of Bonn, Lennéstr. 37, D-53113 Bonn, 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 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, 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 with 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

[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.

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 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 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 a Markov perfect equilibrium (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

[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).

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 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

[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.

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 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 (2008) 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

[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;

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 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. As Markov strategies of player $i = 1, 2$, we allow all functions $k_i : [0, 1] \to \{0, 1\}$

see the discussion in Section 4.6 below.

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

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

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

[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 definition of Markov strategies, 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.

and $u_2(p|k_1,k_2) \geq u_2(p|k_1,\tilde{k}_2)$ for all p in the unit interval and all admissible $(k_1,\tilde{k}_2)$. An 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

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

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

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

[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.

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 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.

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

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 $p = \frac{1}{2}$.

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

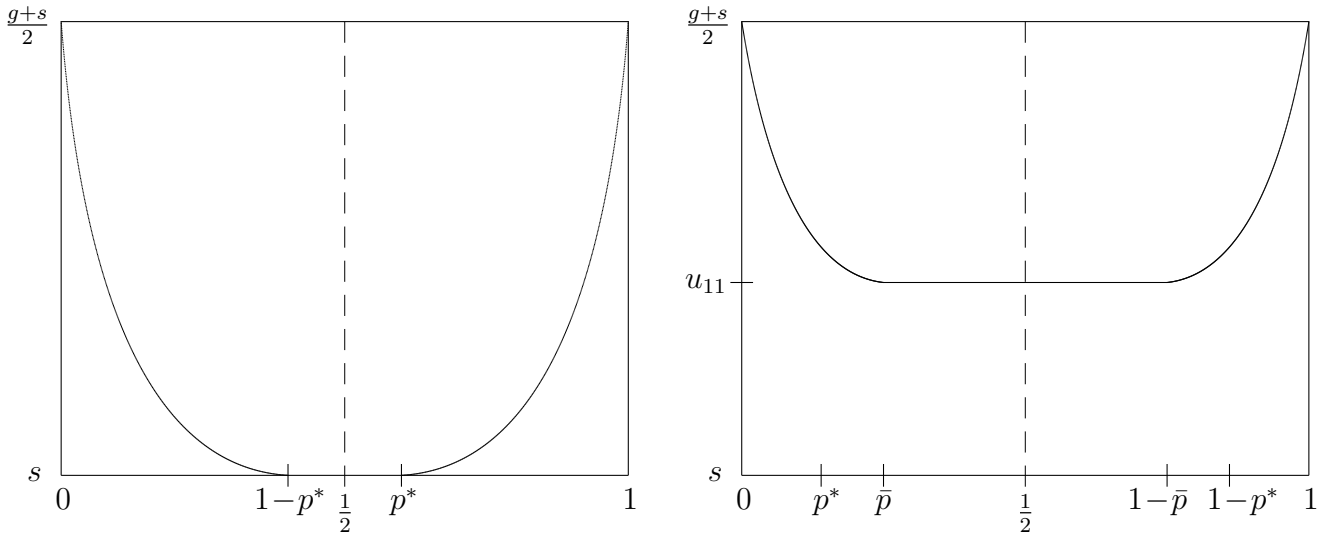

Figure 1: The planner's value function for $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$ (left panel) and $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$ (right panel).

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* $(k_1, k_2) = (1, 0)$ *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

[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.

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, 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 $p=1$.

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 (MPE for low stakes) *When $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$, the unique MPE 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, $p^* < p^m < \frac{1}{2} < 1-p^m < 1-p^*$.

