Strategic Experimentation with Private Payoffs*

Paul Heidhues† Sven Rady‡ Philipp Strack§

May 20, 2015

Abstract

We consider a game of strategic experimentation in which players face identical discrete-time bandit problems with a safe and a risky arm. In any period, the risky arm yields either a success or a failure, and the first success reveals the risky arm to dominate the safe one. When payoffs are public information, the ensuing free-rider problem is so severe that equilibrium experimentation ceases at the same threshold belief at which a single agent would stop, even if players can coordinate their actions through mediated communication. When payoffs are private information and the success probability on the risky arm is not too high, however, the socially optimal symmetric experimentation profile can be supported as a sequential equilibrium for sufficiently optimistic prior beliefs, even if players can only communicate via binary cheap-talk messages.

JEL classification: C73, D83.

Keywords: Strategic Experimentation, Two-Armed Bandit, Bayesian Learning, Information Externality, Mediated Communication, Cheap Talk.

*Our thanks for helpful comments and suggestions are owed to Simon Board, Patrick Bolton, Susanne Goldlücke, Johannes Hörner, Navin Kartik, Nicolas Klein, Moritz Meyer-ter-Vehn, Johannes Münster, Nicolas Vieille, an associate editor, two anonymous referees, and to seminar participants at UCLA, UC Berkeley, Yale, Frankfurt, the 2011 North American Winter Meeting of the Econometric Society, the 14th SFB/TR 15 Conference (Mannheim, April 2012), and the Workshop on Advances in Experimentation (Paris, December 2013). We thank the Economics Department at UC Berkeley and the Cowles Foundation for Research in Economics at Yale University for their hospitality. Financial support from the Deutsche Forschungsgemeinschaft through SFB/TR 15 is gratefully acknowledged.

†Lufthansa Chair in Competition and Regulation, ESMT European School of Management and Technology, Schlossplatz 1, D-10178 Berlin, Germany.

‡Corresponding author. *E-mail address:* rady@hcm.uni-bonn.de.

Department of Economics and Hausdorff Center for Mathematics, University of Bonn, Adenauerallee 24-42, D-53113 Bonn, Germany.

§Department of Economics, University of California at Berkeley, 513 Evans Hall, Berkeley, CA 94720, USA.

1 Introduction

In many real-life situations, economic agents face a trade-off between exploring new options and exploiting their knowledge about which option is likely to be best. A stylized model allowing one to analyze such experimentation problems is the two-armed bandit setting in which a gambler repeatedly decides which of two different slot machines to play with the ultimate goal of maximizing his monetary reward. Consequently, starting with Rothschild (1974), variants of this bandit problem have been studied in a wide variety of economic set-ups; see Bergemann and Välimäki (2008) for an overview, and the references in Section 2 for specific applications.

In this paper, we analyze strategic experimentation problems in which players can learn not only from their own past experiences but also from those of others. In contrast to the previous literature on multi-agent bandit problems, we allow players to communicate with each other. We consider both the case in which they observe each other's actions and payoffs and the case in which they only observe each other's actions. To fix ideas, think of experimental consumption. A consumer can learn by trying different products herself, of course. But she can also learn from observing others' consumption choices and from communicating with others about their experiences. Typically, however, it will be impossible for her to directly observe other agents' payoffs. In contrast, farmers experimenting with various crops may be able to observe not only the crop planted by their neighbors but also whether the crop thrives or not.

Section 3 introduces our strategic experimentation environment with N identical bandit machines, each of them consisting of a safe and a risky arm. The payoff distribution of the risky arm is the same for all machines, and can be either "good" or "bad". In each period, a good arm produces either a "success" or a "failure", while a bad arm always produces failures. Studying the scenario where all these machines are operated by one and the same agent yields the efficient benchmark; the special case where $N = 1$ leads to the autarky solution.

Section 4 turns to the case in which each machine is controlled by a different player, and players' actions *and* payoffs are publicly observable, so that all players always share a common belief about the state of the world. Focusing on continuous-time set-ups, Bolton and Harris (1999, 2000), Keller, Rady and Cripps (2005) and Keller and Rady (2010) show that if the players condition their actions on their common belief only, i.e. if they use Markov strategies, it is impossible to achieve the social optimum. Furthermore, if a single success on the risky arm fully reveals the good state of the world (and players are not allowed to switch actions infinitely often in finite time), then in any Markov perfect equilibrium players stop experimenting once the common belief reaches the single-agent cut-off (Keller, Rady and Cripps 2005).

Maintaining the assumption of fully revealing successes, we work in a discrete-time set-up and introduce a mediator in the spirit of Forges (1986) and Myerson (1986)—thereby allowing for general forms of communication between players. This enables us to considerably strengthen the existing results by showing that in *any*

communication equilibrium, players stop experimenting once the common belief falls beneath the single-agent cut-off. It is worth emphasizing that this impossibility result holds even though our equilibrium concept (Nash equilibrium in the game induced by the mediator's strategy) is permissive in two ways: it allows the players to coordinate through mediation, and it abstains from imposing sequential rationality.

There is a simple and compelling logic behind this finding. Once players are sufficiently pessimistic, it is optimal for them to refrain from experimenting. There is thus some cut-off belief below which players never use the risky arm. A player who is just above this cut-off belief and the last one to experiment then faces the single-agent trade-off; and if a player knows that her fellow players experiment with positive probability, her benefit from doing so herself is even lower. Thus, there is no experimentation below the single-agent cut-off belief.

We then, in Section 5, consider the case in which players observe each other's behavior but not the realized payoffs; they will in general not hold a common belief about the state of the world, therefore. Little is known about the general properties of such stochastic games. In our setting, we could attempt to adapt the arguments in Fudenberg and Levine (1983) and prove existence of equilibrium by truncating the game at finite horizons, invoking existence of sequential equilibrium for finite extensive games, and letting the length of the game go to infinity. As this would reveal little about the structure of equilibria and the set of equilibrium payoffs, however, we use a constructive approach instead.

While even the most sophisticated forms of communication fail to extend the range of equilibrium experimentation beyond the single-agent cut-off when payoffs are public, we show that in the private-payoffs environment the simplest possible form of communication allows the players to reach more efficient equilibria. In fact, we limit the players' ability to communicate by only allowing them to publicly send binary cheap-talk messages. Despite this restriction, we can not only establish the existence of sequential equilibria but also provide conditions under which players fare better in the private-information case then they do in the case of public information with mediation.

More specifically, we first show that for every equilibrium of the experimentation game with public payoffs and no communication, there exists an equilibrium of the game with private payoffs and binary cheap talk that yields the same distribution over experimentation paths, that is, sequences of experimentation choices and results. The key intuition lies in the fact that truthful communication is easy to sustain in our environment in which a single success fully reveals the state of the world. Following a success, a player is certain about the underlying state of the world and is willing to truthfully communicate her success. Furthermore, if a player believes that others are communicating truthfully, she believes that the state of the world is good with probability 1 upon hearing that a fellow player had a success. In this case, she is willing to play the risky arm forever without communicating her future payoffs. But then no player will want to wrongly announce a success because this will make it impossible to learn anything from one's fellow players in the future. Truthful revelation of payoffs, therefore, is always incentive compatible and

any equilibrium outcome for public payoffs can be replicated when payoffs are only privately observable. Comparatively unattractive equilibrium payoffs of the game with public experimentation outcomes can thus be used to penalize deviations in the game with private information.

In a second step, we exploit this insight to prove that the socially optimal symmetric pure-strategy profile can be supported as a sequential equilibrium if the players' initial belief is sufficiently optimistic.[1] To implement this strategy profile, we let all players experiment for as many rounds τ as the social planner would, and thereafter communicate whether they had a success or not. Unilateral deviations from the risky to the safe arm during the first τ rounds are punished by immediate truthful communication of past successes and transition to a continuation equilibrium which induces the same outcome as the symmetric Markov perfect equilibrium of the game with observable payoffs; for sufficiently small success probabilities on a good risky arm, this outcome is indeed worse for each player than the efficient symmetric pure action plan.[2] Intuitively, we make use of the fact that a player observing only his own experiments learns at a slower rate than a social planner who observes all players' experiments. As players learn only from their own experiments prior to round τ, therefore, they still find it myopically optimal to experiment if initially they were sufficiently optimistic, and thus have strictly no incentive to deviate.

