Strategic Experimentation with Public or Private Information*

Martin W. Cripps† Godfrey Keller‡ Sven Rady§

This version: January 2000
Currently under Revision

Abstract

This paper studies a game of strategic experimentation in which the players learn from the experiments of others as well as their own. We first establish the efficient benchmark where the players co-ordinate in order to maximise joint expected payoffs, and then show that, because of free-riding, the strategic problem leads to inefficiently low levels of experimentation in any equilibrium when the players use stationary Markovian strategies. Efficiency can be approximately retrieved provided that the players adopt strategies which slow down the rate at which information is acquired; this is achieved by their taking periodic breaks from experimenting, which get progressively longer. In the public information case (actions and experimental outcomes are both observable), we exhibit a class of non-stationary equilibria in which the ε-efficient amount of experimentation is performed, but only in infinite time. In the private information case (only actions are observable, not outcomes), the breaks have two additional effects: not only do they enable the players to finesse the inference problem, but also they serve to signal their experimental outcome to the other player. We describe an equilibrium with similar non-stationary strategies in which the ε-efficient amount of experimentation is again performed in infinite time, but with a faster rate of information acquisition. The equilibrium rate of information acquisition is slower in the former case because the short-run temptation to free-ride on information acquisition is greater when information is public.

Keywords: Experimentation, Public Goods.

JEL Classification Numbers: C73, D83, H41, O32.

*Our thanks for very helpful suggestions are owed to Dirk Bergemann, Thomas Mariotti and Jeroen Swinkels, and to seminar participants at the IGIER Workshop in Economic Theory, Erasmus University Rotterdam, the University of Warwick, the Stanford GSB Brown Bag Lunch, the University of Pennsylvania, and the University of Wisconsin–Madison.

†Department of Economics, University of Warwick, Coventry CV4 7AL, UK.

‡Department of Economics, LSE, Houghton Street, London WC2A 2AE, UK.

§Graduate School of Business, Stanford University, Stanford, CA 94305-5015, USA, and Department of Economics, University of Munich, Kaulbachstrasse 45, D-80539 Munich, Germany.

1 Introduction

This paper describes a game of strategic experimentation in which the players learn from the experimental outcomes of others as well as from their own. We contrast two possible models, one where a player can directly observe her opponent's experimental outcomes and a second where a player's ability to learn from others is limited, because experimental outcomes are private. In the latter case, a player can draw some inference about her opponent's outcomes by studying the actions he takes. For example, when he uses one action for a long time she can infer that this action is favourable without undertaking a costly experiment herself, so when outcomes are private, his actions determine the information he himself acquires and also how this information is signalled to the other players. We find the following results in such games: first, a level of experimentation which is approximately socially efficient can be induced when the players have public or private information provided the players use strategies which slowly acquire information. For this to hold, it is important for players to take breaks in experimenting which not only slow the rate of information acquisition but also signal information in the private information case. We call these "coffee breaks." Secondly, a move from public to private information will generally increase the amount of experimentation the individuals perform and the rate at which the information is acquired. The increase in the amount is because of the delay in the inference about the opponent's outcomes when they are private. The rate is slower under public information because the short-run temptation to free-ride is greater in that case.

If the results of research are public and two players fund independent experiments, then they are each providing a public good (information) to the other. They must decide how long to continue providing this public good, given they have the option of free-riding on the other's costly information, and the free-riding problem causes the players to underfund the acquisition of information. The benchmark case of strategic experimentation with public information has been well examined by Bolton and Harris (1999), in a more complex environment than that used in this paper.[1] They consider a two-armed bandit problem with Brownian noise; we consider a two-armed bandit problem with Poisson-type uncertainty, and derive comparable results. These are of interest in their own right, and are used as building blocks when we address the issue of non-stationary strategies in the case first of public information and then of private information.

When there is public information there is an obvious state variable in this model – the players' common belief – but there is no obvious state variable in the model with private information because the players' beliefs are not commonly known and, in general, the entire past history of her opponent's play is relevant in determining a player's belief about her opponent's experimental outcomes. Thus, when we move from public to private information it is necessary for us to consider equilibria where the players use more complex history-dependent strategies. We wish to compare the history-dependent equilibria that naturally arise in the game with private information with the equilibria

[1] They consider equilibria where the players use stationary Markovian strategies (with the common 'level of optimism' as the state variable) and show that, because players can immediately see each others' outcomes and can adjust their actions to ensure an individually optimal rate of information acquisition, there is an inefficient amount of experimentation in models with public information.

that arise in the game with public information, and it is not correct to compare these history-dependent equilibria with the stationary Markovian equilibria of the game with public information (otherwise we are changing both the class of strategies we allow the players to use and a change in the information structure). Our benchmark should be the outcomes of the game with public information where the players can also use history-dependent strategies. However, because work has concentrated on equilibria of the game where the players use stationary Markovian strategies, very little is known about the equilibria of the game where the players can use more sophisticated actions. Consequently, the first completely novel result in this paper describes equilibria in the game with public information where the players use history-dependent strategies.

We find that, if experimental outcomes are public information, then there are (non-stationary) equilibria that approach the efficient level and rate of experimentation. These equilibria are characterised by a gradual evolution of cooperation by the two players. The players adjust the rate at which information is acquired, so that the future benefits from participating in the current round of experimentation always outweigh the current costs. For this to happen it must be the case that the players never actually stop experimenting. The equilibria we build divide continuous time into discrete intervals of length Δ. In each interval the players begin by performing some experimentation, and then they stop and use a safe action until the end of the interval. The amount they experiment in each interval of length Δ shrinks to zero, so the rate at which information is acquired converges to zero. The experimentation is costly for the players and in the short run they would rather use the safe action; however, if they fail to perform the experimentation in any interval their opponent never experiments in the future. Thus there is a long-run cost to not abiding by this strategy, because they forgo all the future information that would be generated by their opponent's actions. Provided the strategy is carefully chosen it is possible to make this long-run cost greater than any short-run gain from deviation. It is important for the technical restrictions on strategies in continuous time that players do not have to respond immediately after their opponent deviates. The equilibrium construction does not rely on this – all a player needs to verify is that at the end of the interval (of length Δ) her opponent experimented in the early part of the interval. The reader will notice that two elements are essential for this equilibrium to work: there must be an infinite time horizon and time must be continuous. If there is a point in the future when the players will stop experimenting, then the long-run benefits vanish and the short-run costs force a player to stop experimenting now. Thus this equilibrium can only survive if the long-run level of experimentation is never actually attained. If time is discrete, then players can provide only discrete chunks of information to their opponent at a strictly positive cost. This equilibrium requires the players to make infinitely many arbitrarily small experiments at arbitrarily small cost which is impossible in discrete time. This result is similar to those in dynamic models of the private provision of public goods. It has been noted that an efficient level of the provision of the public good may be possible in the long run if the players are allowed to gradually make smaller contributions (for example, Admati and Perry (1991), Marx and Matthews (1997), Lockwood and Thomas (1999)) and the provision of the public good is irreversible. In those models time is discrete and the players can continuously vary their (irreversible) contributions to the public good. Information is a natural candidate for those models – communicating information is clearly irreversible; however, no previous

model has treated this explicitly.

Once the investigation of equilibria with public experimental outcomes is complete we move on to studying the game with private experimental outcomes. First note that because the coffee breaks have a pre-determined length, the amount of time spent experimenting in any interval is independent of the experimental outcomes, and so the players cannot deduce anything from each other's actions during an activity phase. This breaks the infinite regress of having beliefs about beliefs about .... Next, when experimental outcomes are private each player cannot immediately free-ride on the other's costly research. Her ability to infer her opponent's outcomes is limited by the way his actions signal his outcomes, so the rate of free-riding on experimental effort is endogenous. In particular there is a delay in the players correctly interpreting their opponent's actions. This delay in acquiring information from others leads one to expect there to be more experimentation when the players have private information. Two general principles will apply: (a) If a player observes her opponent engaged in research she will generally revise her belief for successful experimentation upwards as time passes, even when she is not using the action herself. (This, of course, assumes that her opponent will generally choose to experiment more in favourable states of the world.) (b) If a player observes her opponent switching from action R to action S, say, her belief that action R is profitable will move downwards. This is because her knowledge of her opponents strategy and the timing of the switch will give her a lot of information about her opponent's experimental results before the switch. As an example of these processes consider an observer monitoring the behaviour of an individual experimenting with Brownian motion with an uncertain drift, as in Berry and Fristedt (1985, Chapter 8). While the individual uses the risky action, knowledge of the optimal experimental strategy would lead an outside observer to revise upwards her belief for a high drift. The time at which the experimenter moves from the Brownian motion stream to the safe profit stream, when combined with knowledge of the experimenter's (deterministic) strategy, gives the observer all the relevant information from the experiments – without having to perform the experiments herself! In general, these two methods of extracting information from an opponent's experimental strategy will be very effective in uncovering information about an opponent's experimental results. However, unlike the case of direct observation of experimental results, an outside observer will not be able to extract accurate information about experimental results immediately. It is this delay in obtaining accurate information (and the consequent delay in free-riding on others' information provision) that generally increases the amount of experimentation performed in games of strategic experimentation with private information, because players can overcome this delay by experimenting more on their own account.

Two types of dynamic signalling device arise in models with private experimental outcomes. The first is called the "coffee break" mechanism, which arises if both researchers simultaneously take short pre-determined "coffee breaks" from using the risky action if they have been unsuccessful. If a player does not show up for coffee then the other deduces that he has been successful and revises her belief accordingly. This sort of dynamic signalling is self-enforcing if the coffee breaks are sufficiently short – neither player benefits by lying and thereby convincing her opponent that the state is good, because if she does this she will get no more information from her opponent. (In fact

each player would be willing to pay a small cost to be able to truthfully signal by taking a coffee break.) As the players can take breaks very often, the coffee break equilibrium can be used to construct games where players signal their information very frequently. This provides a class of equilibria for the game of strategic experimentation with private information which approximate the equilibria of games of strategic experimentation with public information.

The second type of signalling device (not considered further in this paper) arises when the players alternate in their experimentation, so that one player experiments whilst the other does not and then they switch roles; but a player continues experimenting if she has been successful. Thus, if the players continue to alternate, their actions reveal their private information at discrete points in time, and if the alternation is frequent the asymmetry in the agents' information is also reduced. Thus alternating experimentation is an effective dynamic signalling device. In the limit the players' actions will "chatter" and this chattering approximates perfect communication.

Section 2 is expository: it sets up a simple bandit model and describes the optimal strategy. We establish the efficient benchmark where the players co-ordinate in order to maximise joint expected payoffs, and then show that, because of free-riding, the strategic problem leads to inefficiently low levels of experimentation in any equilibrium when the players use stationary Markovian strategies. The completely novel results of the paper appear in Sections 3 and 4, where we introduce the coffee breaks and consider non-stationary equilibria in which the ε-efficient amount of experimentation is performed, first in the case of public information, and then in the case of asymmetric information when experimental outcomes are private. In Section 5 we discuss whether the results can be extended to more elaborate models. Some of the proofs are relegated to the Appendix.

2 Poisson Bandits

The purpose of this section is to describe the solution to a simple, continuous-time two-armed bandit problem. One arm S is 'safe' and yields a known deterministic *flow* payoff whenever it is played; the other arm R is 'risky' and can be either 'bad' or 'good'. If it is bad, then it always yields 0; if it is good, then it yields a known *lump-sum* reward at random times whenever it is played – the lump-sums arriving according to a Poisson process. We assume that the agent strictly prefers R, if it is good, to S, and strictly prefers S to R, if it is bad, so she has a motive to experiment with the risky action in the hope of discovering that R is indeed good. The problem she faces, however, is that when she plays R she cannot immediately tell whether it is good or bad, because in either case she initially receives no payoff at all, and the longer she waits without getting a lump-sum, the less optimistic she becomes. Of course, if she does eventually receive a lump-sum then she becomes certain that R is good and she will continue with R forever, but if she waits and waits without the lump-sum arriving then there will come a time when it is optimal for her to cut her losses and switch irrevocably to S.

More formally, time $t \in [0, \infty)$ is continuous, and the discount rate is $r > 0$. The

known *flow* payoff of the safe arm is s; the known *lump-sum* payoff of the risky arm if it is good is h, the parameter of the Poisson process which determines the arrival of the lump-sums is λ, and so the expected payoff from a good arm is equivalent to a *flow* payoff of λh. We assume that $0 < s < \lambda h$.

If an agent plays S over a period of time dt then her payoff is $s\,dt$, and if she plays R over this period then her expected payoff is $\lambda\mu\,dt$, where $\mu \in \{0, h\}$ is unknown. Thus, if k indicates her current choice between S ($k = 0$) and R ($k = 1$), then her expected current payoff (conditional on the unknown state μ of the risky arm) is $[(1-k)s + k\lambda\mu]\ dt$. Her overall objective is to choose a strategy $\{k_t\}_{t\geq 0}$ that maximises

$$\mathrm{E}\left[\int_0^\infty r\,e^{-r\,t}\ [(1-k_t)s + k_t\lambda\mu]\ dt \mid p_0\right],$$

which expresses the payoff in per-period terms. Of course, this choice of strategy is subject to the constraint that the action taken at any time t be measurable with respect to the information available at that time.

Let p_t denote the subjective probability at time t that the agent assigns to the risky arm being good, so that $p_t\lambda h$ is her current expectation of the flow equivalent of playing R. By the Law of Iterated Expectations, we can rewrite the above payoff as

$$\mathrm{E}\left[\int_0^\infty r\,e^{-r\,t}\ [(1-k_t)s + k_t p_t\lambda h]\ dt \mid p_0\right].$$

This highlights the potential for beliefs to serve as a state variable.