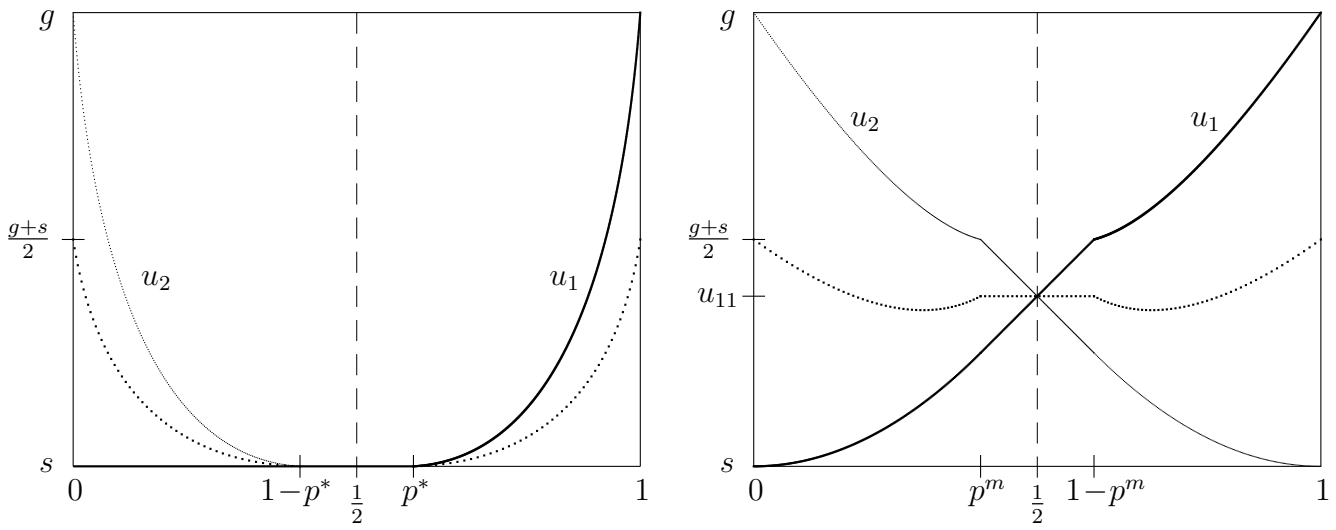

Figure 2: The equilibrium payoff functions for $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$ (left panel) and $\frac{g}{s} > 2$ (right panel). The thick solid curve depicts the payoff function of player 1, the thin solid curve that of player 2, and the dotted curve the players' average payoff function.

Proposition 5 (MPE for high stakes) *When $\frac{g}{s} > 2$, the game has a unique MPE, 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 MPE 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 $\hat{p} \in [\max\{p^*, 1-p^m\}, \min\{p^m, 1-p^*\}]$ 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 $\frac{g}{s} = \frac{2r+\lambda}{r+\lambda}$.*

Proof: For $\max\{p^*, 1-p^m\} < \hat{p} < \min\{p^m, 1-p^*\}$, 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*

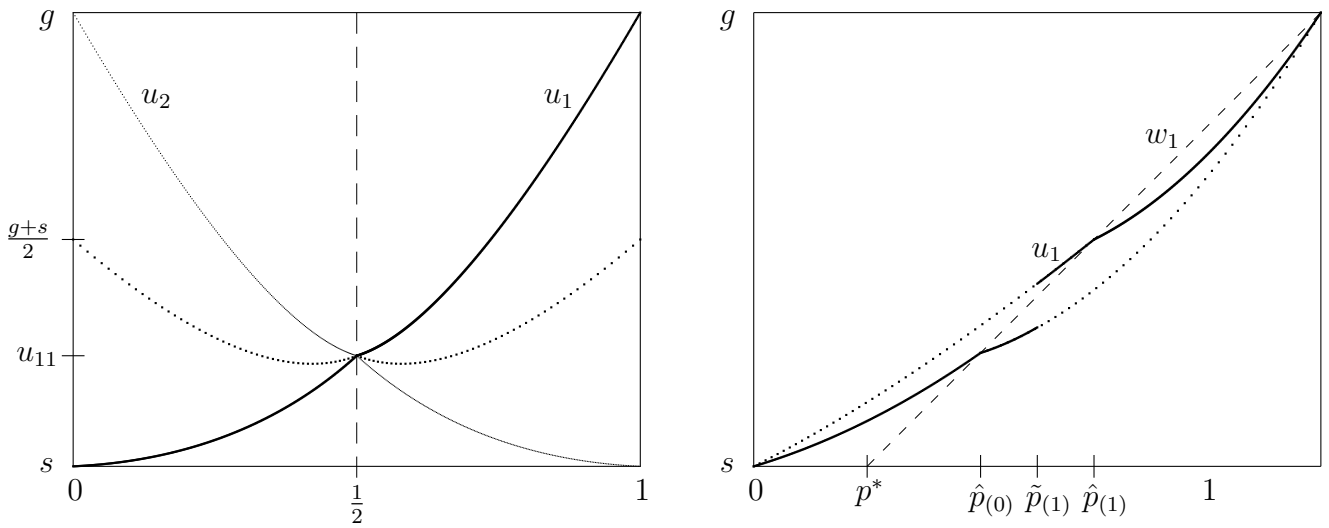

Figure 3: Equilibrium payoff functions for $\frac{2r+\lambda}{r+\lambda} < \frac{g}{s} < 2$. The left panel shows the players' payoff functions and their average in the unique symmetric equilibrium in cutoff strategies. The right panel shows the payoff function of player 1 in an equilibrium that is not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