Section 5 concludes and in particular discusses the role that fully revealing payoffs, observable actions, and communication play for our results, respectively.

The Appendix contains all proofs and shows how to embed our discrete-time model in a version of the continuous-time framework of Keller, Rady and Cripps (2005). The discrete-time model corresponds to imposing an equally-spaced grid of times at which the players can choose actions and exchange messages.

2 Related Literature

To the best of our knowledge, we are the first to introduce (mediated) communication into games of strategic experimentation. A similar approach has been taken in collusion models with imperfect private monitoring or persistent private information; see e.g. Compte (1998), Kandori and Matsushima (1998), and Athey and Bagwell (2008). Like in our scenario with private payoffs, introducing cheap-talk communication in these models facilitates the construction of equilibria and has the potential to improve equilibrium payoffs.

Sugaya and Wolitzky (2014) show that players in a repeated game can be better off with unmediated imperfect monitoring than with mediated perfect monitoring. Like us, therefore, they find that perfect information combined with mediated communication need not generate an upper bound on the equilibrium payoff set. The

[1] Focussing on the socially optimal *symmetric* profile entails only a minor efficiency loss. In fact, we show that in the "unrestricted" social optimum there is at most one period in which the number of experiments is strictly between 0 and N.

[2] This is not trivial because the symmetric Markov perfect equilibrium may involve randomization.

analogy, however, only holds at this very abstract level. Ours is not a repeated game, of course, and private information in our game concerns payoffs, not actions. So different strategic forces are at play in the two settings.

Our work on experimentation with private payoffs is most closely related to Rosenberg, Solan and Vieille (2007). We follow these authors in studying a discrete-time experimentation game with bandits whose risky arm can be of two possible types, and with players who observe each other's actions but not each other's payoffs. There are, however, two important differences: first, Rosenberg et al. assume the decision to stop experimenting to be irreversible, whereas we allow the players to freely switch from one arm to the other and back; second, we permit communication between the players. With irreversible stopping decisions, players cannot free-ride on an opponent's experimentation efforts, so if we assumed irreversible decisions in our framework, the socially optimal symmetric pure-strategy profile could easily be supported as an equilibrium with truthful communication.[3] The ability to switch actions freely thus enriches the players' strategic possibilities considerably, and makes it more difficult to achieve efficiency.

Like in Keller, Rady and Cripps (2005), one type of risky "project" in our model generates a perfectly informative signal at a constant and positive hazard rate while the other type never does. With public information, this signal structure can also be found in the models of R&D competition proposed by Malueg and Tsutsui (1997) and Besanko and Wu (2013), and in the bandit-based experimentation models of Klein and Rady (2011), Thomas (2014) and Forand (2015).[4]

Owing to its tractability, a number of authors have used this signal structure in models of learning and experimentation under private information. Bergemann and Hege (1998, 2005) and Hörner and Samuelson (2013) study the financing of a venture project in a dynamic agency model where the allocation of funds and the learning process are subject to moral hazard; Halac, Kartik and Liu (2013) add adverse selection with respect to the agent's ability. Décamps and Mariotti (2004) analyze a duopoly model of irreversible investment with a learning externality and privately observed investment costs. Acemoglu, Bimpikis and Ozdaglar (2011) investigate the effect of patents on firms' irreversible decisions to experiment themselves or copy a successful rival; they allow for private information about a project's success probability, but assume observable actions and outcomes.[5] In a stopping game with

[3] In fact, a player who is meant to experiment under the efficient strategy profile faces the same trade-off as the social planner: if she experiments, she bears the current cost of one experiment but learns the result of N experiments, while if she refrains from experimenting she obtains the outside payoff. Crucially, the players have no incentive to misrepresent their experimentation results when implementing this optimum: above the social planner's cut-off belief, all players experiment anyhow and benefit from the experimentation of others; below the cut-off, each player wants to cease experimentation even if the others experiment, and therefore has no incentive to misrepresent the fact that she is below the cut-off.

[4] Hazard rate uncertainty with imperfectly informative signals appears in Choi (1991), Keller and Rady (2010, 2015) and Hörner, Klein and Rady (2014).

[5] In their analysis of R&D races as preemption games with private information, Hopenhayn and Squintani (2011) instead let players' private information states increase stochastically over time according to a compound Poisson process.

public actions and private payoffs, Murto and Välimäki (2011) examine information aggregation through observational learning by a large number of players. Allowing for reversible experimentation choices, Bonatti and Hörner (2011) study a team problem in which the players' actions are private information and experimentation outcomes are public, which is precisely the opposite of what we assume here.

The signal structure that we assume in this paper also appears in recent research on optimal disclosure policies in the context of multi-agent Bayesian learning. Bimpikis and Drakopoulos (2014) show that in the set-up of Keller, Rady and Cripps (2005) there exists a time such that if agents commit to sharing no information before that time and disclosing all available information at that time, the incentives to free-ride are overcome and efficiency is restored. Analyzing the design of award structures and disclosure policies in innovation contests, Halac, Kartik and Liu (2015) and Bimpikis, Ehsani and Mostagir (2015) show that the designer may wish to withhold information about past successes from the agents and only release it after a certain amount of time has elapsed. On the path of play, the communication equilibria that we construct in the case of private payoffs also induce the revelation of past signals after a fixed number of periods, and for exactly the same reason: the incentive to free-ride on the information generated by others is lower when this information arrives with a delay.[6] There is no commitment to an information transmission strategy in our model, however, and the players only share past signals when it is sequentially rational for them to do so.[7]

3 A Two-Armed Bandit Model

There is an infinite number of periods $t = 0, 1, \ldots$ and in each period there is a choice between a safe and a risky action (or "arm"). More precisely, in period t, the decision maker chooses an action $k(t) \in \{R, S\}$. If $k(t) = S$, the agent receives a safe payoff normalized to 0; if $k(t) = R$, she receives a risky payoff $X_i(t)$ that is either low (X_L) or high (X_H), where $X_L < 0 < X_H$.

The distribution of the risky payoff depends on an unknown state of the world, which is either good ($\theta = 1$) or bad ($\theta = 0$). Conditional on this state, payoffs are drawn independently across periods. In the good state of the world, the probability of the high payoff is $\mathbb{P}(X_H|\theta = 1) = \pi \in]0, 1[$; in the bad state, it is $\mathbb{P}(X_H|\theta = 0) = 0$. Thus, a single draw of X_H proves that the state of the world is good. This makes our model a discrete-time analog of the model analyzed in Keller, Rady and Cripps (2005).[8] We write E_θ for the conditional expectation $\mathbb{E}[X_i(t)|\theta]$ of the risky

[6] In Wuggenig (2014), this delay is caused by a network structure imposed on the set of experimenting players.

[7] This also differentiates our paper from contributions outside the literature on strategic learning such as Kamien, Tauman and Zamir (1990). Investigating the value of information in finite extensive games, these authors ask how the set of equilibria is affected when they introduce an outsider who at the beginning of the game can commit to a mapping from his private information to messages that the players receive at any of their information sets.

[8] See the Appendix for details.

payoff in any given period, and assume that $E_0 < 0 < E_1$. We say that the agent *experiments* if she chooses the risky action while still being uncertain about the true state of the world.

Given a probability $p(0) = p$ which the agent initially assigns to the good state of the world, her objective is to choose a contingent plan of actions $(k(t))_{t=0}^{\infty}$ that maximizes the expected payoff

$$(1-\delta)\,\mathbb{E}_p\left[\sum_{t=0}^{\infty}\delta^t \mathbf{1}_{\{k(t)=R\}}X_i(t)\right],$$

where the factor $1-\delta$ serves to express the overall payoff in per-period units.

By the dynamic programming principle, it is without loss of performance to restrict the agent to stationary pure Markovian action plans, meaning that $k(t)$ is a time-invariant deterministic function of the posterior belief $p(t)$ which the agent holds at the start of round t. If there were $n \geq 1$ experiments so far and they all failed, then Bayes' rule determines $p(t)$ as

$$B(n,p) = \frac{p(1-\pi)^n}{p(1-\pi)^n + 1 - p};$$

if at least one of these experiments was successful, then $p(t) = 1$; and if there were no experiments so far, then $p(t) = p$, the initial belief.