Were an agent to act myopically over a period of time dt, she would weigh the short-run payoff from playing S, $rs\,dt$, against what she expects from playing R, $rp\lambda h\,dt$. So let us define p^m as the belief that makes her indifferent between these choices,

$$p^m = \frac{s}{\lambda h}.$$

For $p > p^m$ it is myopically optimal to play R; for $p < p^m$ it is myopically optimal to play S. As we shall see below, a forward-looking agent (who values information) continues to play R for some beliefs $p < p^m$, and is said to *experiment*.

We shall consider the cases where there is a single agent, where there are N agents playing as a team, and where there are N players who act strategically but use only Markovian strategies with the state variable being the belief p.

2.1 The single-agent problem

When S is played over a period of time dt, the belief does not change. When R is played over a period of time dt, the lump-sum h arrives with probability $\lambda\,dt$ if the risky arm is good,[2] and the posterior belief jumps to 1; no payoff arrives with probability $1 - \lambda\,dt$ if the risky arm is good, and with probability 1 if the risky arm is bad. If the agent starts

[2] This is up to terms of the order $o(dt)$, which we can ignore here and in what follows.

with the belief p, plays R over a period of time dt and does not obtain a reward, then the updated belief at the end of that time period is

$$p + dp = \frac{p\,(1 - \lambda\,dt)}{1 - p + p\,(1 - \lambda\,dt)}$$

by Bayes' rule. Simplifying, we see that the belief changes by

$$dp = -\lambda p(1 - p)\,dt\,.$$

We now derive the agent's Bellman equation. By the Principle of Optimality, the agent's value function satisfies

$$u(p) = \max_{k \in \{0,1\}} \left\{ r\left[(1-k)s + k\lambda h p\right]\,dt + e^{-r\,dt}\mathrm{E}\left[u(p + dp) \mid p_0\right] \right\}$$

where the first term is the expected current payoff and the second term is the discounted expected continuation payoff.

As to the expected continuation payoff, with subjective probability $pk\lambda\,dt$ the lump-sum arrives and the agent expects a flow payoff of λh in the future; with probability $p(1 - k\lambda\,dt) + (1 - p) = 1 - pk\lambda\,dt$ no lump-sum arrives and she expects $u(p) + u'(p)dp = u(p) - k\lambda p(1 - p)u'(p)\,dt$.

Using $1 - r\,dt$ to approximate $e^{-r\,dt}$, we see that her discounted expected continuation payoff is

$$(1 - r\,dt)\,\left\{u(p) + k\lambda p[\lambda h - u(p) - (1 - p)u'(p)]\,dt\right\}$$

and so her expected total payoff is

$$u(p) + r\left\{(1-k)s + k\lambda h p + k\lambda p[\lambda h - u(p) - (1-p)u'(p)]/r - u(p)\right\}dt\,.$$

When this is maximised it equals $u(p)$. Simplifying and rearranging, we thus obtain the Bellman equation

$$u(p) = \max_{k \in \{0,1\}} \left\{(1-k)s + k\lambda h p + k\lambda p[\lambda h - u(p) - (1-p)u'(p)]/r\right\}\,.$$

Note that the maximand is affine in k, and that the agent is indifferent between the two options when $s - \lambda h p = \lambda p[\lambda h - u(p) - (1-p)u'(p)]/r$, each option resulting in $u(p) = s$. Thus she is effectively unrestricted by the discrete nature of her choice; as usual, there is no scope for randomisation in this single-agent decision problem.

So, when it is optimal to play S ($k^* = 0$), $u(p) = s$ as one would expect; and when it is optimal to play R ($k^* = 1$), u satisfies the first-order ODE

$$\lambda p(1 - p)u'(p) + (r + \lambda p)u(p) = (r + \lambda)\lambda h p, \tag{1}$$

which has the solution

$$V_1(p) = \lambda h p + C\,(1 - p)\left(\frac{1 - p}{p}\right)^{r/\lambda} \tag{2}$$

with C being the constant of integration.

Proposition 2.1 *In the single-agent problem, there is a cut-off belief p_1^* given by*

$$p_1^* = \frac{rs}{(r+\lambda)(\lambda h - s) + rs} < p^m \tag{3}$$

such that below the cut-off it is optimal to play S and above it is optimal to play R. The value function V_1^* for the single-agent is given by

$$V_1^*(p) = \lambda h p + (s - \lambda h p_1^*)\left(\frac{1-p}{1-p_1^*}\right)\left(\frac{1-p}{p}\right)^{r/\lambda}\left(\frac{p_1^*}{1-p_1^*}\right)^{r/\lambda} \tag{4}$$

when $p > p_1^*$, and $V_1^*(p) = s$ otherwise.

PROOF: The expression for p_1^* is derived by using the values $u(p_1^*) = s$ and $u'(p_1^*) = 0$ in equation (1). Note that for any solution u of equation (1), at any p such that $u(p) = s$, it is the case that $u'(p) < 0$ if $p < p_1^*$ and that $u'(p) > 0$ if $p > p_1^*$.

Playing S when $p \in [0, p_1^*)$ gives a payoff of s; since $V_1^*(p) = s$ on that interval, playing R on any interval to the left of p_1^* would give a payoff less than s and is therefore sub-optimal. Playing R when $p \in (p_1^*, 1]$ gives a payoff greater than s; playing S on any interval to the right of p_1^* would give a payoff of s and is therefore also sub-optimal.

Using $V_1^*(p_1^*) = s$ in equation (2) determines the constant of integration C, giving the expression for V_1^*. ■

This solution exhibits all of the familiar properties, which were elegantly described in Rothschild (1974): the optimal strategy has a threshold where the experimenter switches irrevocably from R to S; there are occasions where the experimenter makes a mistake by switching from R to S although the risky action is actually better (R is good); the probability of mistakes decreases as the experimenter becomes more patient, and as the expected reward from the risky action increases.

2.2 The N-agent team problem

Now suppose that there are $N \geq 2$ identical agents (same prior belief, same discount rate), each with identical two-armed bandits, who are working as a team, i.e. they want to maximise the *average* expected payoff. Information is public: the players can observe each other's actions and outcomes, so the players' beliefs remain identical throughout time.

If K of them play R over a period of time dt, then, if a lump-sum arrives they all switch to R else their belief decays K-times as fast. Whenever that arm is good, the probability of *none* of them getting a lump-sum is $(1 - \lambda\,dt)^K = 1 - K\lambda\,dt$, the probability of *exactly one* of them getting a lump-sum is $K\lambda\,dt(1 - \lambda\,dt)^{K-1} = K\lambda\,dt$, and the probability of *more than one* of them getting a lump-sum is negligible.[3]

[3] Again, we are ignoring terms of order $o(dt)$.

Lemma 2.1 *In the N-agent team problem, it is optimal either for* all *players to play R or for* none *of them to do so.*

PROOF: Let u be the value function for the team problem, expressed as average payoff per team member. When the current belief is p and the current choice is for K agents to play R, the average expected current payoff is $r\left[(1-\frac{K}{N})s+\frac{K}{N}\lambda hp\right]\,dt$. Paralleling the calculation for the single-agent problem, we see that the discounted expected continuation payoff is

$$(1-r\,dt)\,\{u(p)+K\lambda p[\lambda h-u(p)-(1-p)u'(p)]\,dt\}$$

and so the average expected total payoff is

$$u(p)+r\left\{(1-\tfrac{K}{N})s+\tfrac{K}{N}\lambda hp+K\lambda p[\lambda h-u(p)-(1-p)u'(p)]/r-u(p)\right\}dt\,.$$

Thus the value function satisfies the Bellman equation

$$u(p)=\max_{K\in\{0,1,\ldots,N\}}\left\{(1-\tfrac{K}{N})s+\tfrac{K}{N}\lambda hp+K\lambda p[\lambda h-u(p)-(1-p)u'(p)]/r\right\}\,.$$

Note that the maximand is affine in K, and that the team is indifferent between all levels of K when $s-\lambda hp=N\lambda p[\lambda h-u(p)-(1-p)u'(p)]/r$, all of them resulting in $u(p)=s$. Thus at all beliefs either $K^*=N$ or $K^*=0$ is optimal. ■

So, when it is optimal for them all to play S, $u(p)=s$ as usual; and when it is optimal for them all to play R, u satisfies

$$N\lambda p(1-p)u'(p)+(r+N\lambda p)u(p)=(r+N\lambda)\lambda hp \tag{5}$$

which is like equation (1) with λ replaced by $N\lambda$ (reflecting an N-times faster rate of information acquisition), and h replaced by h/N (reflecting the fact that lump-sum rewards are shared amongst the N team members). This has the solution

$$V_N(p)=\lambda hp+C\,(1-p)\left(\frac{1-p}{p}\right)^{r/N\lambda}\,. \tag{6}$$

Proposition 2.2 *In the N-agent team problem, there is a cut-off belief p_N^* given by*

$$p_N^*=\frac{rs}{(r+N\lambda)(\lambda h-s)+rs}<p_1^* \tag{7}$$

such that below the cut-off it is optimal for all to play S and above it is optimal for all to play R. The value function V_N^* for the N-agent team is given by

$$V_N^*(p)=\lambda hp+(s-\lambda hp_N^*)\left(\frac{1-p}{1-p_N^*}\right)\left(\frac{1-p}{p}\right)^{r/N\lambda}\left(\frac{p_N^*}{1-p_N^*}\right)^{r/N\lambda} \tag{8}$$

when $p>p_N^*$, and $V_N^*(p)=s$ otherwise.

PROOF: As in the proof of Proposition 2.1, p_N^* is determined by setting $u(p_N^*) = s$ and $u'(p_N^*) = 0$ in equation (5). Determining the constant of integration C as before gives the expression for V_N^*. ■

The above proposition determines the *efficient* experimentation strategies for N players acting as a team. We can distinguish two aspects of efficiency here. Given a strategy profile $\{(k_{1,t}, \ldots, k_{N,t})\}_{t\geq 0}$ for the team members, the integral $\int_0^\infty \sum_{n=1}^N k_{n,t}\, dt$ measures the overall amount of time that a risky arm was used. We will call this number the *amount* of experimentation that is performed. On the other hand, the sum $K_t = \sum_{n=1}^N k_{n,t}$ measures how many risky arms are used at a given time t. We will call this number the *intensity* of experimentation.

Proposition 2.2 shows that the efficient amount of experimentation is N times the time it takes for the agents' common belief to decay to p_N^* when all players use the risky arm all the way through. A simple calculation shows that for priors $p_0 > p_N^*$ this efficient amount is $[\ln \frac{1-p_N^*}{p_N^*} - \ln \frac{1-p_0}{p_0}]/\lambda$. The efficient intensity of experimentation exhibits a bang-bang feature, being N when the current belief is above p_N^*, and 0 when it is below. Thus, the efficient intensity is maximal at early stages, and minimal later on.

As we shall see next, equilibria of the N-player *strategic* problem are never efficient. Although it is possible to generate the efficient amount of experimentation in such an equilibrium, the intensity of experimentation will always be inefficient because of each player's incentive to free-ride on the efforts of the others.

2.3 The N-player strategic problem – pure strategies

We continue to assume that the players have the same prior belief, the same discount rate, identical two-armed bandits, and that information is public. We consider stationary Markovian pure strategies with the common belief as the state variable. Our concise treatment largely reflects the approach taken in Bolton and Harris (1993, 1999).[4]

Let $k_n \in \{0, 1\}$ indicate the current choice of player n between S ($k_n = 0$) and R ($k_n = 1$); let $K = \sum_{n=1}^N k_n$ and $K_{\neg n} = K - k_n$, so that $K_{\neg n}$ summarises the current choices of the other players. Taking into account the information generated if the other players play R, we see that player n's value function satisfies the Bellman equation

$$u_n(p) = \max_{k_n \in \{0,1\}} \{(1 - k_n)s + k_n \lambda h p + (k_n + K_{\neg n})\lambda p[\lambda h - u_n(p) - p(1-p)u_n'(p)]/r\} \ .$$

Immediately we see that the best response, $k_n^*(p)$, is determined by comparing the opportunity cost of playing R, $s - \lambda h p$, with $\lambda p \left[\lambda h - u_n(p) - (1-p)u_n'(p)\right]/r$:

$$k_n^*(p) \begin{cases} = 0 & \text{if } s - \lambda h p > \lambda p \left[\lambda h - u_n(p) - (1-p)u_n'(p)\right]/r \, , \\ \in \{0,1\} & \text{if } s - \lambda h p = \lambda p \left[\lambda h - u_n(p) - (1-p)u_n'(p)\right]/r \, , \\ = 1 & \text{if } s - \lambda h p < \lambda p \left[\lambda h - u_n(p) - (1-p)u_n'(p)\right]/r \, . \end{cases}$$

[4]The reader is referred to this work for a careful and detailed description of many stationary Markov equilibria in the case where the agents sample Brownian motion with an unknown drift.

If the best response is to play R ($k_n^* = 1$) then player n's value function u_n satisfies

$$K\lambda p(1-p)u'(p) + (r + K\lambda p)u(p) = (r + K\lambda)\lambda hp \tag{9}$$

with $K = K_{\neg n} + 1$.[5] The solution to (9) is

$$V_K(p) = \lambda hp + C\,(1-p)\left(\frac{1-p}{p}\right)^{r/K\lambda}. \tag{10}$$

If the best response is to free-ride by playing S ($k_n^* = 0$) then u_n satisfies

$$K\lambda p(1-p)u'(p) + (r + K\lambda p)u(p) = rs + K\lambda^2 hp \tag{11}$$

with $K = K_{\neg n}$. The solution to (11) is

$$F_K(p) = s + \frac{K\lambda(\lambda h - s)}{r + K\lambda}p + C\,(1-p)\left(\frac{1-p}{p}\right)^{r/K\lambda}. \tag{12}$$

Finally, for $K_{\neg n} > 0$, player n is indifferent if and only if $u_n(p) = s + K_{\neg n}(s - \lambda hp)$. Note that the equation $u = (K+1)s - K\lambda hp$ defines a diagonal line $\mathcal{D}_K$ in the (p,u)-plane which cuts the safe payoff line $u = s$ at $p = p^m$, the myopic switch-point. If the graphs of $F_{K_{\neg n}}$ and $V_{K_{\neg n}+1}$ meet $\mathcal{D}_{K_{\neg n}}$ at the same belief p_c then $F'_{K_{\neg n}}(p_c) = V'_{K_{\neg n}+1}(p_c)$, which is a manifestation of the usual smooth-pasting property.