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* $[\max\{p^*, 1-p^m\}, \min\{p^m, 1-p^*\}]$ *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

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

The asymptotics of the learning process depend both on the true state of the world and the initial belief. Let us suppose 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 MPE 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 MPE 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 MPE 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 MPE 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 MPE. 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 MPE 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, 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 MPE 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 deter-

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

mined 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 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

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

The corresponding differential equations are

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

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 $k_i \colon \mathcal{P} \to [0,1]$ with the following properties: (i) the interior in $\mathcal{P}$ of each of the sets $k_i^{-1}(0)$ and $k_i^{-1}(1)$ has a finite number of connected components; (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 MPE 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).

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

[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.

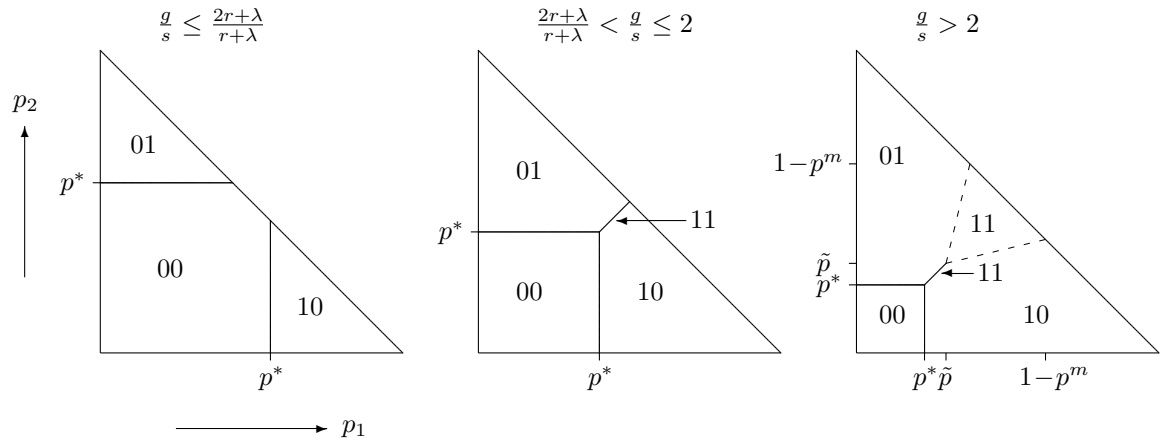

Figure 4: Cutoff equilibria of the experimentation game with imperfect negative correlation between the types of the risky arms.

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

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

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 MPE 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 $p_{1,0} \geq p_{2,0}$.

[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.

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.

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 $c_i(p_1,p_2) = s-p_ig$ 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: $u_i = p_ig + (1-p_i)\frac{\lambda}{r+\lambda}s$. 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 MPE 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. ∎