For future reference, we define the *myopic* cut-off belief

$$p^m = \frac{|E_0|}{|E_0| + E_1}.$$

This is the belief at which the expected current payoff from the risky option just equals the safe payoff, so that a myopic agent would find it optimal to play risky when $p(t) \geq p^m$, and safe when $p(t) < p^m$.

A standard argument establishes that for a forward-looking agent, the optimal cut-off belief is

$$p^a = \frac{(1-\delta)|E_0|}{(1-\delta)(E_1 + |E_0|) + \delta\pi E_1},$$

where the superscript "a" indicates the *single-agent* or *autarky* solution. It is easy to see that $p^a < p^m$, reflecting the fact that a forward-looking decision maker values the information that her actions generate, and hence conducts more experiments than a myopic agent would.

More generally, consider an agent who operates $N \geq 2$ bandit machines. Let the state of the world θ be the same for all of them, and assume that conditional on $\theta = 1$, risky payoffs are drawn independently across bandits and periods. A (potentially random) action plan now takes values in $\{R,S\}^N$, and the payoff to be maximized is

$$(1-\delta)\,\mathbb{E}_p\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{\infty}\delta^t \mathbf{1}_{\{k_i(t)=R\}}X_i(t)\right],$$

where the bandit machines are labelled $i = 1, \ldots, N$ and $k_i(t)$ denotes the action taken on bandit i at time t. If this agent is restricted to plans that treat all machines symmetrically and do not involve any mixing, then she optimally uses all risky arms above the cut-off

$$p^{se} = \frac{(1-\delta)|E_0|}{(1-\delta)(E_1 + |E_0|) + \delta[1-(1-\pi)^N]E_1} < p^a,$$

and all safe arms below it. Here, the superscript "se" indicates the *efficient symmetric pure action plan*. If the agent is not restricted to symmetric plans, experimentation will optimally cease once $p(t)$ falls below the *efficient* cut-off

$$p^e = \frac{(1-\delta)|E_0|}{(1-\delta)(E_1 + |E_0|) + \delta N \pi E_1} < p^{se}.$$

While p^{se} is the belief at which the agent is indifferent between conducting N experiments in one round and stopping all experimentation for good, p^e is the belief at which she is just willing to conduct one last experiment. We show in the appendix that the efficient action plan involves at most one period in which the number of experiments lies strictly between 0 and N.

The single-agent, symmetric efficient, and efficient solutions provide useful benchmarks for settings in which $N \geq 2$ *different* players operate the bandit machines. We turn to such settings now.

4 Strategic Experimentation with Public Payoffs and Mediated Communication

Suppose that each of $N \geq 2$ players operates a bandit machine as in Section 3 with a common unknown state of the world and conditionally independent payoff processes. Assume that the players' actions and payoffs are public information and that players can coordinate their behavior through mediated communication. More precisely, they can send private messages to a mediator at the start of each round; having received all these messages, the mediator in turn sends a private message to each player, who then chooses an action. These messages can be random and, while players must mix independently of each other, the mediator's messages can be correlated across players.

Formally, let I_i be the set of "input" messages that player $i = 1, \ldots, N$ can send to the mediator, M_i the set of messages that the mediator can send to player i, and $O_i = \{(R, X_L), (R, X_H), (S, 0)\}$ the set of possible combinations of player i's actions and payoffs in any period. In addition, let $I = I_1 \times \ldots \times I_N$, $M = M_1 \times \ldots \times M_N$, and $O = O_1 \times \ldots \times O_N$. A strategy for the mediator is a mapping μ which assigns a probability distribution on M to each history in $\bigcup_{t=0}^{\infty} (I^t \times M^{t-1} \times O^{t-1})$.[9] A

[9]Note that we allow our mediator to observe public information—i.e. past actions—directly, while usually a mediator is assumed to only observe what players report. This strengthens our impossibility result regarding experimentation below the single-agent cutoff.

strategy for player i is a mapping σ_i which assigns a probability distribution on I_i to each history in $\bigcup_{t=0}^{\infty}\left(I_i^{t-1}\times M_i^{t-1}\times O^{t-1}\right)$, and a probability distribution on $\{R,S\}$ to each history in $\bigcup_{t=0}^{\infty}\left(I_i^{t}\times M_i^{t}\times O^{t-1}\right)$. We write σ for a profile of such strategies.

A pair (μ,σ) is a *communication equilibrium* if σ is a Nash equilibrium of the game with payoff functions

$$u_i(\sigma|p,\mu)=(1-\delta)\,\mathbb{E}_{p,\mu,\sigma}\left[\sum_{t=0}^{\infty}\delta^t\mathbf{1}_{\{k_i(t)=R\}}X_i(t)\right],$$

where $k_i(t)$ denotes the action chosen by player i at time t and the expectation is taken with respect to the distribution over player i's actions and payoffs induced by the prior belief p, the mediator's strategy μ, and the players' strategies σ.

Since any profile of payoffs $u_i(\sigma|p,\mu)$ can also be achieved by an agent who operates N bandit machines simultaneously (and is not restricted to symmetric action plans), the maximal average payoff $v^e(p)$ per bandit machine that this agent can attain constitutes an upper bound on the average payoff $\frac{1}{N}\sum_{i=1}^{N}u_i(\sigma|p,\mu)$ in any communication equilibrium. And since each player always has the option to ignore the information contained in the opponent's actions and payoffs and in the mediator's messages, the maximum payoff $v^a(p)$ of an agent operating a single bandit machine in autarky constitutes a lower bound on each player's individual equilibrium payoff. As $v^a(p)=v^e(p)=0$ for $p\leq p^e$, we can conclude that starting from such a prior, any communication equilibrium gives each player a payoff of zero. What is more, given that the agent operating N machines *strictly* prefers not to conduct any experiments below p^e, there can be no communication equilibrium in which an agent plays risky with positive probability once the posterior belief has fallen below p^e.

To get a first intuition for why equilibrium experimentation cannot even go beyond the single-agent cut-off p^a, consider pure-strategy communication equilibria. Since players do not experiment below p^e, there are only finitely many periods in which such an equilibrium can require players to experiment in the absence of a prior success. Consider the last period in which a player is meant to experiment. If this player is the only one required to use the risky arm, she knows that in the event of a failure, no player will experiment in future. Hence, she will only be willing to experiment if this is individually optimal, that is, if the current belief is at least p^a. When other players are also meant to experiment in this last round, the belief must be above p^a because the value of experimenting in this round is lower if a fellow player also experiments. Hence, in any pure-strategy equilibrium all experimentation must stop once the posterior beliefs falls below the single-agent cut-off. Conversely, it cannot be the case that all players permanently stop experimenting at a belief above p^a. The reason is simply that each player—believing that the other players stopped experimenting—would then face the single-agent trade-off.

The following proposition exploits this logic and extends it to arbitrary communication equilibria.

Proposition 1. *In any communication equilibrium of the experimentation game with public payoffs and mediated communication, there will almost surely be another*

experiment on the path of play when $p(t) > p^a$, and no further experiment when $p(t) < p^a$.

In a continuous-time set-up without communication, Keller, Rady and Cripps (2005) establish that with fully revealing successes on the risky arm, any Markov perfect equilibrium in which players change actions only finitely often has the properties stated in Proposition 1. In our discrete-time set-up, we can drop the Markov assumption and show that their result extends to all Nash equilibria of the game with public payoffs. Moreover, this finding is robust to the introduction of mediated communication.[10]

While Proposition 1 shows that communication does not help in terms of the total number of experiments carried out in equilibrium, one may ask whether it might increase efficiency by reducing the time in which these experiments are performed. In general, this is not the case. Hörner, Klein and Rady (2014) show that, as the period length goes to zero, the following outcome can be supported as an equilibrium of the game with observable actions and payoffs and no communication: all players use the risky arm until the common posterior belief falls below p^a and from then on use the safe arm only. As this outcome maximizes the players' aggregate payoffs subject to the constraint that there cannot be any experimentation below p^a, introducing communication thus neither decreases delay nor increases overall payoffs in the game with observable realizations on the risky arm.[11]

5 Strategic Experimentation with Private Payoffs and Direct Communication

We now change the set-up in two ways. First, while actions remain publicly observable, we assume from now on that the realizations of the risky payoffs $X_i(t)$ are private information. Second, we restrict the players' communication to direct exchange of simple cheap-talk messages after actions have been chosen and payoffs are realized, taking the message space to be as small as possible. This restriction strengthens the *possibility* result which we are going to prove in this section; by contrast, the *impossibility* result in Proposition 1 was strengthened by our allowing the most powerful communication protocol and using the most permissive equilibrium concept. Following this logic, we shall apply a more restrictive solution concept in this section, and construct equilibria that satisfy sequential rationality as well as consistency of off-path beliefs in the spirit of Kreps and Wilson (1982).