2.4 The N-player strategic problem – inefficiency

Using the above results, we can now show the following.

Proposition 2.3 *All Markov Perfect Equilibria of the N-player strategic game are inefficient.*

PROOF: All we need to show is that the efficient strategies from Proposition 2.2 are *not* an equilibrium. Suppose therefore that players $1, \ldots, N-1$ use the risky arm at beliefs above the cut-off p_N^* and the safe arm below. If player N adopts the same strategy, her payoff function is V_N^*. If, for some $\epsilon > 0$, she deviates by switching to S on the interval $]p_N^*, p_N^* + \epsilon]$, the restriction of her payoff function u_N to this interval solves (11) with $u_N(p_N^*) = s$. Evaluating (11) at p_N^* shows that

$$(N-1)\lambda p_N^*(1-p_N^*)u'_N(p_N^*) = (N-1)\lambda p_N^*(\lambda h - s) > 0.$$

[5] Note that equation (9) for the strategic problem is the same ODE as that for the team problem with K players; cf. equation (5). To see why, suppose that the risky arm is good. Then, whenever K agents play the risky arm, a lump-sum arrives with probability $K\lambda\,dt$ over the next instant. In the K-agent team problem, this lump-sum is shared amongst K players, so the expected lump-sum reward over the next instant is $\frac{h}{K}K\lambda\,dt = h\lambda\,dt$ per player. In the strategic problem, the lump-sum arrives with probability $\lambda\,dt$ on player n's arm and with probability $(K-1)\lambda\,dt$ on someone else's arm. Since player n keeps her own lump-sum in full and receives no share of someone else's, her expected lump-sum reward is also $h\lambda\,dt$.

Since $u_N(p_N^*) = V_N^*(p_N^*) = s$ and $u_N'(p_N^*) > (V_N^*)'(p_N^*) = 0$, the deviation is clearly profitable for beliefs sufficiently close to p_N^*. (In fact, by choosing ϵ sufficiently small, player N can achieve a payoff function $u_N > V_N^*$ on the *whole* of $]p_N^*, 1[$.) ■

Thus, the incentive to free-ride on the experimentation efforts of the other players makes it impossible to reach efficiency. In the following two subsections, we turn to the question as to what *can* be achieved in Markov perfect equilibria. We shall consider symmetric mixed strategy equilibria of the N-player game and asymmetric pure strategy equilibria of the 2-player game. In Section 3, we shall then show that equilibria in non-stationary strategies can come arbitrarily close to efficiency.

2.5 The N-player strategic problem – symmetric MPE

Since the efficient strategy profile is symmetric and Markovian with the belief as state variable, it is natural to ask what outcomes can be achieved in symmetric Markovian equilibria of the N-player game. We maintain the assumptions of the previous subsections, but allow for mixed strategies now. Following Bolton and Harris (1999), we actually consider the time-division game in which agent n allocates a fraction κ_n of the current period $[t, t+dt)$ to R, and the remainder to S; this is isomorphic to the player using the mixed strategy that places probability κ_n on playing R, and the remainder on S.

So, let $\kappa_n \in [0,1]$ indicate the current decision of player n, $K = \sum_{n=1}^{N} \kappa_n$, and $K_{\neg n} = K - \kappa_n$. Once again taking into account the information generated by the other players, we see that player n's value function satisfies the Bellman equation

$$u_n(p) = \max_{\kappa_n \in [0,1]} \left\{ (1-\kappa_n)s + \kappa_n \lambda h p + (\kappa_n + K_{\neg n})\lambda p \left[\lambda h - u_n(p) - (1-p)u_n'(p)\right]/r \right\} .$$

Again the best response, $\kappa_n^*(p)$, is determined by comparing the opportunity cost of experimentation, $s - \lambda h p$, with $\lambda p \left[\lambda h - u_n(p) - (1-p)u_n'(p)\right]/r$:

$$\kappa_n^*(p) \begin{cases} = 0 & \text{if } s - \lambda h p > \lambda p \left[\lambda h - u_n(p) - (1-p)u_n'(p)\right]/r \,, \\ \in [0,1] & \text{if } s - \lambda h p = \lambda p \left[\lambda h - u_n(p) - (1-p)u_n'(p)\right]/r \,, \\ = 1 & \text{if } s - \lambda h p < \lambda p \left[\lambda h - u_n(p) - (1-p)u_n'(p)\right]/r \,. \end{cases}$$

In any Markov perfect equilibrium player n's value function will be defined piecewise: when all the time is devoted to S it satisfies equation (11) with $K = K_{\neg n}$ and is of the form $F_{K_{\neg n}}$; when all the time is devoted to R it satisfies equation (9) with $K = K_{\neg n} + 1$ and is of the form $V_{K_{\neg n}+1}$; and when the time is divided strictly between S and R it satisfies

$$\lambda p(1-p)u'(p) + \lambda p u(p) = (r+\lambda)\lambda h p - rs, \tag{13}$$

which has the solution

$$W(p) = s + \frac{r+\lambda}{\lambda}(\lambda h - s) + \frac{rs}{\lambda}(1-p)\,\ln\left(\frac{1-p}{p}\right) + C\,(1-p). \tag{14}$$

Note that when $K_{\neg n} > 0$, player n is indifferent if and only if $(p, W_n(p)) \in \mathcal{D}_{K_{\neg n}}$; and if the graphs of W_n and $V_{K_{\neg n}+1}$ meet $\mathcal{D}_{K_{\neg n}}$ at the same belief p_c then $W_n'(p_c) = V_{K_{\neg n}+1}'(p_c)$.

Our next result describes the unique symmetric Markov perfect equilibrium of the strategic experimentation game. Let

$$\Omega(p) = \frac{1-p}{p}$$

denote the 'odds ratio' corresponding to the belief p.

Proposition 2.4 (N players, symmetric strategies) *In the N-player time-division game with public information, there is a unique symmetric equilibrium in Markovian strategies with the common posterior belief as the state variable. In this equilibrium, all time is devoted to the safe arm at beliefs below the single-player cut-off p_1^*; all time is devoted to the risky arm at beliefs above a cut-off $\tilde{p}_N > p_1^*$ solving*

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

and a positive fraction of time is devoted to each arm at beliefs strictly between p_1^* and $\tilde{p}_N$. The fraction of time that each player allocates to the risky arm at such a belief is

$$\kappa^*(p) = \frac{1}{N-1}\left(\frac{W^*(p)-s}{s-\lambda h p}\right) \tag{15}$$

with

$$W^*(p) = s + \frac{rs}{\lambda}\left[\Omega(p_1^*)\left(1 - \frac{1-p}{1-p_1^*}\right) - (1-p)\,\ln\left(\frac{\Omega(p_1^*)}{\Omega(p)}\right)\right]. \tag{16}$$

PROOF: Suppose that all N players using the Markov strategy $\kappa : [0,1] \to [0,1]$ constitutes an equilibrium of the time-division game with common value function u. Let $p_c = \inf\{p \in [0,1] : u(p) > s\}$, so that $u(p) = s$ and $\kappa(p) = 0$ on $[0, p_c]$, and $u(p) > s$ and $\kappa(p) > 0$ on $]p_c, 1]$. It is easy to see that $p_c \leq p_1^*$, for if we had $p_c > p_1^*$, it would be a profitable deviation to switch from S to R on the interval $]p_1^*, p_c[$.

On the interval $]p_c, 1[$ the value function satisfies an ODE that is a combination of (13) to the left of $\mathcal{D}_{N-1}$ and (9) with $K = N$ to the right of $\mathcal{D}_{N-1}$. Relevant to us are all solutions of this ODE whose graphs in the (p,u)-plane lie in the triangle with corners $(0,s)$, (p^m,s) and $(1,\lambda h)$. It is straightforward to see that all these solutions can be parameterised by the point where they cross the diagonal $\mathcal{D}_{N-1}$ and that, depending on this point, two possibilities arise: either the solution stays strictly above the level s or it reaches this level. Exactly one solution has a point of tangency with the level s; let p_c' be the corresponding belief. All other solutions that reach the level s coming down and to the left from $\mathcal{D}_{N-1}$ do so at a belief to the right of p_c'. Thus, the cut-off p_c that we defined for the equilibrium value function satisfies $p_c' \leq p_c \leq p_1^*$.

Now, p_c' is determined by setting $u(p_c') = s$ and $u'(p_c) = 0$ in the ODE (13). As this equation differs from (1) only by having s instead of $u(p)$, it is immediate that $p_c' = p_1^*$, the single-player cut-off given in Proposition 2.1. This in turn proves that $p_c = p_1^*$.

Finally, using $W(p_1^*) = s$ in equation (14) determines the constant of integration C, giving the expression (16) for the value function over the range where both actions are used for a positive fraction of time. Given this function, the expression (15) for the share of time κ^* allocated to R follows from each player's Bellman equation and (13). As W^* is strictly convex, κ^* is strictly increasing to $+\infty$ as $p \uparrow p^m$. Thus there is a unique cut-off $\tilde{p}_N < p^m$ where $\kappa^*(\tilde{p}_N) = 1$. Simplifying $W^*(\tilde{p}_N) - s = (N-1)(s - \lambda h \tilde{p}_N)$ gives the equation satisfied by $\tilde{p}_N$. ■

Several points are noteworthy. First, the lower cut-off belief at which all experimentation in the symmetric MPE stops does not depend on the number of players; quite surprisingly, it equals the optimal cut-off from the single-player problem. This means that we do not have the *encouragement effect* of Bolton and Harris (1999). In their model, an agent who on his own would be indifferent between the two actions, strictly prefers the risky action when other players are present. In fact, his own experimentation may produce favourable information that will make everybody more optimistic, and thus encourage the other players to perform some more experimentation themselves from which the first player will eventually benefit. Note that it is crucial for this argument that there be a sufficiently good chance for beliefs to become more optimistic over the next instant. This is the case in Bolton and Harris' framework where changes in beliefs are driven by the increments of a Wiener process, so that the chance of an upward revision of beliefs over a short time interval stays bounded away from zero even as the length of this interval shrinks to zero. With our Poisson bandits, beliefs can only become more optimistic if a success occurs, and the chance of this happening over a given time interval goes to zero with the length of the interval.[6] Almost surely, beliefs an instant later will be more pessimistic, so there is no hope that an extra bit of experimentation might be reciprocated by the other players.

Second, the expected equilibrium payoff that each player obtains at beliefs where both arms are used a positive fraction of time does not depend on the number of players either; cf. equation (16). The reason for this is that the relevant ODE, equation (14), is just the indifference condition of a *single* player, and that the boundary condition at the lower cut-off belief is the same for any number of players. Put differently, the combined intensity of experimentation by $N-1$ players when both arms are used a positive fraction of time is independent of the total number of players, N; cf. equation (15). Over that range of beliefs, therefore, a player's best response and payoff do not depend on N either. What does depend on N is the upper cut-off belief, of course: with more players, the temptation to free-ride becomes stronger, and $\tilde{p}_N$ increases. Formally, this is most easily seen from equation (15): given that $\kappa^*(\tilde{p}_N) = 1$ and W^* is a strictly increasing function, $\tilde{p}_N$ must increase with N. Informally, the indifference diagonal $\mathcal{D}_{N-1}$ rotates clockwise as N increases.

Third, the proposition implies that there is no symmetric MPE in pure strategies. In fact, any candidate for such an equilibrium unravels because of free-riding at lower beliefs. What sort of behaviour can arise in a pure-strategy MPE will be addressed next. For ease of exposition, we restrict ourselves to the two-player case from now on.

[6] Technically speaking, the difference is that the belief process in Bolton and Harris (1999) is of infinite variation, while ours is of finite variation.

2.6 The 2-player strategic problem - pure-strategy equilibria

We will present two types of asymmetric Markov perfect equilibrium in pure strategies. The first type of MPE consists of strategies where the action of each player switches at finitely many beliefs. As a consequence, there is a last point in time at which any player is willing to experiment. As in the symmetric MPE, the belief at which this happens (provided no success has been observed) will be the single-player cut-off p_1^*. So the same inefficiency arises: both the amount and the intensity of experimentation are too low.

In the second type of MPE, each player's strategy has infinitely many switching points, and although there is a finite time after which no player ever experiments again, no single player has a *last* time for experimentation. That is, immediately prior to reaching a certain cut-off belief, the players switch roles increasingly fast, and infinitely often. We will see that we can take this cut-off belief to be the efficient cut-off p_2^* that would be chosen by a two-player team. Still, the equilibrium is inefficient: although the efficient amount of experimentation is performed, it is performed with an inefficient intensity.

With finitely many beliefs at which a player changes his action, a Markov perfect equilibrium has three phases. When the players are optimistic ($p > \bar{p}_r$), both play R; when they are pessimistic ($p \leq p_1^*$), both play S; in between, one of them free-rides by playing S while the other is playing R. This mid-range of beliefs further splits into two regions. The roles of free-rider and 'lone ranger' are assigned for the whole of the upper region ($p > \bar{p}_\ell$); in the lower region ($p \leq \bar{p}_\ell$), players can swap roles. Note that the lower threshold belief at which all experimentation stops is the single-agent cut-off; in particular, it is the same for all equilibria of this type, whereas the higher threshold beliefs are determined endogenously by how the burden of experimentation is shared in the lower region $]p_1^*, \bar{p}_\ell]$.

The proposition below first describes the 'simplest' such equilibrium, in which one particular player experiments and the other free-rides throughout the lower region. This equilibrium exhibits the least amount of experimentation, in that the part of the state space where both players experiment is smallest, i.e. the high threshold is as close to 1 as it can be in an equilibrium of this type. We characterise the thresholds using the notation $\Omega(p) = \frac{1-p}{p}$ again.