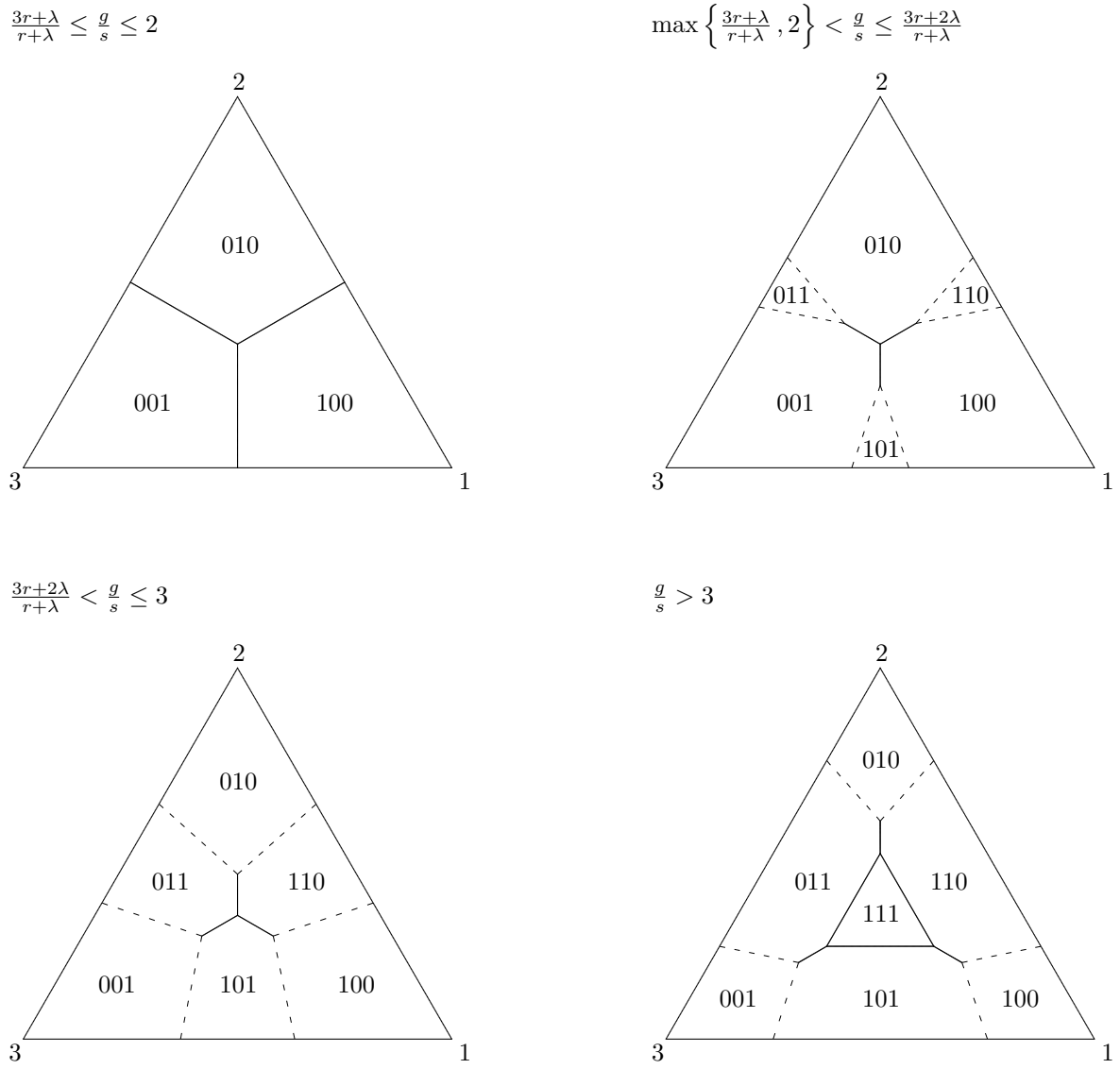

Figure 5: Cutoff equilibria of the experimentation game with three players when $p^* \leq \frac{1}{3}$.

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 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 MPE 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, Klein (2010) 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 MPE 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 $u_1(p) = pg + C_1u_0(p)$ and $u_2(p) = s + C_2u_0(p)$ with constants C_1 and C_2.

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 $u_1(p) = s + D_1u_0(1-p)$ and $u_2(p) = (1-p)g + D_2u_0(1-p)$ 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 $(1,0)$:

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

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 $-\frac{h(p_1)}{1-p_1} \le h'(p_1) \le \frac{h(p_1)}{p_1}$. 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

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

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

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

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

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

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

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* $[\max\{p^*, 1-p^m\}, \min\{p^m, 1-p^*\}]$. *(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. ∎

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

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

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 $\frac{g}{s} \geq 2$, 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 $I = [\max\{p^*, 1 - p^m\}, \min\{p^m, 1 - p^*\}]$. 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 $1 - \bar{p}$, 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. ∎

References

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

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.

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

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.

Dewatripont, M. and J. Tirole (1999): "Advocates," *Journal of Political Economy*, 107, 1–39.

Harrington, J.E. Jr. (1995): "Experimentation and Learning in a Differentiated-Products Duopoly," *Journal of Economic Theory*, 66, 275–288.

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. (2010): "Strategic Learning in Teams," SFB/TR 15 Discussion Paper No. 333.

Murto, P. and J. Välimäki (2008): "Learning and Information Aggregation in an Exit Game," Helsinki Center of Economic Research Discussion Paper No. 235.

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.

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.

Simon, L.K. and M.B. Stinchcombe (1989): "Extensive Form Games in Continuous Time: Pure Strategies," *Econometrica*, 57, 1171–1214.