[10] Keller, Rady and Cripps (2005) also construct equilibria which feature some experimentation below the single-agent cut-off; these equilibria rely on players switching actions an infinite number of times during a finite time interval. Proposition 1 confirms that they cannot be obtained as the limit of equilibria of the discretized experimentation game; see Hörner, Klein and Rady (2014) for a related discussion.

[11] To be precise, the equilibrium construction in Hörner et al. relies on a public randomization device. So the mediator in the form considered here (observing public payoffs) helps in terms of delay and aggregate payoffs to the extent that it serves to replicate the public randomization device that we did not add to our model as a primitive element.

Let each player's message space be $\{0,1\}$, set $M = \{0,1\}^N$ and recall that $O_i = \{(R, X_L), (R, X_H), (S, 0)\}$ is the set of possible combinations of player i's actions and payoffs in any period. Now, a strategy for player i is a mapping σ_i which assigns a probability distribution on $\{R, S\}$ to each history in $\bigcup_{t=0}^{\infty} \left(O_i^{t-1} \times \{R,S\}^{t-1} \times M^{t-1}\right)$, and a probability distribution on $\{0,1\}$ to each history in $\bigcup_{t=0}^{\infty} \left(O_i^t \times \{R,S\}^t \times M^{t-1}\right)$. We write σ for a profile of such strategies.

Our solution concept is *sequential equilibrium*. Such an equilibrium consists of a strategy profile and a belief system such that each player acts optimally after every history given her beliefs and the other players' strategies (sequential rationality), and beliefs are consistent in the sense of Kreps and Wilson (1982): there is a sequence of completely mixed strategy profiles converging to the equilibrium profile such that the associated belief system (which is determined by Bayes' rule everywhere) converges to the beliefs that support the equilibrium. The infinite-horizon setting requires us to specify a notion of convergence here; we use pointwise convergence, that is, convergence for each history.

Recalling that O denotes the Cartesian product of the sets O_i, $i = 1, \ldots, N$, we call O^t the set of possible *experimentation paths* of length t. We call two strategy profiles *path-equivalent* if they induce the same distribution over the set of all experimentation paths, $\bigcup_{t=1}^{\infty} O^t$. Note that all path-equivalent strategy profiles give rise to the same payoff profile. What is more, this equivalence relation also applies in situations where one strategy profile is from the game with private payoffs and direct communication, and the other from the game with public payoffs and mediated communication that we analyzed before.

We begin by arguing that truthful communication can be made incentive compatible when payoffs are privately observed and fully revealing. Suppose that, upon a first success, players send the message 1. After the first occurrence of an action-message pair $(k_i(t), m_i(t)) = (R, 1)$, let all players babble forever, that is, send each of the messages 0 and 1 with probability $\frac{1}{2}$ in any round. As long as player i has no success and none of the other players has chosen the action-message pair $(k_j(t), m_j(t)) = (R, 1)$, let player i send the message 0 after a failed experiment and babble if she used the safe arm. Intuitively, then, if all players communicate this way, a first success is publicly observable, and—because play after a first success is obvious—we are back in the case of public payoffs *without* any communication.

The key observation now is that if players anticipate this communication strategy, truthful communication is indeed incentive compatible. Following a first success on player i's risky arm, player i knows the state of the world and hence is indifferent as to what other players believe. So truthfully announcing a success is optimal for player i following a success of her own. After such an announcement, player j believes with certainty that the state of the world is good, and hence will play risky in all future periods irrespectively of what player i does after the announcement. If player i incorrectly announces a success, she cannot infer anything from player j's future behavior, so she is at last weakly better off telling the truth.[12]

[12]Since player i cannot learn from players $-i$ after announcing a success she did not have, she solves the autarky problem given her beliefs about the state of the world.

The corresponding belief system looks as follows. As long as no success is announced, all players use Bayes' rule to update their beliefs based on the presumption that all experiments carried out so far were failures. After the first time that some player i announces a success, players $-i$ assign probability 1 to the good state of the world; if no other player announces a success at the same time, i updates her beliefs using Bayes' rule. Once a player is subjectively certain of the good state, she never revises this belief again.

This specification of communication strategies and beliefs yields the following result.

Proposition 2. *For every sequential equilibrium of the experimentation game with public payoffs and no communication, there exists a path-equivalent sequential equilibrium of the game with private payoffs and binary cheap-talk messages.*

With public payoffs and no communication, consistency of beliefs simply requires the application of Bayes' rule after all histories. With private payoffs and binary cheap talk, the off-equilibrium beliefs specified above are such that a player who observes a deviation from the risky to the safe action and is not yet certain of the good state of the world believes that the deviating player had only failures before the deviation. This is in line with the fact that the safe action is strictly dominated for any player who already had a success. A player who was made subjectively certain of the good state by an opponent's announcement of a success, however, maintains this belief even in the face of the opponent's subsequent use of the safe arm. As shown in the Appendix, these beliefs are indeed consistent in the sense of Kreps and Wilson (1982). The key is to make "wrong" actions so much more likely than "wrong" messages that for any given off-path history, it becomes infinitely more likely in the limit that the history was caused by mistakes in actions alone.

Combining Propositions 1 and 2, we can construct equilibria of the game with private payoffs and direct communication that have the potential to deter deviations from a more efficient path of play. By Proposition 1, we can restrict attention to finitely many beliefs when we consider equilibria of the game with public payoffs and no communication. When we take the set of these beliefs as the state space (and think of the lowest of these beliefs as an absorbing state), there are only finitely many pure Markov perfect strategies. So we can consider a finite auxiliary game in which each player's action set is the set of pure Markov perfect strategies and payoffs are defined as in the original game. This auxiliary game is symmetric and hence possesses a symmetric (possibly mixed-strategy) Nash equilibrium which corresponds to a symmetric Markov perfect equilibrium of the experimentation game with public payoffs and no communication. By Proposition 2, the game with private payoffs and direct communication has a path-equivalent sequential equilibrium.

Using this equilibrium to punish early deviations from the risky to the safe arm, we can build an equilibrium that implements the efficient symmetric pure action plan for sufficiently optimistic initial beliefs. To see the intuition, recall that at the symmetric optimum, and conditional on all experiments failing, an agent who operates N bandit machines simultaneously updates her belief on the basis of N

experiments in every round until the belief falls below p^{se}. When payoffs are private and messages are uninformative, by contrast, each player updates her belief using only the result of her own experimentation, and so the belief decreases at a slower rate. The key observation is that for high enough priors, this slower learning implies that the symmetric optimum reaches p^{se} before players who do not exchange any information reach the myopic cut-off p^m.

Above p^m, players myopically prefer to experiment, and in the equilibrium that we are about to construct, they experiment and babble as long as the symmetric optimum prescribes full experimentation when all experiments fail. At the end of that time span, players communicate truthfully and continue playing risky if and only if at least one player announces a prior success. During that time span, a player who deviates from the risky to the safe arm reduces her expected current payoff and triggers a transition to a continuation equilibrium that is path-equivalent to a symmetric Markov perfect equilibrium of the game with public payoffs and no communication.