Proposition 2.5 (Two players, pure strategies, finite number of switches) *In the two-player strategic experimentation problem with public information, there is a pure-strategy Markov perfect equilibrium where the players' actions depend as follows on the common posterior belief. There are three cut-off beliefs $p_1^* < \tilde{p}_\ell < \tilde{p}_r$ such that: on $(\tilde{p}_r, 1]$, both players play R; on $(\tilde{p}_\ell, \tilde{p}_r]$, player 1 plays R and player 2 plays S; on $(p_1^*, \tilde{p}_\ell]$, player 1 plays S and player 2 plays R; on $[0, p_1^*]$, they both play S. The low cut-off, p_1^*, is given in Proposition 2.1; the other two are given by the solution to*

$$\left(\frac{\Omega(\tilde{p}_\ell)}{\Omega(p_1^*)}\right)^{r/\lambda+1} + \frac{r+\lambda}{\lambda}\left[\frac{\Omega(\tilde{p}_\ell)}{\Omega(p^m)} - 1\right] - 1 = 0$$

and the solution to

$$\left\{\frac{(r+\lambda)(2r+\lambda)}{r\lambda}\,\frac{\Omega(\tilde{p}_\ell)}{\Omega(p^m)}-\frac{r^2+(r+\lambda)(r+2\lambda)}{r\lambda}\right\}\left(\frac{\Omega(\tilde{p}_r)}{\Omega(\tilde{p}_\ell)}\right)^{r/\lambda+1}+\frac{r+\lambda}{\lambda}\left[\frac{\Omega(\tilde{p}_r)}{\Omega(p^m)}-1\right]-1=0.$$

Moreover, in any pure-strategy MPE with finitely many switching points there are three cut-off beliefs $p_1^* < \bar{p}_\ell \leq \bar{p}_r$, with $\tilde{p}_\ell \leq \bar{p}_\ell$ and $\bar{p}_r \leq \tilde{p}_r$, such that: on $(\bar{p}_r, 1]$, both players play R; throughout $(\bar{p}_\ell, \bar{p}_r]$, one player plays R and the other plays S; on $(p_1^*, \bar{p}_\ell]$, the players share the burden of experimentation; on $[0, p_1^*]$, they both play S.

PROOF: See the Appendix. ■

Note that in the 'simplest' and, at the same time, 'worst' equilibrium (with cut-offs p_1^*, $\tilde{p}_\ell$ and $\tilde{p}_r$) the player who uses the risky arm longer has the lower expected payoff. As the burden of experimentation is shared more and more equally in the lower region of the mid-range of beliefs, the upper region of this range shrinks – we approach payoff symmetry but not the payoffs of the 2-player symmetric MPE of Proposition 2.4, since here the fraction of time each player allocates to the risky arm approaches $\frac{1}{2}$ for the entire region of strict mixing.

If we allow players to switch between actions at infinitely many beliefs, they can take turns experimenting in such a way that no player ever has a last time (or lowest belief) at which he is supposed to use the risky arm. Surprisingly, it is possible to reach cut-off beliefs below p_1^* in such an equilibrium. In fact, it is possible to attain the efficient cut-off p_2^*, but it is reached too slowly.

Proposition 2.6 (Two players, pure strategies, infinite number of switches) *There is a strictly decreasing sequence of beliefs $\{p_i^\dagger\}_{i=0}^\infty$ with $p_0^\dagger = p_1^*$ and $\lim_{i\to\infty} p_i^\dagger = p_2^*$ such that the following pure strategies constitute a Markov perfect equilibrium of the two-player strategic experimentation problem with public information: on $]p_1^*, 1]$, both players play R; on $]p_{i+1}^\dagger, p_i^\dagger]$, player 1 plays R and player 2 plays S if i is even, whereas player 1 plays S and player 2 plays R if i is odd; on $[0, p_2^*]$, they both play S.*

PROOF: To be added. ■

A simple calculation shows that the amount of experimentation performed in this equilibrium equals $[\ln\frac{1-p_2^*}{p_2^*} - \ln\frac{1-p_0}{p_0}]/\lambda$, which is the efficient amount according to our remark after Proposition 2.2. However, the intensity of experimentation is efficient only at times before p_1^* and after p_2^* is reached; at times in between, it is 1, and therefore first too low, then too high relative to the efficient benchmark.

3 Non-Stationary Equilibria with Public Information

We will now move on to describing a class of equilibria where the players do not use Markovian strategies. There are three main reasons for doing this. The first is that we are interested in comparing equilibria of the game with public information to equilibria of the game with private information. In the game with private information there are no natural state variables, so we are obliged to consider equilibria where the players use strategies that are more complex and history-dependent. To make a fair comparison, therefore, we ought to consider equilibria of the game with public information when the players are able to use more sophisticated strategies. The second reason is that the equilibria we construct for games with private information have a natural analogue in games with public information, so we would be remiss in not describing these. Thirdly, models of strategic experimentation with public information are essentially models of private provision of public goods and there is already a literature on this topic showing that non-Markovian strategies are useful in generating efficiency.[7]

Focusing on the two-player case, we construct a family of equilibria of the experimentation game with public information where, through the use of non-Markovian strategies, players achieve payoffs arbitrarily close to the efficient level given by the team solution. On the equilibrium path the players will both use action R as long as their common belief satisfies $p \geq p_1^*$. When their common belief is below p_1^*, time is partitioned into a sequence of finite intervals of equal length (except possibly for the very first interval, which can be shorter if the player's belief at the beginning of the game is already below p_1^*). On the equilibrium path, both players experiment during the first part of each interval; if, during that time, no player has received a reward, both players take a "coffee break" from using the risky arm during the remaining part of the interval. The length of the breaks is strictly bounded away from zero and grows from interval to interval, expanding in such a way that the players' belief cannot reach the efficient cut-off p_2^* in finite time. A deviation from the risky to the safe arm during the first part of such an interval is deterred by the opponent's credible threat never to use the risky arm again.

The notation used to describe this non-stationary equilibrium is as follows. The length of the time intervals that come into play at beliefs below p_1^* is $\Delta > 0$. The ith interval ($i = 0, 1, 2, \ldots$) consists of an "activity phase" of length a_i followed by a break of length $b_i = \Delta - a_i$. We impose an upper bound $\bar{a} < \Delta$ on the experimentation time, which means a lower bound on the length of the coffee breaks.

The sequence of intervals with their activity phases and coffee breaks leads to a decreasing sequence of beliefs $\hat{p}_i$ ($i = 0, 1, \ldots$) that, as long as no success is observed, are reached consecutively on the equilibrium path. We anchor this sequence at the one-player cut-off by setting $\hat{p}_0 = p_1^*$. We then define recursively:

$$\hat{p}_{i+1} = \frac{\hat{p}_i \exp(-2\lambda a_i)}{1 - \hat{p}_i + \hat{p}_i \exp(-2\lambda a_i)}.$$

Thus, if the players enter the ith activity phase with belief $\hat{p}_i$ and both use the risky

[7] E.g., Marx and Matthews (1997).

arm for the length of time a_i, their belief at the beginning of the ith coffee break is $\hat{p}_{i+1}$. The sequence $\{\hat{p}_i\}$ decreases and is bounded below by zero, so there is a limiting belief, denoted by $\hat{p}_\infty$. To deal with possible deviations at beliefs above p_1^*, we will need to introduce two more beliefs, $\hat{p}_{-1}$ and $\hat{p}_{-1/2}$. These will be determined endogenously once the sequence of the a_i is given. For the moment, all we need to know is that $\hat{p}_{-1} > \hat{p}_{-1/2} > p_1^*$.

Given the interval length Δ and sequences $\{a_i\}_{i=0}^{\infty}$ and $\{\hat{p}_i\}_{i=-1,-1/2,0,1,\ldots}$ as above, we specify the players' strategies as follows. Always assuming that no success has been observed so far, we build these strategies recursively from a few basic steps labelled $[Start]$, $[-2]$, $[-2D]$ $(D = 1, 2)$, $[-1]$, $[-1D]$ $(D = 1, 2)$, $[i]$ $(i = 0, 1, \ldots)$, and $[End]$. The ideas behind these steps are simple. $[Start]$ initialises play by positioning the current common belief with respect to the grid given by the beliefs $\{\hat{p}_i\}$. If the current belief is still above $\hat{p}_{-1}$, we enter Step $[-2]$ where both players are supposed to play the risky arm until their common belief has decayed to $\hat{p}_{-1}$. If one of the players, say player D, deviates during Step $[-2]$, he is eventually punished in Step $[-1D]$, but this punishment is delayed (Step $[-2D]$) as long as the risky arm is still dominant for either player; this is the case as long as the common belief remains above $\hat{p}_{-1/2}$. If the current belief is between $\hat{p}_{-1}$ and $\hat{p}_0$, we enter Step $[-1]$ where both players are supposed to play the risky arm until their common belief has decayed to $\hat{p}_0 = p_1^*$. If one of the players, say player D, deviates during Step $[-1]$, he is punished in Step $[-1D]$. This step has player D play the risky arm whereas the other player free-rides on D's effort. The punishment phase only ends when the common belief has reached p_1^*. Once beliefs are at or below the one-player cut-off, we go through Steps $[i]$ corresponding to the intervals described earlier. Players are supposed to play the risky arm in the first part of each interval, and the safe arm in the second part. If a player deviates in the activity phase of an interval, he is punished in Step $[End]$ where both players switch to the safe arm for good. Deviations during a break do not trigger punishments: if one player performs more experimentation than required, his results are evaluated at the end of the break, beliefs are updated accordingly, and play is re-initialised in Step $[Start]$.

More precisely, the steps and the transitions between them are:

• $[Start]$: If the current belief is strictly greater then $\hat{p}_{-1}$, go to Step $[-2]$. If the current belief is in $[\hat{p}_{-1}, \hat{p}_0[$, go to Step $[-1]$. If the current belief is in $[\hat{p}_i, \hat{p}_{i+1}[$ with $i \geq 0$, go to Step $[i]$.

• $[-2]$: Let the current time be t. Let τ_{-1} be the time needed for the common belief to reach $\hat{p}_{-1}$ when both players use R all the way. The strategy of both players prescribes R on $[t, t + \tau_{-1}[$. If $p_{t+\tau_{-1}} = \hat{p}_{-1}$, go to Step $[-1]$. If $p_{t+\tau_{-1}} > \hat{p}_{-1}$ and only player D deviated (i.e., played S on a set of times of positive measure during $[t, t + \tau_{-1}[$), go to Step $[-2D]$. If $p_{t+\tau_{-1}} > \hat{p}_{-1}$ and both players deviated, go to $[Start]$.

• $[-2D]$: Let the current time be t. Let $\tilde{\tau}_{-1/2}$ be the time needed for the common belief to reach $\hat{p}_{-1/2}$ when both players use R all the way. The strategy of both players prescribes R on $[t, t + \tilde{\tau}_{-1/2}[$. If $p_{t+\tilde{\tau}_{-1/2}} = \hat{p}_{-1/2}$, go to Step $[-1D]$. If $p_{t+\tilde{\tau}_{-1/2}} > \hat{p}_{-1/2}$ and only player D has ever deviated, repeat Step $[-2D]$. If $p_{t+\tilde{\tau}_{-1/2}} > \hat{p}_{-1/2}$ and both players have now deviated, go to $[Start]$.

• $[-1]$: Let the current time be t. Let τ_0 be the time needed for the common belief to reach $\hat{p}_0$ when both players use R all the way. The strategy of both players prescribes R on $[t, t+\tau_0[$. If $p_{t+\tau_0} = \hat{p}_0$, go to Step [0]. If $p_{t+\tau_0} > \hat{p}_0$ and only player D deviated (i.e., played S on a set of times of positive measure during $[t, t+\tau_0[$), go to Step $[-1D]$. If $p_{t+\tau_0} > \hat{p}_0$ and both players deviated, go to $[Start]$.

• $[-1D]$: Let the current time be t. Let $\tilde{\tau}_0$ be the time needed for the common belief to reach $\hat{p}_0$ when exactly one player uses R all the way. The strategy of player D (the deviator being punished) prescribes R on $[t, t+\tilde{\tau}_0[$. The strategy of player $\neg D$ (the one who has carried out the prescribed experimentation in step $[-1]$) prescribes S on $[t, t+\tilde{\tau}_0[$. If $p_{t+\tilde{\tau}_0} = \hat{p}_0$, go to Step [0]. If $p_{t+\tilde{\tau}_0} > \hat{p}_0$, repeat Step $[-1D]$.

• $[i]$: Let the current time be t. Let τ_i be the time needed for the common belief to reach $\hat{p}_{i+1}$ when both players use R all the way. (Unless the previous step was $[Start]$, $\tau_i = a_i$.) The strategy of both players prescribes R on $[t, t+\tau_i[$ and S on $[t+\tau_i, t+\tau_i+b_i[$. (Recall that $b_i = \Delta - a_i$ is the length of the ith coffee break.) If $p_{t+\tau_i} = \hat{p}_{i+1}$, go to $[Start]$ at time $t+\tau_i+b_i$. If $p_{t+\tau_i} > \hat{p}_{i+1}$ and only one player deviated on $[t, t+\tau_i[$ by playing S on a subset of positive measure, go to $[End]$ at time $t+\tau_i+b_i$. If $p_{t+\tau_i} > \hat{p}_{i+1}$ and both players deviated in the above sense on $[t, t+\tau_i[$, go to $[Start]$ at time $t+\tau_i+b_i$.

• $[End]$: The strategy of each player prescribes S forever. All further deviations are ignored.