To see which level of initial optimism suffices, let $\tau^{se}(p)$ denote the number of rounds in which, given the prior p, the symmetric optimum prescribes full experimentation when all experiments fail. If $B(\tau^{se}(p), p) > p^m$, then the belief updated on the basis of N failures per round reaches p^{se} before the belief updated on the basis of just one failure per round reaches p^m. The requirement that $B(\tau^{se}(p), p) > p^m$ is equivalent to the inequality

$$\ln \frac{p}{1-p} + \tau^{se}(p) \ln(1-\pi) > \ln \frac{|E_0|}{E_1}.$$

It is straightforward to verify that

$$\tau^{se}(p) \leq \frac{1}{N \ln(1-\pi)} \left(\ln \frac{1-\delta}{1-\delta(1-\pi)^N} + \ln \frac{1-p}{p} + \ln \frac{|E_0|}{E_1} \right) + 1;$$

as $\ln(1-\pi) < 0$, the desired inequality thus is easily seen to hold for all p greater than

$$\bar{p} = \frac{\left(1-\delta(1-\pi)^N\right)^{\frac{1}{N-1}} |E_0|}{\left(1-\delta(1-\pi)^N\right)^{\frac{1}{N-1}} |E_0| + (1-\delta)^{\frac{1}{N-1}} (1-\pi)^{\frac{N}{N-1}} E_1}.$$

By construction, $\bar{p} > p^m$.

Proposition 3. *Let $(1-\pi)^{N-1} \geq 1/N$. Then, for all initial beliefs $p > \bar{p}$, the experimentation game with private payoffs and binary cheap-talk messages has a sequential equilibrium that is path-equivalent to the efficient symmetric pure action plan.*

The condition on the number of players N and the success probability π ensures that the players' common payoff function in a symmetric Markov perfect equilibrium of the experimentation game with public payoffs and no communication (which may involve randomization) is bounded above by the value function associated with

the efficient symmetric pure action plan.[13] Under this condition, deviations before $\tau^{se}(p)$ can be punished through a continuation equilibrium that is path-equivalent to a symmetric Markov perfect equilibrium. Of course, there may be harsher punishments for a deviating player. For example, we could search for an equilibrium of the game with public payoffs and no communication that minimizes the payoff of a given player. By playing a continuation equilibrium that is path-equivalent to this asymmetric equilibrium, we would punish early deviations more severely. Thus, the interval $]\bar{p}, 1]$ constitutes just part of the set of initial beliefs for which the outcome of the efficient symmetric pure action plan can be sustained in an equilibrium.

For a range of initial beliefs below $\bar{p}$, the same logic as in Proposition 3 allows us to construct equilibria in which the players may not achieve the symmetric optimum but do perform more experiments, and hence are better off, than in any equilibrium with public payoffs. Suppose that $B(N, p^a) > p^{se}$ and let

$$\tau^a(p) = \min\{\tau \in \mathbb{N} : B(N\tau, p) < p^a\}.$$

Then, starting from a prior p such that $B(\tau^a(p) + 1, p) > p^m$ and defining strategies as in the proof of Proposition 3 but with τ^{se} replaced by $\tau^a + 1$, we obtain an equilibrium of the game with private payoffs in which the players experiment some way below p^a. In view of the similarity with the above construction, we omit the details.

So far, we have focused on inducing socially desirable experimentation. One can also exploit the above logic to induce over-experimentation in equilibrium. To construct a simple example, observe first that for any prior belief $p \in]p^m, 1[$, there exists a number $\tilde{N}(p)$ such that for any $N > \tilde{N}(p)$ the optimal symmetric pure-strategy profile involves at most one round of experimentation and $B(N-1, p) < p^a$. This follows from the fact that the optimal symmetric cut-off belief p^{se} is bounded away from zero as $N \to \infty$, while $B(N-1, p)$ goes to zero as $N \to \infty$. Now choose a p very close to 1 and a number of players $N > \tilde{N}(p)$, and let players on the path of play experiment in the first two rounds, communicate truthfully at the start of the third round, and babble in all other rounds. If any player deviates in period 1 and does not experiment, let all players communicate immediately and choose the safe arm in all future periods whenever there was no success; as the public belief is below the autarky cut-off if all first-period experiments failed, this is indeed sequentially rational. If the first unilateral deviation occurs in round 2, let players communicate and continue as on the path of play—that is, play risky if and only if there has been a prior success (experienced on one's own risky arm or announced by a fellow player). If all players experimented in period 1, it is clearly optimal for each of them to experiment in period 2: it is myopically optimal and the player will be better informed than when he refrains from experimenting. It is also optimal to experiment in period 1: at a prior belief p sufficiently close to 1, a player is not willing

[13] Heidhues, Rady and Strack (2012) show that the condition can be dispensed with when $N = 2$. For arbitrary N, the condition holds if the model is viewed as a discretization of the set-up in Keller, Rady and Cripps (2005), and players interact sufficiently often per unit of time; see the Appendix for details.

to forgo the myopic benefit of using the risky arm even if doing so would provide him with perfect information after the first period.[14] Note that this equilibrium involves massive over-experimentation conditional on all experiments failing—in fact, twice as much experimentation as in the optimal symmetric pure-strategy profile.

Given that equilibrium can involve over-experimentation, one may also wonder whether moving from public to private breakthroughs can decrease the lowest equilibrium payoff. Hörner, Klein and Rady (2014) show that, as the period length goes to zero, there exist symmetric equilibria of the game with observable actions and payoffs which take each player's expected payoff arbitrarily close to the single-agent value, and hence to the infimum payoff over all equilibria of the experimentation game, be it with public or private breakthroughs. Thus, at least in the frequent-action limit, private information together with simple direct communication unambiguously improves the set of symmetric equilibrium payoffs.

6 Conclusion

We analyzed a discrete-time experimentation game with two-armed bandits. For publicly observable payoffs, the free-rider problem is so severe that in any communication equilibrium, no player experiments below the single-agent cut-off belief. Privately observed payoffs mitigate the free-rider problem to the point where for sufficiently optimistic prior beliefs, it becomes possible to sustain the socially optimal symmetric pure-strategy profile as an equilibrium with simple cheap-talk communication.

Our results clearly hinge on the assumption of fully revealing successes. We actually exploit this signal structure in two ways: first, after observing a success, a player is willing to communicate truthfully; second, after a player has announced a success, there is no interest in further communication. It would be interesting to investigate the impact of communication in games of strategic experimentation with imperfectly informative signals and to check to what extent our findings carry over to a set-up in which a bad risky arm generates successes at a very low, but positive rate. As such an investigation raises entirely new challenges, however, we leave it to future work.

Throughout, we assumed that players cannot prove the results of their own experiments. If we suppose instead that a player can provide hard evidence of any prior success, our equilibrium construction in Proposition 3 still carries through. Intuitively, whenever players communicate, they do so truthfully and, hence, it is unnecessary to show a proof. Moreover, a player who has had a success is indifferent as to the other players' behavior and therefore willing not to reveal hard information.

It is easy to see that if actions and payoffs are both unobservable, we can still support the efficient symmetric pure action plan as a sequential equilibrium for sufficiently optimistic prior beliefs. We merely need to modify the construction

[14] Because play following simultaneous deviations is irrelevant for the incentives to deviate unilaterally, we do not discuss such histories; in our environment with communication, it is straightforward to prescribe sequentially rational behavior following such histories.

underlying Proposition 3 to take account of the fact that other players' deviations from the supposed experimentation path are unobservable, so off-path play must be specified after own deviations only. Following a deviation from the risky to the safe arm during the first τ rounds (the number required in the symmetric optimum), we can simply let the player in question experiment up to and including period $\tau - 1$ and then—exactly as on the equilibrium path—announce at the beginning of period τ whether or not she had a success. We let players babble in all other rounds and follow their autarky strategy from round τ on. It is then straightforward to verify that this constitutes an equilibrium for initial beliefs as in Proposition 3.

Although we believe that in most strategic experimentation problems the presumption that players can communicate is realistic, one may wonder exactly what role communication plays for our results. The answer is somewhat subtle. In the equilibrium of Proposition 3, players only communicate at a single point in time on the path of play. That is, after a given number of rounds of experimentation—say 100—players announce truthfully whether or not they had a success. Intuitively, one could replace this communication by one round of play in period 101 in which each player uses the risky arm if and only if she had a prior success, ensuring that all necessary information is exchanged within one round. The role of communication may thus seem very limited. Truthful communication, however, plays another important technical role: it ensures the existence of a simple continuation equilibrium following a deviation—including one in the first period. What kind of equilibria exist absent communication remains, in our view, an interesting question for further research.