To show that these strategies form an equilibrium, it will suffice to show that no player has an incentive to deviate in any given step, assuming that the above strategies are followed in all subsequent steps (this is a variant of the familiar one-stage-deviation principle). At beliefs above the single-player cut-off p_1^*, i.e. in Steps $[-2]$, $[-2D]$, $[-1]$ and $[-1D]$, robustness to deviations will follow from arguments that build on the Markovian case developed in Section 2. The same holds for Step $[End]$. As to Steps $[i]$, robustness to deviations requires that the sequence $\{a_i\}_{i=0}^{\infty}$ (or, equivalently, $\{\hat{p}_i\}_{i=1}^{\infty}$) be chosen in a particular way. Given $\bar{a}$, in fact, we will find a natural number I such that $a_i = \bar{a}$ for $i \leq I$, and $a_i < \bar{a}$ otherwise. Thus, in the early intervals, the above strategies have both players experiment as much as possible, and they will strictly prefer following their strategy to deviating. For all intervals $i > I$, the players experiment less than the maximum amount, and they will be just indifferent between following their strategy and the most advantageous deviation. This indifference imposes a restriction on the beliefs $\{\hat{p}_i\}_{i=I+1}^{\infty}$ in form of a second-order difference equation. We can use this equation to anchor the sequence "at infinity", i.e. show existence of a stable solution and provide bounds on the limit belief $\hat{p}_\infty$, and then work backwards towards the endogenously determined I.

The bounds on $\hat{p}_\infty$ will show that, as Δ tends to zero and the "maximal activity ratio" $\bar{a}/\Delta$ tends to one, the limit belief tends to the efficient cut-off p_2^*. This is quite intuitive: As the intervals become shorter and the intensity of experimentation in the early stages becomes higher, the players can achieve an amount of experimentation, measured by the sum of the lengths of time that the players are using the risky arm, arbitrarily close to the efficient amount, which is twice the time needed for the initial belief to decay to p_2^* when both players use the risky arm all the way through. In the early stages of the equilibrium,

moreover, the players can achieve an intensity of experimentation arbitrarily close to the efficient rate, which is 2. Taken together, these two facts imply that as intervals and coffee breaks become shorter, the players' equilibrium payoffs converge to the team solution.

We have the following result.

Proposition 3.1 *For any Δ and $\bar{a}$ with $0 < \bar{a} < \Delta$, there exists a non-increasing sequence of positive real numbers $\{a_i\}_{i=0}^{\infty}$ such that the strategies defined in Steps [Start], [−2], [−2D], [−1], [−1D] ($D = 1, 2$), [i] ($i = 0, 1, ...$), and [End] constitute a perfect Bayesian equilibrium of the experimentation game with public information. The outcome of this equilibrium is that the players both use the risky arm whenever their common belief is above or at the single-player cut-off p_1^*; otherwise, they alternate jointly between activity phases (where both use the risky arm) of diminishing length $\leq \bar{a}$ and coffee breaks (where both use the safe arm) of increasing length $\geq \Delta - \bar{a}$. Provided no success is observed, the common belief in this equilibrium converges to a limit $\hat{p}_\infty$; as $(\Delta, \frac{\bar{a}}{\Delta}) \to (0, 1)$, the limit belief $\hat{p}_\infty$ tends to p_2^* and the players' equilibrium payoffs converge to the two-player team payoffs.*

PROOF: See the Appendix, where we establish the sequence $\{a_i\}_{i=0}^{\infty}$ and show that

$$\Omega(p_2^*) - (r\Delta + 2\lambda\bar{a})\frac{\lambda}{r}\Omega(p^m) \leq \Omega(\hat{p}_\infty) < \Omega(p_2^*) - r\Delta\frac{\lambda}{r}\Omega(p^m)$$

when Δ and $\bar{a}$ are close to 0. This clearly implies that $\hat{p}_\infty \to p_2^*$ as $(\Delta, \frac{\bar{a}}{\Delta}) \to (0, 1)$. ■

The fact that these non-Markovian equilibria of the game with public information are arbitrarily close to efficient, stands in sharp contrast to the inefficient Markovian equilibria of the previous section. This result is due to the introduction of (infinitely many) coffee breaks, which promote cooperation. In fact the breaks are arranged so that the short-run benefit of free-riding never exceeds the long-run cost. That is, the information to be gained from free-riding today is less important than the information a player knows she will get in the future if her opponent continues to play out the equilibrium. In particular, the current payoff from a deviation is small because the use of R within each interval is shrinking sufficiently fast relative to the remaining amount of experimentation.

4 Equilibria with Private Information

In this section the players' experimental outcomes are private information. Nevertheless, we show that there is an equilibrium where the players' strategies are similar to those described in the previous section. Again, the equilibrium payoffs can be made arbitrarily close to the efficient ones.

When the players' experimental outcomes are private, but their actions are observable, the inference problem faced by players is severe and intricate. In general all orders of belief will be important in solving this inference problem. First, consider one player

observing her opponent's actions. In general, the longer he is observed to use the risky action, R, the more likely it is that he has received a prize h. Thus without knowing his experimental results, she will tend to revise her belief about the state upwards as he uses R for longer. If the observer is using R at the same time as making these observations, however, the fact that she is using R itself encourages her opponent to use R and to disregard his own information. If she is to infer the probability of a success correctly, she must recognise that her own actions will have an encouraging effect on her opponent and reduce the importance he attaches to his own information. This second-order effect can lead to extreme inferences. In some cases she will get no information at all from his use of R, because he is basing his actions solely on her behaviour. At the same time, he is choosing to ignore his own information at the expense of the information he believes he is getting from his opponent. This is a form of herding.

There are two key intuitions about how the players use actions to signal, which are extensively used in the equilibrium constructed below. Both of these intuitions again arise out of the coffee breaks. First, if one player chooses to use R for a fixed amount of time independent of the experimental outcomes, then her opponent cannot deduce anything from his observations of her actions. By Bayes' Theorem, if a player's actions are independent of a random variable then his actions must be uninformative about that random variable. Thus, a fixed time spent using R in equilibrium will force a player to disregard the actions of his opponent and to obtain information only from his own experimental outcomes. The coffee breaks divide the players' use of R into fixed time chunks that are independent of their immediate experimental results, which gives players an incentive to collect observations on their own account. The second role they play is to coordinate the signalling. At the end of a fixed time a_i using R the players either switch to S for an amount of time b_i or they continue to use R if they had a prize h. Thus, at the start of a coffee break a player's action will signal his experimental outcomes to his opponent. The experimental outcomes of the players are signalled in a simultaneous burst of information – either no success thus far, or a success. The breaks are the points at which the players' accumulated information is exchanged. The exchange of information at the end of the fixed period of experimentation delays the inferences the players make about the information of their opponent and thereby delays their ability to free-ride on the information acquired by their opponent. They must wait until the next coffee break until they can infer what their opponent has observed.

One important question that must be addressed is: do the players have an incentive to signal correctly when a coffee break arises? First, does a player have an incentive to use S during the break when he has received a prize h? Certainly not, because once a prize has been received it is a strictly dominant strategy to use R at every instant of time. Beliefs in the equilibrium below will indeed be such that a deviation from R to S will be considered evidence of failure so far. Second, does a player have an incentive to continue to use R during the break even when he has not received a prize h during the previous a_i periods? The answer to this question is also no, because of the herding effect described in the earlier paragraph. In fact, the equilibrium that we construct will be supported by the belief that whoever deviates in this particular way must have received a prize in the preceding activity phase. In other words, if a player misrepresents his information in this way his opponent will infer that the state is good and then use R

forever (disregarding her own observations to the contrary). This harms the deviator, because his opponent's signals are now entirely uninformative and he cannot obtain any further information from her actions.

The main result in this section is Proposition 4.1. It shows that the equilibrium described in Proposition 3.1 essentially carries over to the case where experimental outcomes are private information, although there are different constraints on the sequence $\{a_i\}$. In the early stages of play both players use R until their 'pooled' belief (which each player calculates assuming that the other has not had a prize so far) approaches the threshold p_1^*, then they use R for pre-determined short periods of time followed by coffee breaks if no success has been observed. We will see that given a sequence $\{a_i\}$ of lengths of the activity phases, the payoff from a deviation is lower in the game with private information than in the game with public information. When a player deviates in the game with public information he can continue to observe his opponent's experimental outcomes. However, under private information the player cannot immediately observe what his opponent's outcomes are. Instead he must wait until the next coffee break to see whether his opponent continues to use R or abandons this entirely. Thus there is a delay in his ability to free-ride on her information acquisition. This delay makes deviation less attractive under private information. As the payoff to a deviation from the equilibrium with private information is lower, the players should be willing to experiment more in the short run at this equilibrium. This suggests that for given Δ and $\bar{a}$, the equilibria with private information should have a higher intensity of experimentation, and hence be more efficient, than the equilibria with public information.

The construction of equilibrium strategies follows that in the public information case very closely, with common beliefs replaced by pooled beliefs when it comes to the decision when to stop for a break. One difference is that a deviation from S to R during a break is not simply ignored; rather, it triggers a reaction whereby the other player revises his belief to certainty that the state is good and consequently switches to R forever.

Proposition 4.1 *For any Δ and $\bar{a}$ with $0 < \bar{a} < \Delta$, there exists a non-increasing sequence of positive real numbers $\{a_i\}_{i=0}^{\infty}$ such that the following is the outcome of a perfect Bayesian equilibrium of the game with private information. The players both use the risky arm whenever their pooled belief is above or at the single-player cut-off p_1^*; otherwise, they alternate jointly between activity phases (where both use the risky arm) of diminishing length $\leq \bar{a}$ and coffee breaks (where both use the safe arm) of increasing length $\geq \Delta - \bar{a}$. Provided no success is observed, the common belief in this equilibrium converges to a limit $\hat{p}_\infty$; as $(\Delta, \frac{\bar{a}}{\Delta}) \to (0,1)$, the limit belief $\hat{p}_\infty$ tends to p_2^* and the players' equilibrium payoffs converge to the two-player team payoffs.*

PROOF: See the Appendix, where we again establish the sequence $\{a_i\}_{i=0}^{\infty}$ and show that

$$\Omega(p_2^*) - (r\Delta + 2\lambda\bar{a})\frac{\lambda}{r}\Omega(p^m) \leq \Omega(\hat{p}_\infty) < \Omega(p_2^*) - r\Delta\frac{\lambda}{r}\Omega(p^m)$$

when Δ and $\bar{a}$ are close to 0. This clearly implies that $\hat{p}_\infty \to p_2^*$ as $(\Delta, \frac{\bar{a}}{\Delta}) \to (0,1)$. ■

One issue that arises in games of private information is the formation of beliefs off the equilibrium path. If player 1 deviates from the above equilibrium by using S instead of R,

player 2 infers that player 1's experimentation before the deviation had been unsuccessful. Although, this inference seems intuitively appealing it may be that the equilibrium is not robust to a different inference from such a deviation. If beliefs were revised so that such a deviation leads player 2 to conclude that player 1's experiments were successful, then player 2 would play R forever believing that it was good. Player 1 then derives no further benefit from the presence of player 2 and cannot see player 2's experimental results. The deviation would be very harmful to player 1 and the equilibrium would be robust to this alternative inference. The worst belief player 2 can have (from player 1's perspective) is to be convinced the state is good, because then player 1 gets no more information from player 2. Using these beliefs off the equilibrium path makes building equilibria much easier, but it is counter-intuitive: if player 1 really had had a success then she could never benefit by deviating; however, if her experimentation had been unsuccessful she could conceivably benefit. This argument suggests that the alternative beliefs revision process is inconsistent with the spirit of a number of equilibrium refinements. Consequently, we do not consider equilibria that use this type of belief revision.

On the other hand, in the above equilibrium a player witnessing a deviation from S to R is supposed to conclude that the deviator has had a success in the previous activity phase. A jump in beliefs to full subjective certainty seems an extreme assumption to make. In fact, this assumption is not necessary: a belief jumping sufficiently close, but not all the way, to certainty will do as well. A deviation from S to R will then imply that the other player switches to R for a finite period of time. All we have to ensure is that this period is sufficiently long to deter this type of deviation.

5 How General are these Results?

There are some trivial and obvious generalisations of the model that follow with very little additional work. First, none of the arguments requires only two players so the results are robust to many players. Secondly, the results apply to bandit problems where the known arm generates a stationary non-deterministic stream of payoffs – we can simply reinterpret s as the expected flow payoff.

At first sight the results appear to be very dependent on the specification of the process for the risky arm. However, a more careful investigation reveals that suitably amended versions of the propositions apply to other bandit problems too. Suppose that the risky arm is in one of two states both of which generate Poisson rewards, and that in the good state the Poisson rewards arrive more frequently than in the bad state. In this case, whenever a reward arrives the belief jumps upwards but never hits unity; when a reward does not arrive the belief decays, as before. With public information a version of the coffee-break equilibrium can be built: for high beliefs both players use R and for lower beliefs they use the risky arm intermittently, the payoff from doing this again being the future experimental input from their opponent and the possibility of an upward jump in the belief. A direct translation of the equilibrium can be constructed provided the arrival of a reward leads the belief to jump above the one-player threshold p_1^*; if this fails it is necessary to define a new coffee-break equilibrium that is played out after each upward jump. With private information it is harder to generalise the results

here in a simple way. The intuition that the fixed period spent using the risky arm is uninformative still holds true, but in this case the players in general need to signal not only whether a reward has arrived but also the number of rewards, and it is harder to see how the players can signal the *number* of rewards when they have only two actions – one resolution is to make the intervals small, in which case the probability of two or more arrivals in any activity phase is negligible relative to the probability of there being just one arrival. The next issue is the incentive-compatibility of the signalling: would a player necessarily want to turn up for a coffee break? By signalling an arrival a player will in general lead her opponent to use R for a longer time without pooling any information. This will be harmful to the signaller and may well deter false arrival signals provided the signaller cannot herself free-ride on this extra experimental effort from her opponent. Taking a break when an arrival occurred is also harmful because the break requires the player to take a sub-optimal action. Thus at least in the domain of Poisson bandits the results do seem to generalise.

The final generalisation we consider is to a *multi*-armed bandit. In this case the players can take the equilibrium strategies above and apply them to the safe arm and one of the risky arms at a time, using the safe arm for the coffee breaks. Instead of switching to the safe arm at the end of the trials with the sole risky arm, they just switch to the next risky arm that needs to be investigated according to the familiar Gittins index rule.

Appendix

Proof(s) from Section 2

Proof of Proposition 2.5: First note that each player's value function is continuous as a function of p and takes the value λh at $p = 1$ and s at $p = 0$; moreover it is differentiable *wherever he/she chooses optimally to switch* (from playing R to playing S, or *vice versa*) and the other player does not switch – if the right derivative is smaller, the player should switch at a larger p; if the right derivative is larger, the player should switch at a smaller p.

Our aim is to show that the region bounded below by the myopic payoff in the (p, u)-plane contains three regions, as in Figure 1. In one region (when the players are optimistic) it is dominant for each of them to play R and in another region (when the players are pessimistic and $u = s$) it is dominant for each of them to play S; in between, S and R are mutual best responses.

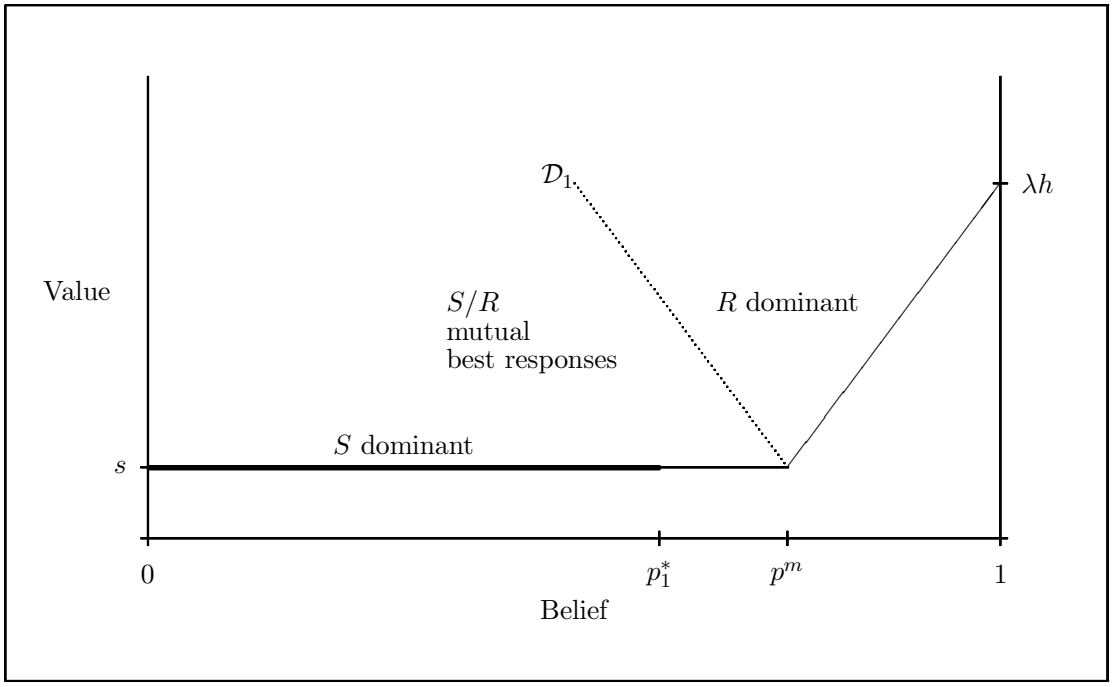

Figure 1: Three regions

The solid kinked line is the payoff from the myopic strategy.

Assume that player n's continuation value is given by $u_n(p)$.

• Assume that player A (she) is playing R when the belief is in some interval $[p_\ell, p_r]$, and consider the best response of player B (he) on $[p_\ell, p_c] \subseteq [p_\ell, p_r]$. If it is also R then his value function on $[p_\ell, p_c]$ is given by V_2 from equation (10) with $V_2(p_\ell) = u_B(p_\ell)$; if his best response is S then his value function on $[p_\ell, p_c]$ is given by F_1 from equation (12) with $F_1(p_\ell) = u_B(p_\ell)$. Now, if $V_2(p) = F_1(p) = u$, say, then $V_2'(p) > F_1'(p)$ if $u > 2s - \lambda hp$, and $V_2'(p) < F_1'(p)$ if $u < 2s - \lambda hp$. Thus, if $u_B(p_\ell) > 2s - \lambda hp_\ell$, then his best response to R is to "join in" by playing R on $[p_\ell, p_r]$; if $u_B(p_\ell) < 2s - \lambda hp_\ell$, then his best response to R is to free-ride by playing S on

$[p_\ell, p_c]$ for any p_c such that $F_1(p_c) < 2s - \lambda h p_c$; and he can only switch optimally at a belief $p_c \in [p_\ell, p_r]$ where $(p_c, u_B(p_c)) \in \mathcal{D}_1$.

• Now, assume that player A (she) is playing S when the belief is in some interval $[p_\ell, p_r]$, and consider the best response of player B (he) on $[p_\ell, p_c] \subseteq [p_\ell, p_r]$. If it is R then his value function on $[p_\ell, p_c]$ is given by V_1 from equation (2) with $V_1(p_\ell) = u_B(p_\ell)$; if his best response is also S then the belief no longer changes, so it must be the case that $u_B(p_\ell) = s$ and his value function on $[p_\ell, p_c]$ is simply s. Now, if $V_1(p) = s$, then $V_1'(p) > 0$ if $p > p_1^*$, and $V_1'(p) < 0$ if $p < p_1^*$. Thus, if $u_B(p_\ell) = s$, then his best response to S is to act unilaterally: if $p_\ell > p_1^*$ then play R on $[p_\ell, p_r]$; if $p_\ell < p_1^*$ then play S on $[p_\ell, p_c]$ for any p_c such that $p_c < p_1^*$; and he can only switch optimally at the belief p_1^*. However, if $u_B(p_\ell) > s$, then his best response to S must be to play R on $[p_\ell, p_r]$ (but note that $V_1'(p) < 0$ if $(r + \lambda p)V_1(p) > (r + \lambda)\lambda h p$).

Let $\bar{p}_r$ denote the smallest belief where each player's continuation value is (weakly) above $\mathcal{D}_1$, and let $\bar{p}_\ell$ denote the largest belief where each player's continuation value is (weakly) below $\mathcal{D}_1$; necessarily, $p_1^* < \bar{p}_\ell \leq \bar{p}_r < p^m$.

For a belief in a neighbourhood of 1, specifically $p \in (\bar{p}_r, 1]$, R is the dominant strategy; and for a belief in a neighbourhood of 0, specifically $p \in [0, p_1^*]$, S is the dominant strategy. (We know that $u_n(0) = s$, and so S is a dominant response on any interval $[0, p_c] \subseteq [0, p_1^*]$). For beliefs $p \in (p_1^*, \bar{p}_\ell]$, the best response to S is to play R (act unilaterally), and the best response to R is to play S (free-ride). Now consider beliefs $p \in (\bar{p}_\ell, \bar{p}_r]$; let A be the player whose continuation value crosses $\mathcal{D}_1$ at $\bar{p}_\ell$ and let B be the player whose continuation value crosses $\mathcal{D}_1$ at $\bar{p}_r$. If B plays S, then A's best response is to play R (act unilaterally), and if B plays R, then A's best response is to play R ("join in"); thus R is the dominant response for A. So, given A plays R, B's best response is to play S (free-ride). To summarise:

Belief p	0	p_1^*	$\bar{p}_\ell$	$\bar{p}_r$	1
A's strategy	S	S/R	R	R	
B's strategy	S	R/S	S	R	
A's continuation value	s	$F_{1,A}/V_{1,A}$	$V_{1,A}$	$V_{2,A}$	
B's continuation value	s	$V_{1,B}/F_{1,B}$	$F_{1,B}$	$V_{2,B}$	

and the strategies on $(p_1^*, \bar{p}_\ell]$ determine $\bar{p}_\ell$ endogenously, which player plays R and which player plays S on $(\bar{p}_\ell, \bar{p}_r]$, and $\bar{p}_r$ endogenously. If the players have the above continuation values, then the above strategies are best responses to each other; and if the players are using the above strategies, then the continuation values are indeed those given above. Thus the above strategies constitute an equilibrium with the equilibrium value functions given by the continuation values.

The 'simplest' equilibrium is where one player, say player 1, plays S on $(p_1^*, \tilde{p}_\ell]$, and the other player, player 2, plays R on this interval. Then player 1's value function F_1 satisfies equation (12) and player 2's value function V_1 satisfies equation (2), with $F_1(p_1^*) = V_1(p_1^*) = s$. So $F_1'(p_1^*) > V_1'(p_1^*)$, since whenever $F_1(p) = V_1(p) = u$, say, $F_1'(p) > V_1'(p)$ iff $\lambda h p < s$, i.e. iff $p < p^m$. Furthermore, it can be shown that F_1 is concave and V_1 is convex[8] and so if F_1 and V_1 take the same value again, say at $p_c > p_1^*$, then $F_1'(p_c) \leq V_1'(p_c)$, which implies that $p_c \geq p^m$. This shows that F_1 meets $\mathcal{D}_1$ at a *smaller* belief than does V_1, and that $F_1 > V_1$ on $(p_1^*, \tilde{p}_\ell]$; that is, player 1 must be A and switch from playing R on $(\tilde{p}_\ell, \tilde{p}_r]$, and player 2 must be B and

[8]It transpires that the second derivative of the functions F_1, V_1 and V_2 has the same sign as the constant of integration (in (12), (2) and (10) respectively) and thus the convexity/concavity of the solution is determined by that sign.

switch from playing S on $(\tilde{p}_\ell, \tilde{p}_r]$. This equilibrium is thus given by:

Belief p	0	p_1^*	$\tilde{p}_\ell$	$\tilde{p}_r$	1
A's strategy	S	S	R	R	
B's strategy	S	R	S	R	
A's value function	s	$F_{1,A}$	$V_{1,A}$	$V_{2,A}$	
B's value function	s	$V_{1,B}$	$F_{1,B}$	$V_{2,B}$	

and the components of the value functions, and the switch-points, are determined as follows:

(1) C in $F_{1,A}$ from $F_{1,A}(p_1^*) = s$

(2) C in $V_{1,B}$ from $V_{1,B}(p_1^*) = s$

(3) $\tilde{p}_\ell$ from $F_{1,A}(\tilde{p}_\ell) = 2s - \lambda h \tilde{p}_\ell$

(4) C in $V_{1,A}$ from $V_{1,A}(\tilde{p}_\ell) = F_{1,A}(\tilde{p}_\ell) = 2s - \lambda h \tilde{p}_\ell$

(5) C in $F_{1,B}$ from $F_{1,B}(\tilde{p}_\ell) = V_{1,B}(\tilde{p}_\ell)$

(6) $\tilde{p}_r$ from $F_{1,B}(\tilde{p}_r) = 2s - \lambda h \tilde{p}_r$

(7) C in $V_{2,A}$ from $V_{2,A}(\tilde{p}_r) = V_{1,A}(\tilde{p}_r)$

(8) C in $V_{2,B}$ from $V_{2,B}(\tilde{p}_r) = F_{1,B}(\tilde{p}_r) = 2s - \lambda h \tilde{p}_r$

Note that the boundary condition at $p = 1$ is automatically satisfied because $V_{2,A}(1) = V_{2,B}(1) = \lambda h$ regardless of the constants of integration.

Noting that when $V_2(p) = V_1(p) = u$, say, $V_2'(p) > V_1'(p)$ iff $u > \lambda h p$ (the payoff from always playing R), we see that

- $0 < F'_{1,A}(p_1^*), \quad F'_{1,A}(\tilde{p}_\ell) > V'_{1,A}(\tilde{p}_\ell), \quad V'_{1,A}(\tilde{p}_r) < V'_{2,A}(\tilde{p}_r)$;
- $0 = V'_{1,B}(p_1^*), \quad V'_{1,B}(\tilde{p}_\ell) < F'_{1,B}(\tilde{p}_\ell), \quad F'_{1,B}(\tilde{p}_r) = V'_{2,B}(\tilde{p}_r)$.

Thus, as the common belief decays, B switches smoothly from R to S against R at $\tilde{p}_r$ (where A has a kink), both A and B switch at $\tilde{p}_\ell$ (each with a kink), and B switches smoothly again from R to S against S at p_1^* (where A again has a kink).

Following steps (1) and (3) determines the equation for $\tilde{p}_\ell$ given in the statement of the proposition; following steps (2), (5) and (6) determines the equation for $\tilde{p}_r$ given in the statement of the proposition; the remaining steps are for completeness only.[9]

Other equilibria for the two-player strategic problem

Any finite partition of the interval to the right of p_1^* can be used to construct a pure strategy equilibrium of the two-player strategic problem.

Take any finite (measurable) partition of $(p_1^*, p^m]$ and divide this into two subsets I_n, $n = 1, 2$. Build the continuous functions X_n on $[p_1^*, p^m]$ as follows: $X_n(p_1^*) = s$, X_n satisfies equation (12) on I_n (free-rider), X_n satisfies equation (2) on $I_{\neg n}$ (lone ranger).

Define $\bar{p}_\ell = \min\{p \in [p_1^*, p^m] : X_1(p) \vee X_2(p) = 2s - \lambda h p\}$. If $X_n(\bar{p}_\ell) \geq X_{\neg n}(\bar{p}_\ell)$ then $A = n$, else $A = \neg n$; $B = \neg A$.

Define $\bar{p}_r$ by $X_B(\bar{p}_r) = 2s - \lambda h \bar{p}_r$, so $\bar{p}_\ell \leq \bar{p}_r$.

Now take the partition $J_1 \cup J_2$ of $(p_1^*, \bar{p}_\ell]$, where $J_n = \{p \leq \bar{p}_\ell : p \in I_n\}$, i.e. J_n and I_n agree on $(p_1^*, \bar{p}_\ell]$.

[9] Details are available from the authors on request.