A natural question is what payoffs are achievable in the private-information case if beliefs are not "sufficiently optimistic". In Heidhues, Rady and Strack (2012) we provide an answer for the continuous-time limit of our discrete-time model, showing that for *any* given prior, a sufficiently small period length will ensure that the (symmetric) optimum can be achieved in an equilibrium with private payoffs and cheap-talk communication, in marked contrast to the case of public payoffs. For this result, however, we rely on a construction of beliefs that—although consistent in the spirit of sequential equilibrium—we do not find compelling. To see the logic behind this construction, suppose that the optimal symmetric action plan requires $\tau > 1$ periods of experimentation. Player i may be tempted to pause prior to period τ in order to free-ride on her fellow players' costly experimentation. Following such a deviation to the safe arm by player i, let the other players believe that player i had a prior success, and then use the risky arm forever without communicating.[15] This implies that player i cannot learn anything from her fellow players' subsequent actions, and thereby deters free-riding. As these beliefs seem peculiar, we focus on a different equilibrium construction in this paper, but the result nevertheless establishes that absent a stronger belief refinement the social optimum can always be achieved in the continuous-time limit.

[15]This "punishment via beliefs" construction is similar in spirit to the one sustaining collusion in Blume and Heidhues (2006).

Appendix

Discretizing the model of Keller, Rady and Cripps (2005)

We can embed our discrete-time model in a continuous-time framework that coincides (up to a normalization of the safe payoff to zero) with the set-up in Keller, Rady and Cripps (2005).

Let time be continuous and suppose that operating the risky arm comes at a flow cost of $s > 0$ per unit of time. In the good state ($\theta = 1$), the risky arm yields lump-sum payoffs which arrive at the jump times of a Poisson process with intensity $\lambda > 0$. These lump-sums are independent draws from a time-invariant distribution with known mean $h > 0$, and the Poisson processes in question are independent across players. In the bad state ($\theta = 0$), the risky arm never generates a lump-sum payoff. The safe arm does not produce any such payoffs either, but is free to use.

Given the common discount rate $r > 0$, a player's payoff increment from using a bad risky arm for a length of time $\Delta > 0$ is

$$\int_0^\Delta r\, e^{-r\,t}\,(-s)\,dt = \left(1 - e^{-r\,\Delta}\right)(-s).$$

The expected discounted payoff increment from a good risky arm is

$$\mathbb{E}\left[\int_0^\Delta r\, e^{-r\,t}\,(h\,dN_t - s\,dt)\right] = \int_0^\Delta r\, e^{-r\,t}\,(\lambda h - s)\,dt = \left(1 - e^{-r\,\Delta}\right)(\lambda h - s);$$

here N_t is a standard Poisson process with intensity λ, and the first equality follows from the fact that $N_t - \lambda t$ is a martingale. We assume $\lambda h > s$ so that a good risky arm dominates the safe arm. Finally, the probability of observing at least one lump-sum on a good risky arm during a time interval of length Δ is $1 - e^{-\lambda\Delta}$.

If we let the players adjust their actions only at the times $t = 0, \Delta, 2\Delta, \ldots$ for some fixed $\Delta > 0$, we are back in our discrete-time framework with $\pi = 1 - e^{-\lambda\Delta}$, $E_0 = -s$, $E_1 = \lambda h - s$, and $\delta = e^{-r\,\Delta}$.

Letting $\Delta \to 0$, we can study the impact of vanishing time lags. The cut-offs p^{se} and p^e from Section 3 converge in a monotonically decreasing fashion to one and the same limit as Δ vanishes; this limit is the efficient N-player cut-off from Keller, Rady and Cripps (2005),

$$p_N^* = \frac{r|E_0|}{r(E_1 + |E_0|) + N\lambda E_1}\,.$$

Thus, the difference between the efficient symmetric strategy profile and its unrestricted counterpart completely disappears in the limit, and implementing the symmetric optimum as we do in Proposition 3 fully solves the free-rider problem. The belief $\bar{p}$ in that proposition converges to a belief strictly above p^m which can be determined in closed form.

Efficiency implies at most one period of asymmetric experimentation

Given the belief p, an agent who operates $N \geq 2$ bandit machines is indifferent between conducting K experiments in one round and stopping all experimentation for good if and only if

$$(1-\delta)\tfrac{K}{N}E_p + \delta p\left[1-(1-\pi)^K\right]E_1 = 0,$$

where $E_p = pE_1 + (1-p)E_0$ is the expected current payoff from using the risky arm. This equation is solved by the cut-off belief

$$p^{(K)} = \frac{(1-\delta)|E_0|}{(1-\delta)(E_1+|E_0|) + \delta\frac{N}{K}\left[1-(1-\pi)^K\right]E_1}.$$

As $[1-(1-\pi)^K]/K$ is decreasing in K,[16] we obtain the cut-offs

$$p^e = p^{(1)} < p^{(2)} < \ldots < p^{(N-1)} < p^{(N)} = p^{se}.$$

To prove that there is at most one round in which the optimal number of experiments is neither 0 nor N, it is enough to show that starting from the cut-off $p^{(K)}$, K failed experiments take the belief below p^e, i.e., $B(K, p^{(K)}) < p^e$. Straightforward computations show that this inequality is equivalent to

$$(1-\delta)(1-\pi)^{-K} + \delta\tfrac{N}{K}\left[(1-\pi)^{-K}-1\right] > 1-\delta+\delta N\pi;$$

as $(1-\pi)^{-K} \geq 1+K\pi$,[17] this is always satisfied.

Proof of Proposition 1

First, consider any on-path history $h_1^t \in I_1^{t-1} \times M_1^{t-1} \times O^{t-1}$ of length t for player 1 such that $p(t) = p$ with $p^a < p < 1$. Let $v^a(p) > 0$ be the value of the single-agent problem at the belief p. Let τ be the smallest integer such that $\delta^\tau E_1 < v^a(p)$, and ψ be the probability that player 1 assigns to the event that on the path of play there will be at least one experiment (conducted by any of the N players) in the periods $t, t-1, \ldots, t+\tau-1$. The period t continuation value of player 1 is then trivially bounded above by $(1-\psi)\delta^\tau E_1 + \psi E_1$. So ψ cannot be smaller than $\bar{\psi} = [v^a(p) - \delta^\tau E_1]/[(1-\delta^\tau)E_1]$, because otherwise it would be profitable for the player to deviate to the single-agent solution.

For $\ell = 1, 2, \ldots$ let A_ℓ denote the event that no player experiments in the $\ell\tau$ periods $t, t+1, \ldots, t+\ell\tau-1$ on the path of play. We have just shown that $\mathbb{P}[A_1|h_1^t] \leq 1-\bar{\psi}$ where the probability measure $\mathbb{P}$ is the one induced by the communication equilibrium. If $\mathbb{P}[A_1|h_1^t] = 0$, then $\mathbb{P}[A_2|h_1^t] = 0$ as well. If $\mathbb{P}[A_1|h_1^t] > 0$, then $\mathbb{P}[A_2|h_1^t] = \mathbb{P}[A_2|h_1^t \wedge A_1]\mathbb{P}[A_1|h_1^t]$ and, since $p(t+\tau) = p(t)$ in the event A_1,

[16] For $0 < a < 1$ and $x > 0$, the derivative of the function $x \mapsto (1-a^x)/x$ has the same sign as $(1+\alpha x)\,a^x - 1$ where $\alpha = |\ln a|$. This is negative because $a^x = e^{-\alpha x} = 1/e^{\alpha x} < 1/(1+\alpha x)$.

[17] For $x > -1$, we have $\ln(1+x) \leq x$, and so $-K\ln(1-\pi) \geq K\pi \geq \ln(1+K\pi)$. Applying the exponential function yields the desired inequality.