Let A's strategy be as follows:
play S on $[0, p_1^*]$; play S on J_A and R on J_B; play R on $(\bar{p}_\ell, \bar{p}_r]$; play R on $(\bar{p}_r, 1]$.

Let B's strategy be as follows:
play S on $[0, p_1^*]$; play R on J_A and S on J_B; play S on $(\bar{p}_\ell, \bar{p}_r]$; play R on $(\bar{p}_r, 1]$.

Build the continuous functions Y_n on $[0,1]$ as follows:

$Y_A(p) = s$ on $[0, p_1^*]$; Y_A satisfies equation (12) on J_A (free-rider) and satisfies equation (2) on J_B (lone ranger); Y_A satisfies equation (2) on $(\bar{p}_\ell, \bar{p}_r]$ (lone ranger); Y_A satisfies equation (10) on $(\bar{p}_r, 1]$.

$Y_B(p) = s$ on $[0, p_1^*]$; Y_B satisfies equation (2) on J_A (lone ranger) and satisfies equation (12) on J_B (free-rider); Y_B satisfies equation (12) on $(\bar{p}_\ell, \bar{p}_r]$ (free-rider); Y_B satisfies equation (10) on $(\bar{p}_r, 1]$.

If the continuation values are given by Y_n, then the above strategies are best responses to each other; and if the players are using the above strategies, then the continuation values are indeed given by Y_n. Thus the above strategies constitute an equilibrium with the equilibrium value functions given by Y_n.

Y_A and Y_B lie between $F_{1,A}$ and $V_{1,B} \cup F_{1,B}$ below and to the left of $\mathcal{D}_1$. Thus $\tilde{p}_\ell \leq \bar{p}_\ell \leq \bar{p}_r \leq \tilde{p}_r$, and so the 'simplest' equilibrium exhibits the least experimentation. ■

Proof(s) from Section 3

Proof of Proposition 3.1: We focus on the case where the belief has reached p_1^* and consider Steps $[i]$ $(i = 0, 1, \ldots)$.

Define $x(p) = \Omega(p)/\Omega(p^m)$, the normalised odds ratio when the belief is p. We will establish an increasing convergent sequence $\{x_i\}_{i=0}^{\infty}$ from which we can retrieve the decreasing sequence of beliefs by inverting $x_i = x(\hat{p}_i)$, so $x_0 = \Omega(p_1^*)/\Omega(p^m) = (r+\lambda)/\lambda$. Also define the sequence $\{y_i\}_{i=0}^{\infty}$ by $y_i = x_i/x_{i+1}$, so $y_i = e^{-2\lambda\, a_i}$, from which we can retrieve the length of each "activity phase". The restrictions $a_i \leq \bar{a} < \Delta$ obviously cascade onto restrictions on x_i.

Let $\delta = e^{-r\Delta} < 1$ be the discount factor for an interval of length Δ; let $\beta = e^{-r\bar{a}} > \delta$ be the discount factor for a subinterval of length $\bar{a}$; and let $\gamma = r/2\lambda$, the exponent of the odds ratio in the payoff function of an agent when there are 2 players experimenting.

Let the common belief be $\hat{p}_i$. Assume that player 2 (he) will play R for the next a_i periods, followed by S for the subsequent b_i periods if no lump-sum arrives. (If the other player has done likewise, he will continue; but if the other player has deviated, he will stop and play S forever.) The first step is to calculate the payoff E_i of player 1 (she) when she uses the same strategy, and then her payoff D_i when she deviates by playing S for the first a_i periods.

When both players play R, the common belief at the end of a_i periods will be $\hat{p}_{i+1}$, and her payoff during this time is given by (10), that is

$$E_i(p) = \lambda h p + C_i(1-p)\Omega(p)^\gamma, \quad \hat{p}_{i+1} \leq p \leq \hat{p}_i\,,$$

or equivalently

$$\text{(A.1)} \quad \frac{E_i(p) - s}{s(1-p)} = \frac{1}{x(p)}\left[1 - \left(\frac{x(p)}{x_{i+1}}\right)^{\gamma+1}\right] - \left[1 - \left(\frac{x(p)}{x_{i+1}}\right)^{\gamma}\right] + \left(\frac{x(p)}{x_{i+1}}\right)^{\gamma} \frac{E_i(\hat{p}_{i+1}) - s}{s(1-\hat{p}_{i+1})}\,.$$

When both players switch to S, the common belief at the end of b_i periods will still be $\hat{p}_{i+1}$, but her continuation value will rise to $u_{i+1}(\hat{p}_{i+1})$, where

$$E_i(\hat{p}_{i+1}) = \int_0^{b_i} r\, e^{-rt}\, s\, dt + e^{-r\, b_i}\, u_{i+1}(\hat{p}_{i+1})$$

or equivalently $E_i(\hat{p}_{i+1}) - s = e^{-r\,b_i}(u_{i+1}(\hat{p}_{i+1}) - s)$. Using this in equation (A.1), we see that

$$\frac{E_i(\hat{p}_i) - s}{s(1-\hat{p}_i)} = \frac{1}{x_i}\left[1 - y_i^{\gamma+1}\right] - [1 - y_i^{\gamma}] + y_i^{\gamma} e^{-r\,b_i} \frac{u_{i+1}(\hat{p}_{i+1}) - s}{s(1-\hat{p}_{i+1})}\,,$$

and using the fact that $y_i^{\gamma}\, e^{-r\,b_i} = \delta$ the above equation becomes

$$\frac{E_i(\hat{p}_i) - s}{s(1-\hat{p}_i)} = \frac{1}{x_i}\left[1 - y_i^{\gamma+1}\right] - [1 - y_i^{\gamma}] + \delta \frac{u_{i+1}(\hat{p}_{i+1}) - s}{s(1-\hat{p}_{i+1})}\,. \tag{A.2}$$

If player 1 deviates by playing S (free-riding) for the first a_i periods and if no lump-sum arrives, then player 2 will stop, both players will play S forever and get a payoff of s. Because only one player has been using R, the players' belief has decayed to $p_{i+1/2} = \hat{p}_i e^{-\lambda\, a_i}/(1-\hat{p}_i+\hat{p}_i e^{-\lambda\, a_i})$. Player 1's expected payoff from this deviation, $D_i(p)$, is of the form (12), that is,

$$D_i(p) = s + \frac{\lambda(\lambda h - s)}{r+\lambda} p + B_i(1-p)\Omega(p)^{2\gamma}, \quad p_{i+1/2} \le p \le \hat{p}_i\,,$$

or equivalently

$$\frac{D_i(p) - s}{s(1-p)} = \frac{\lambda}{r+\lambda}\frac{1}{x(p)}\left[1 - \left(\frac{x(p)}{x(p_{i+1/2})}\right)^{2\gamma+1}\right] + \left(\frac{x(p)}{x(p_{i+1/2})}\right)^{2\gamma} \frac{D_i(p_{i+1/2}) - s}{s(1-p_{i+1/2})}\,. \tag{A.3}$$

But $D_i(p_{i+1/2}) = s$ and $x_i/x(p_{i+1/2}) = (x_i/x_{i+1})^{1/2}$, which leads to the payoff from the deviation

$$\frac{D_i(\hat{p}_i) - s}{s(1-\hat{p}_i)} = \frac{\lambda}{r+\lambda}\frac{1}{x_i}\left[1 - y_i^{\gamma+\frac{1}{2}}\right]\,. \tag{A.4}$$

For deviation not to be optimal, $D_i(\hat{p}_i) \le E_i(\hat{p}_i)$ for all i. From equations (A.4) and (A.2), the general condition is

$$\frac{\lambda}{r+\lambda}\frac{1}{x_i}\left[1 - y_i^{\gamma+\frac{1}{2}}\right] - \frac{1}{x_i}\left[1 - y_i^{\gamma+1}\right] + [1 - y_i^{\gamma}] \le \delta\, U_{i+1}(\hat{p}_{i+1}) \tag{A.5}$$

for all i, where the term on the RHS denotes the final term in equation (A.2). The condition (A.5) is sufficient to show that no deviation is optimal, because any deviation that occurs later than the start of an interval yields the deviator a lower payoff from deviation.

Clearly the sequence $x_i = x_0$ (so $y_i = 1$) for all i satisfies condition (A.5), but we shall generate an *increasing* sequence such that x_i (and thus a_i) is as large as possible. It may be that for low values of i condition (A.5) is a strict inequality[10] (when a_i will be at its maximum), and for i larger than some critical value I, condition (A.5) holds with equality (when a_i will be less than its maximum).

The LHS of condition (A.5) is 0 when $x_{i+1} = x_i$, and its first derivative with respect to x_{i+1} has the same sign as $\lambda(x_{i+1}/x_i)^{\frac{1}{2}} - (r+2\lambda) + r x_{i+1}$. When $x_{i+1} = x_i$, this is greater than or equal to 0 iff $r x_i \ge r + \lambda$, and the second derivative with respect to x_{i+1} is strictly positive. Since $r x_0 = r + \lambda$, it follows by induction that the LHS of condition (A.5) is increasing in x_{i+1}. This enables us to define x_{i+1} recursively as follows. Set $x_{i+1} = \beta^{-1/\gamma} x_i$ (corresponding to $a_i = \bar{a}$) if this does not violate condition (A.5), else set $x_{i+1} < \beta^{-1/\gamma} x_i$ (corresponding to $a_i < \bar{a}$) so that condition (A.5) holds with equality.

Define I as the smallest integer such that it holds with equality. For $i < I$ deviation is strictly suboptimal, whereas for $i \ge I$ she is indifferent and so we can use the payoff to

[10] It *should* hold strictly for $i = 0$ if we take δ and β uniformly closer to 1.

a deviation ($D_{i+1}(\hat{p}_{i+1})$ – see equation (A.4)) to replace the continuation payoff (implicitly $u_{i+1}(\hat{p}_{i+1})$) in condition (A.5):

$$\text{(A.6)} \qquad \frac{\lambda}{r+\lambda}\frac{1}{x_i}\left[1-y_i^{\gamma+\frac{1}{2}}\right]-\frac{1}{x_i}\left[1-y_i^{\gamma+1}\right]+\left[1-y_i^{\gamma}\right]=\delta\,\frac{\lambda}{r+\lambda}\frac{1}{x_{i+1}}\left[1-y_{i+1}^{\gamma+\frac{1}{2}}\right]$$

for $i \geq I$.

Consider the equality when $i = I$. As the LHS is increasing in x_{i+1}, if we replace x_{I+1} by $\beta^{-1/\gamma}x_I$ on the LHS and at its first occurrence on the RHS we have the following inequality:[11]

$$\frac{\lambda}{r+\lambda}\frac{1}{x_I}\left[1-\beta^{1+1/2\gamma}\right]-\frac{1}{x_I}\left[1-\beta^{1+1/\gamma}\right]+\left[1-\beta\right]\geq\delta\,\frac{\lambda}{r+\lambda}\frac{\beta^{1/\gamma}}{x_I}\left[1-y_{I+1}^{\gamma+\frac{1}{2}}\right],$$

leading to a lower bound for x_I when δ and β are close to 1 (i.e. when Δ and $\bar{a}$ are close to 0):

$$x_I \geq x(p_2^*)-(r\Delta+2\lambda\bar{a})\,\frac{\lambda}{r}\,.$$

Having constructed the increasing sequence $\{x_i\}_{i=0}^{I}$, we have to show that there exists an increasing sequence $\{x_i\}_{i=I+1}^{\infty}$ that satisfies condition (A.6) and converges to a finite limit ξ. We have already seen that the LHS of the condition is positive when $x_{i+1} > x_i > x_0$, implying that the RHS is also positive and so $x_{i+2} > x_{i+1}$. By induction, any sequence that satisfies condition (A.6) will be increasing if $x_{I+1} > x_I$.

We can write condition (A.6) as the two-variable, first-order system:

$$\begin{aligned} x_{i+1} &= x_i y_i^{-1} \\ \delta\left[1-y_{i+1}^{\gamma+\frac{1}{2}}\right] &= y_i^{-1}\left[1-y_i^{\gamma+\frac{1}{2}}\right]-\frac{r+\lambda}{\lambda}\left\{y_i^{-1}\left[1-y_i^{\gamma+1}\right]-x_i y_i^{-1}\left[1-y_i^{\gamma}\right]\right\} \end{aligned}$$

which has a steady state at $(x, y) = (\xi, 1)$ for any ξ. The linear approximation to this system in a neighbourhood of $(\xi, 1)$ is

$$\begin{pmatrix} x_{i+1}-\xi \\ y_{i+1}-1 \end{pmatrix} = \begin{pmatrix} 1 & -\xi \\ 0 & \delta^{-1}\frac{r}{\lambda}\left(\xi-\frac{r+\lambda}{r}\right) \end{pmatrix}\begin{pmatrix} x_i-\xi \\ y_i-1 \end{pmatrix}.$$

The point $(\xi, 1)$ is a non-hyperbolic steady state (Devaney 1987), since one of the eigenvalues of this linear approximation is unity. We show that there exists a sequence $\{(x_i, y_i)\}$ converging to $(\xi, 1)$ provided that the other eigenvalue of this system is strictly between 0 and 1. This is equivalent to the condition:

$$\text{(A.7)} \qquad \frac{r+\lambda}{r} < \xi < \frac{r+(1+\delta)\lambda}{r}\,.$$

The lower bound is simply x_0, and is satisfied because x_i is increasing. The upper bound is equal to $x(p_2^*) - (1-\delta)\,\lambda/r$.