$\mathbb{P}[A_2|h_1^t \wedge A_1] \leq 1-\bar{\psi}$ by the same argument as in the previous paragraph. By induction, this proves that $\mathbb{P}[A_\ell|h_1^t] \leq (1-\bar{\psi})^\ell$ for all ℓ. Letting $\ell \to \infty$, we see that player 1 assigns probability $\mathbb{P}[B|h_1^t] = 1$ to the event B that there will be another experiment on the path of play starting in period t. The true probability of this event is $\mathbb{P}[B|h_1^t, h_2^t, \ldots, h_N^t]$; as $1 = \mathbb{P}[B|h_1^t] = \mathbb{E}\left[\mathbb{P}[B|h_1^t, h_2^t, \ldots, h_N^t]|\, h_1^t\right]$, we must have $\mathbb{P}[B|h_1^t, h_2^t, \ldots, h_N^t] = 1$ as well.

Next, consider an on-path history of length t for player 1 such that $p(t) = p < p^a$. Suppose that the equilibrium requires player 1 to experiment with positive probability at time t. Conditional on the good state of the world, moreover, let player 1 assign probability ϕ to the event that at least one other player has a success in round t. As player 1 cannot gain from switching to the strategy of playing safe now and, in case all other players are unsuccessful, continuing to play safe forever, we must have

$$\delta p\phi E_1 \quad \leq \quad (1-\delta)E_p + \delta\left\{p[\pi+\phi-\pi\phi]E_1 + (1-p[\pi+\phi-\pi\phi])\,v\right\},$$

where $E_p = pE_1 + (1-p)E_0$ is the expected current payoff from using the risky arm, $\pi+\phi-\pi\phi$ is the probability that player 1 assigns to at least one success occurring when he plays risky himself, and v is player 1's expected continuation value conditional on him playing risky and all experiments failing. As $0 \leq v \leq E_1$, this in turn requires that

$$0 \leq (1-\delta)E_p + \delta\left\{p\pi E_1 + (1-p\pi)v\right\}.$$

As $p < p^a$, we have $(1-\delta)E_p + \delta p\pi E_1 < 0$, and hence $v > 0$. So on the path of play, some player must experiment with positive probability in round $t+1$ or later. Repeating this step until a time $t+\tau$ at which $p(t+\tau) < p^e$ in the absence of a success, we obtain a contradiction with our earlier finding that the equilibrium probability of experimentation is zero below the efficient cut-off.

Consistency of beliefs in Proposition 2

We want to show that the beliefs supporting the equilibrium in Proposition 2 can be derived as the limit of beliefs induced by completely mixed strategies that converge to the equilibrium strategies. To do so, we assign "trembles" after every history in which players do not randomize under the equilibrium strategies. After any action history of this kind, let players choose the "opposite" action with probability $\epsilon > 0$. Given any message history after which the players are meant to communicate truthfully, let them send the "wrong" message with probability $\rho(\epsilon) = \exp(-\epsilon^{-1})$. Starting from the fact that $\lim_{x\to\infty} x^q e^{-x} = 0$ for all natural numbers q, it is straightforward to show that $\lim_{\epsilon\to 0} \rho(\epsilon)/P(\epsilon) = 0$ for any polynomial $P\colon \mathbb{R} \to \mathbb{R}$.

Consider a history $h_i(t)$ for player i such that (i) player i had no success, (ii) at least one other player announced a first success, and (iii) one such player subsequently played S at least once. From player i's point of view, the conditional probabilities of this history are $\mathbb{P}(h_i(t)|\theta = 0) = \rho(\epsilon)P_0(\epsilon)$ and $\mathbb{P}(h_i(t)|\theta = 1) = [1-\rho(\epsilon)]P_1(\epsilon)$ for some polynomials P_0 and P_1. Now, the fact that $\lim_{\epsilon\to 0} \rho(\epsilon)/P_1(\epsilon) = 0$ implies $\lim_{\epsilon\to 0} \mathbb{P}(\theta = 1|h_i(t))) = 1$.

Proof of Proposition 3

Recall that there exists a symmetric Markov perfect equilibrium in the case of public payoffs. Choose such an equilibrium for any starting belief p. By Proposition 2 there exists an equilibrium with private payoffs that is path-equivalent. Denote this equilibrium by $\sigma(p)$, and the players' common payoff in this equilibrium by $v_\sigma(p)$.

We are now ready to specify strategies.[18] Fix an initial belief $p > \bar{p}$, and let $\tau = \tau^{se}(p)$. For $n = 0, 1, \ldots$, let $p(n) = B(n, p)$. In all periods $t \leq \tau - 1$, all players babble and use the risky arm provided no player chose the safe arm in an earlier period $t' \leq t - 1$. If a single player is the first to deviate to the safe arm in a period $t' \leq \tau - 1$, the players truthfully communicate at the beginning of period $t' + 1$. Then, if player i communicated truthfully in period $t' + 1$, she subsequently plays the strategy prescribed by $\sigma(q)$, where $q = p(Nt' - 1)$ if no player announced a success, and $q = 1$ otherwise. If player i did not communicate truthfully in period $t' + 1$, she either had a success she did not announce or she incorrectly announced a success she did not have. In the former case, she uses the risky arm from period $t' + 1$ on; in the latter case, she plays the autarky strategy.

We are left to specify strategies in case all players used the risky arm in all periods $t \leq \tau - 1$. In this case, all players truthfully announce at the beginning of period τ whether they had a success in any of the previous rounds; and independent of how the play proceeds from period τ onwards, players babble in every period $t > \tau$. If a player $j \neq i$ announced a success, or if player i observed a success herself in any prior period, player i uses the risky arm in all periods $t \geq \tau$. If no player announced a success in period τ, player i uses the risky arm in period $t \geq \tau$ if and only if she had a prior success herself or some player $j \neq i$ used the risky arm in a period $t' \in \{\tau, \ldots, t - 1\}$ when no player did so in periods $t = \tau, \ldots, t' - 1$.

Next, we specify the players' beliefs about the state of the world.[19] Any player who had a prior success believes the state of the world to be good with probability 1. At the start of any period $t' \leq \tau - 1$ such that all players used the risky arm in all periods $t < t'$, the belief of player i is $p(t' - 1)$ if all her experiments failed. If a player is the first to deviate to the safe arm in a period $t' \leq \tau - 1$, she believes the state of the world to be good with probability 1 if some other player announces a success in period $t' + 1$ (or she had a prior success herself); otherwise her belief is $p(Nt' - 1)$. Thereafter, beliefs are as in the equilibrium $\sigma(p(Nt' - 1))$.

Now suppose that all players used the risky arm in all periods $t \leq \tau - 1$. Each player believes the others' messages in period τ to be truthful, and all subsequent messages to be uninformative. If player i deviates in period τ by incorrectly an-

[18] For brevity, we do not explicitly specify behavior following observable simultaneous deviations as this is irrelevant for the incentives to deviate unilaterally.

[19] We thus do not follow the usual practice of specifying beliefs about nodes in information sets (where a node can be identified through whether and when the other players had a success when experimenting in addition to one's own observations). Such beliefs can be constructed in the obvious way from the probability that the state of the world is good together with how often the other players used the risky arm; these beliefs are unique but for the fact that we can arbitrarily prescribe when another player had a success following an out-of-equilibrium observation.

nouncing a success she did not have, and no player $j \neq i$ announces a success in period τ, then player i's belief in period $t \geq \tau$ equals 1 if she experiences a success in one of the periods $\tau, \ldots, t-1$, and equals $p(N\tau+n)$ if she carries out n experiments in periods $\tau, \ldots, t-1$ and they all fail. If player $j \neq i$ is the first to deviate (by using the risky arm) in period $t' \geq \tau$ after no player announced a success, player i believes the state of the world to be good with probability 1. If player i is the first to deviate in period $t' \geq \tau$ after no player announced a success, player i's belief in round $t \geq t'+1$ equals 1 if she experiences a success in one of the periods $t', \ldots, t-1$, and equals $p(N\tau+n)$ if she carries out n experiments in periods $t', \ldots, t-1$ and they all fail.

To verify that our beliefs are consistent, we assign "trembles" after every history in which players do not randomize under the equilibrium strategies in exactly the same way as in the proof of Proposition 2. After any action history of this kind, let players choose the "opposite" action with probability $\epsilon > 0$. Given any message history after which the players are meant to communicate truthfully, let them send the "wrong" message with probability $\rho(\epsilon) = \exp(-\epsilon^{-1})$. Then the exact same argument as the one in the proof of Proposition 2 proves that in the limit as $\epsilon \to 0$, the beliefs converge to the ones specified above.