If condition (A.7) holds, then the Mean Value Theorem and the above linear approximation to y_{i+1} imply that there exists $\epsilon > 0$ and $0 < \alpha < 1$ such that $(1-y_{i+1}) < \alpha(1-y_i)$ for all (x_i, y_i) for which $\|(x_i, y_i)-(\xi, 1)\| < \epsilon$. The sequence $\{y_i\}$ is converging to unity as long as (x_i, y_i) is in this neighbourhood, but it may be that at some point the x_i coordinate becomes further than ϵ away from ξ and so the sequence leaves the neighbourhood. We will use induction to show that this cannot happen. Choose (x_i, y_i) so that $\|(x_i, y_i)-(\xi, 1)\| < C\epsilon(1-\alpha)$ and assume that

[11] In general, it will be strict; however, it will be weak in the knife-edge case where $x_{I+1} = \beta^{-1/\gamma}x_I$.

for $i = i+1, \ldots, j-1$ the sequence (x_i, y_i) does not leave this neighbourhood; we will then show that (x_j, y_j) also lies in the neighbourhood.

$$\begin{aligned}
\ln\left(\frac{x_j}{x_i}\right) \;=\; \ln\left(\prod_{m=1}^{j-i-1}\frac{1}{y_{i+m}}\right) \;&=\; -\sum_{m=1}^{j-i-1}\ln y_{i+m} \\
&<\; -\sum_{m=1}^{j-i-1}\ln[1-\alpha(1-y_{i+m-1})] \\
&<\; \sum_{m=1}^{j-i-1}\alpha(1-y_{i+m-1}) \\
&<\; -\alpha\sum_{m=1}^{\infty}\alpha^{m-1}(1-y_i) = \frac{\alpha}{1-\alpha}(1-y_i) < C\epsilon
\end{aligned}$$

(The first line applies the definition of y_i; the second line uses the fact that $(1-y_{i+1}) < \alpha(1-y_i)$ in the neighbourhood; the third line uses $\ln(1+x) \leq x$; the final line applies $(1-y_{i+m}) < \alpha^m(1-y_i)$ and then uses the initial assertion.) If $C < 1/\xi$, the inequality $\ln(x_j/x_i) < C\epsilon$ implies that $x_j < \xi + \epsilon$ for small ϵ. Thus y_i converges to unity, x_i is increasing and thus also converges to ξ. By working backwards from a neighbourhood of the steady state to the initial condition $x_{I+1} > \beta^{-I/\gamma}x_0$, we have shown that there is an increasing solution to condition (A.6) such that $\lim_{i\to\infty} x_i = \xi$ for any ξ satisfying condition (A.7). ■

Proof(s) from Section 4

Proof of Proposition 4.1: We again focus on the case where the belief has reached p_1^* and consider Steps $[i]$ $(i = 0, 1, \ldots)$.

For the case of private information (unobservable outcomes), we retain the definition of $x(p)$ and will establish a (different) increasing convergent sequence $\{x_i\}_{i=0}^{\infty}$. As before $x_0 = \Omega(p_1^*)/\Omega(p^m) = (r+\lambda)/\lambda$, $y_i = x_i/x_{i+1}$, and the proof has a similar structure.

Let the common belief be $\hat{p}_i$. Assume that player 2 (he) will play R for the next a_i periods, followed by S for the subsequent b_i periods if no lump-sum arrives. (If the other player has done likewise, he will continue; but if the other player has deviated, he will stop and play S forever.) The first step is to calculate the payoff E_i of player 1 (she) when she uses the same strategy, and then her payoff D_i when she deviates by playing S for the first a_i periods.

When both players play R for the first a_i periods and then neither continues to play R (thereby signalling no success), the common belief will become $\hat{p}_{i+1}$. The expected payoff from playing this strategy, $E_i(\hat{p}_i)$, is the same as that calculated in Proposition 3.1:

$$\frac{E_i(\hat{p}_i) - s}{s(1-\hat{p}_i)} = \frac{1}{x_i}\left[1 - y_i^{\gamma+1}\right] - [1 - y_i^{\gamma}] + \delta\frac{u_{i+1}(\hat{p}_{i+1}) - s}{s(1-\hat{p}_{i+1})}\,. \tag{A.8}$$

This is because the players are in effect playing the same strategy as before: if either has a success, they keep on playing R; if neither has a success, they take a break by playing S. In the private information case where one of them has a success, the other one takes an instantaneous break but reverts immediately *and was playing the appropriate arm in the periods preceding the prescribed break.*[12]

[12] A more formal argument is as follows. When player 1 plays R, her belief at the end of a_i periods will be $p_{i+1/2} = \hat{p}_i e^{-\lambda a_i}/(1-\hat{p}_i+\hat{p}_i e^{-\lambda a_i})$, and her payoff during this time is given by (2), that is

$$E_i(p) = \lambda h p + C_i(1-p)\Omega(p)^{2\gamma}, \quad p_{i+1/2} \leq p \leq \hat{p}_i\,,$$

If player 1 deviates by playing S for the first a_i periods and if no lump-sum arrives (for player 2), then player 2 will stop, both players will play S forever and get a payoff of s; however, if a lump-sum arrives (for player 2), then player 2 will continue to play R (thereby signalling success), both players will play R forever and get a payoff of λh. Player 1's expected payoff from this deviation, $D_i(\hat{p}_i)$, is

$$D_i(\hat{p}_i) = \int_0^{a_i} r\, e^{-rt}\, s\, dt + e^{-r\, a_i}\, [(1-\pi_i)s + \pi_i \lambda h]$$

or equivalently $(D_i(\hat{p}_i) - s)/s = e^{-r\, a_i}\, \pi_i\, \Omega(p^m)$, where π_i is the subjective probability that player 1 attaches to a success for player 2 having occurred during a_i periods *given her current (unchanged) belief* $\hat{p}_i$. Now $\pi_i = \hat{p}_i(1 - e^{-\lambda\, a_i})$ and using the fact that $y_i = e^{-2\lambda\, a_i}$ leads to the payoff from the deviation

$$\frac{D_i(\hat{p}_i) - s}{s(1-\hat{p}_i)} = \frac{1}{x_i}\, y_i^{\gamma} \left[1 - y_i^{\frac{1}{2}}\right]. \tag{A.9}$$

(Note that this is less than equation (A.4), so the payoff to deviation with private information is less than the payoff to deviation with public information.)

For deviation not to be optimal, $D_i(\hat{p}_i) \leq E_i(\hat{p}_i)$ for all i. From equations (A.9) and (A.8), the general condition is

$$\frac{1}{x_i}\, y_i^{\gamma} \left[1 - y_i^{\frac{1}{2}}\right] - \frac{1}{x_i} \left[1 - y_i^{\gamma+1}\right] + \left[1 - y_i^{\gamma}\right] \leq \delta\, U_{i+1}(\hat{p}_{i+1}) \tag{A.10}$$

for all i, where the term on the RHS denotes the final term in equation (A.8).

As before, the sequence $x_i = x_0$ (so $y_i = 1$) for all i satisfies condition (A.10), but we shall generate an *increasing* sequence such that x_i (and thus a_i) is as large as possible: for low values of i condition (A.10) is a strict inequality (when a_i will be at its maximum), and for i larger than some critical value J, condition (A.10) holds with equality (when a_i will be less than its maximum).

The LHS of condition (A.10) is 0 when $x_{i+1} = x_i$, and its first derivative with respect to x_{i+1} has the same sign as $-r(x_{i+1}/x_i) + (r+\lambda)(x_{i+1}/x_i)^{\frac{1}{2}} - (r+2\lambda) + r x_{i+1}$. When $x_{i+1} = x_i$, this is greater than or equal to 0 iff $r x_i \geq r + \lambda$, and the second derivative with respect to x_{i+1} is strictly positive. AS before, since $r x_0 = r + \lambda$, it follows by induction that the LHS of condition (A.10) is increasing in x_{i+1}, which enables us to define x_{i+1} recursively with $x_{i+1} = \beta^{-1/\gamma} x_i$ (corresponding to $a_i = \bar{a}$) if this does not violate condition (A.10), or else $x_{i+1} < \beta^{-1/\gamma} x_i$ (corresponding to $a_i < \bar{a}$) so that condition (A.10) holds with equality.

---

or equivalently

$$\frac{E_i(p) - s}{s(1-p)} = \frac{1}{x(p)} \left[1 - \left(\frac{x(p)}{x_{i+1/2}}\right)^{2\gamma+1}\right] - \left[1 - \left(\frac{x(p)}{x_{i+1/2}}\right)^{2\gamma}\right] + \left(\frac{x(p)}{x_{i+1/2}}\right)^{2\gamma} \frac{E_i(p_{i+1/2}) - s}{s(1-p_{i+1/2})},$$

where $x_{i+1/2} = x(p_{i+1/2})$. At the break time, the common belief either jumps to 1 with a continuation value of λh, or is revised down to $\hat{p}_{i+1}$ with a continuation value rising to $u_{i+1}(\hat{p}_{i+1})$ by the end of the break, where

$$E_i(\hat{p}_{i+1/2}) = \pi_{i+1/2} \lambda h + (1 - \pi_{i+1/2}) \left[\int_0^{b_i} r\, e^{-rt}\, s\, dt + e^{-r\, b_i}\, u_{i+1}(\hat{p}_{i+1})\right]$$

or equivalently $E_i(\hat{p}_{i+1/2}) - s = \pi_{i+1/2}[\lambda h - s] + (1 - \pi_{i+1/2}) e^{-r\, b_i}(u_{i+1}(\hat{p}_{i+1}) - s)$, where $\pi_{i+1/2}$ is the subjective probability that player 1 attaches to a success for player 2 having occurred during a_i periods *given her current belief* $\hat{p}_{i+1/2}$. Now $\pi_{i+1/2} = p_{i+1/2}(1 - e^{-\lambda\, a_i})$ and using the fact that $y_i = e^{-2\lambda\, a_i}$ and $x_i / x(p_{i+1/2}) = y_i^{1/2}$ leads to the payoff given in equation (A.8).

Define J as the smallest integer such that it holds with equality. For $i < J$ deviation is strictly suboptimal, whereas for $i \geq J$ she is indifferent and so we can use the payoff to a deviation ($D_{i+1}(\hat{p}_{i+1})$ – see equation (A.9)) to replace the continuation payoff (implicitly $u_{i+1}(\hat{p}_{i+1})$) in condition (A.10):

$$\frac{1}{x_i}\, y_i^{\gamma}\left[1-y_i^{\frac{1}{2}}\right]-\frac{1}{x_i}\left[1-y_i^{\gamma+1}\right]+\left[1-y_i^{\gamma}\right]=\delta\,\frac{1}{x_{i+1}}\, y_{i+1}^{\gamma}\left[1-y_{i+1}^{\frac{1}{2}}\right] \tag{A.11}$$

for $i \geq J$.

Consider the equality when $i = J$. As the LHS is increasing in x_{i+1}, if we replace x_{J+1} by $\beta^{-1/\gamma}x_J$ on the LHS and at its first occurrence on the RHS we have the following inequality:[13]

$$\frac{1}{x_J}\,\beta\left[1-\beta^{1/2\gamma}\right]-\frac{1}{x_J}\left[1-\beta^{1+1/\gamma}\right]+\left[1-\beta\right]\geq\delta\,\frac{\beta^{1/\gamma}}{x_J}\, y_{J+1}^{\gamma}\left[1-y_{J+1}^{\frac{1}{2}}\right],$$

leading to a lower bound for x_J when δ and β are close to 1 (i.e. when Δ and $\bar{a}$ are close to 0):

$$x_J \geq x(p_2^*)-(r\Delta+2\lambda\bar{a})\,\frac{\lambda}{r}\,.$$

This bound is the same as the bound in Proposition 3.1 for the public information case.

Having constructed the increasing sequence $\{x_i\}_{i=0}^{J}$, as before we have to show that there exists an increasing sequence $\{x_i\}_{i=J+1}^{\infty}$ that satisfies condition (A.11) and converges to a finite limit ξ. By induction, as before, any sequence that satisfies condition (A.11) will be increasing if $x_{J+1} > x_J$.

We can write condition (A.11) as the two-variable, first-order system:

$$\begin{aligned} x_{i+1} &= x_i y_i^{-1} \\ \delta y_{i+1}^{\gamma}\left[1-y_{i+1}^{\frac{1}{2}}\right] &= y_i^{-1}y_i^{\gamma}\left[1-y_i^{\frac{1}{2}}\right]-\left\{y_i^{-1}\left[1-y_i^{\gamma+1}\right]-x_i y_i^{-1}\left[1-y_i^{\gamma}\right]\right\} \end{aligned}$$

which has a steady state at $(x, y) = (\xi, 1)$ for any ξ. The linear approximation to this system in a neighbourhood of $(\xi, 1)$ is

$$\begin{pmatrix} x_{i+1}-\xi \\ y_{i+1}-1 \end{pmatrix} = \begin{pmatrix} 1 & -\xi \\ 0 & \delta^{-1}\,\frac{r}{\lambda}\left(\xi-\frac{r+\lambda}{r}\right) \end{pmatrix}\begin{pmatrix} x_i-\xi \\ y_i-1 \end{pmatrix}.$$

Since this linearisation is exactly the same as for the public information case, the remainder of the proof concerning the sequence $\{x_i\}_{i=0}^{\infty}$ replicates that of Proposition 3.1. ■

[13] Again, it will be strict except in the knife-edge case where $x_{J+1} = \beta^{-1/\gamma}x_J$.

References

Admati, A.R. and M. Perry (1991): "Joint Projects without Commitment", *Review of Economic Studies*, **58**, 259–276.

Berry, D.A. and B. Fristedt (1985): *Bandit Problems* (New York: Chapman and Hall).

Bolton, P. and C. Harris (1993): "Strategic Experimentation" (STICERD Discussion Paper No. TE/93/261, London School of Economics).

Bolton, P. and C. Harris (1999): "Strategic Experimentation", *Econometrica*, **67**, 349–374.

Devaney, R.L. (1987): *An Introduction to Chaotic Dynamical Systems* (Redwood City, California: Addison Wesley).

Lockwood, B. and J.P. Thomas (1999): "Gradual Cooperation in Repeated Games with Reversibilities" (mimeo).

Marx, L. and S. Matthews (1997): "Dynamic Contribution to a Public Project" (mimeo).

Rothschild, M. (1974): "A Two-Armed Bandit Theory of Market Pricing", *Journal of Economic Theory*, **9**, 185–202.