It remains to show sequential rationality. Each player uses the risky arm whenever she assigns probability 1 to the good state of the world, which is clearly optimal. If a single player is the first to deviate to the safe arm in a period $t' \leq \tau - 1$, it is optimal for all players to communicate truthfully at the beginning of period $t'+1$ by the argument underlying Proposition 2. If player i does communicate truthfully in period $t'+1$, she believes with probability 1 that players $-i$ play according to $\sigma(q)$ with q specified above. As $\sigma(q)$ constitutes an equilibrium, it is optimal for player i to play according to $\sigma(q)$ as well. If player i announces a success in period $t'+1$ that she did not have, players $-i$ use the risky arm forever and babble, so player i finds herself in an autarky situation and optimally plays the autarky strategy. If player i does not announce a success she had, it is clearly optimal for her to use the risky arm forever.

If player i is the first to deviate in a period $t' \leq \tau - 1$, she does so with the belief $p(t') > p^m$, and her expected total payoff from period t' on is

$$\delta \Big\{ p(t') \left[1 - (1-\pi)^{(N-1)(t'+1)} \right] E_1 + \left(1 - p(t') \left[1 - (1-\pi)^{(N-1)(t'+1)} \right] \right) v_\sigma(p(Nt' + N - 1)) \Big\},$$

where $p(t') \left[1 - (1-\pi)^{(N-1)(t'+1)} \right]$ is the probability of at least one player $j \neq i$ having a success in any of the periods $0, \ldots, t'$ (which is announced truthfully in period t'). We shall show below that for $(1-\pi)^{N-1} \geq 1/N$, the payoff function v_σ is bounded above by the payoff function v^{se} associated with the efficient symmetric pure action plan. So player i's expected continuation payoff from the deviation is

no larger than $\delta w_i'$ where

$$\begin{aligned} w_i' &= p(t')\left[1-(1-\pi)^{(N-1)(t'+1)}\right]E_1 \\ &\quad + \left(1-p(t')\left[1-(1-\pi)^{(N-1)(t'+1)}\right]\right)v^{se}(p(N(t'+1)-1)). \end{aligned}$$

If player i does not deviate, her expected total payoff from period t' on equals $(1-\delta)E_{p(t')}+\delta w_i$ where

$$\begin{aligned} w_i &= p(t')\left[1-(1-\pi)^{(N-1)(t'+1)+1}\right]E_1 \\ &\quad + \left(1-p(t')\left[1-(1-\pi)^{(N-1)(t'+1)+1}\right]\right)v^{se}(p(N(t'+1))). \end{aligned}$$

Being the upper envelope of linear functions, v^{se} is convex. This implies that

$$v^{se}(p(N(t'+1)-1)) \leq p(N(t'+1)-1)\pi E_1 + [1-p(N(t'+1)-1)\pi]\, v^{se}(p(N(t'+1))).$$

Using this inequality and the fact that

$$\begin{aligned} &\left(1-p(t')\left[1-(1-\pi)^{(N-1)(t'+1)}\right]\right)[1-p(N(t'+1)-1)\pi] \\ &= 1-p(t')\left[1-(1-\pi)^{(N-1)(t'+1)+1}\right], \end{aligned}$$

one finds $w_i' \leq w_i$. As $E_{p(t')} > 0$, deviating is suboptimal, therefore.

We are left to rule out deviations by player i in a period $t' \geq \tau$ following a history in which all players used the risky arm in all periods $t \leq \tau - 1$, player i had no success himself, no player announced a success, and no player $j \neq i$ was the first to deviate in any of the periods $\tau, \ldots, t'-1$. In this case, player i's belief is below p^{se}. If player i is the first to deviate in a period $t'' \in \{\tau, \ldots, t'-1\}$, players $-i$ use the risky arm in all subsequent periods independent of their experimentation results. The same holds if there was no deviation in any round $t \leq t'-1$ and player i deviates to the risky arm in round t'. If there was no prior deviation and player i uses the safe arm in round t', finally, she expects players $-i$ to use the safe arm in all future periods. Whatever she does in round t', therefore, player i cannot learn anything from the other players' future behavior and so finds herself in an autarky situation. As $p^{se} < p^a$, it is thus optimal for player i to use the safe arm.

Finally, we show that the stated condition on N and π ensures that $v_\sigma \leq v^{se}$. Consider an agent who operates N bandit machines simultaneously subject to the restriction that only safe arms must be used at beliefs below the autarky cut-off p^a; suppose that this agent is not required to treat the bandit machines symmetrically and can randomize independently across machines. As this agent could replicate the experimentation path of the equilibrium σ, her value function $\hat{v}$ (computed as the average payoff per bandit machine) satisfies $v_\sigma \leq \hat{v}$. In turn, one has $\hat{v} \leq v^{se}$ if it is optimal for this agent to use all N risky arms with probability 1 at all beliefs $p \geq p^a$. This is the case if the agent is willing to conduct N experiments at the

belief p^a, that is, if

$$\begin{aligned}&(1-\delta)E_{p^a}+\delta p^a\left[1-(1-\pi)^N\right]E_1\\ \geq\ &(1-\delta)\frac{N-1}{N}E_{p^a}+\delta p^a\left[1-(1-\pi)^{N-1}\right]E_1,\end{aligned}$$

which simplifies to

$$(1-\delta)\frac{1}{N}E_{p^a}+\delta p^a(1-\pi)^{N-1}\pi E_1\geq 0$$

and, since $(1-\delta)E_{p^a}=-\delta p^a\pi E_1$, to $(1-\pi)^{N-1}\geq 1/N$.

References

Acemoglu, D., K. Bimpikis and A. Ozdaglar (2011): "Experimentation, Patents, and Innovation," *American Economic Journal: Microeconomics*, 3, 37–77.

Athey, S. and K. Bagwell (2008): "Collusion with Persistent Cost Shocks," *Econometrica*, 76, 593–540.

Bergemann, D. and U. Hege (1998): "Venture Capital Financing, Moral Hazard and Learning," *Journal of Banking and Finance*, 22, 703–735.

Bergemann, D. and U. Hege (2005): "The Financing of Innovation: Learning and Stopping," *RAND Journal of Economics*, 36, 719–752.

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

Besanko, D. and J. Wu (2013): "The Impact of Market Structure and Learning on the Tradeoff between R&D Competition and Cooperation," *Journal of Industrial Economics*, 61, 166–201.

Bimpikis, K. and K. Drakopoulos (2014): "Disclosing Information in Strategic Experimentation," working paper, Stanford University and MIT.

Bimpikis, K., S. Ehsani and M. Mostagir (2015): "Designing Dynamic Contests," working paper, Stanford University and University of Michigan.

Blume, A. and P. Heidhues (2006): "Private Monitoring in Auctions," *Journal of Economic Theory*, 131, 179–211.

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

Bolton, P. and C. Harris (2000): "Strategic Experimentation: the Undiscounted Case," in: *Incentives, Organizations and Public Economics – Papers in Honour of Sir James Mirrlees*, ed. by P.J. Hammond and G.D. Myles. Oxford: Oxford University Press, 53–68.

Bonatti, A. and J. Hörner (2011): "Collaborating," *American Economic Review*, 101, 632–663.

Choi, J.P. (1991): "Dynamic R&D Competitionu Under "Hazard Rate" Uncertainty," *RAND Journal of Economics*, 22, 596–610.

Compte, O. (1998): "Communication in Repeated Games with Imperfect Private Monitoring," *Econometrica*, 66, 597–626.

Décamps, J.-P. and T. Mariotti (2004): "Investment Timing and Learning Externalities," *Journal of Economic Theory*, 118, 80–102.

Forand, G. (2015): "Keeping Your Options Open," *Journal of Economic Dynamics and Control*, 53, 47–68.

Forges, F. (1986): "An Approach to Communication Equilibria," *Econometrica*, 54, 1375–1385.

Fudenberg, D. and D. Levine (1983): "Subgame-Perfect Equilibria of Finite- and Infinite-Horizon Games," *Journal of Economic Theory*, 31, 251–268.

Halac, M., N. Kartik and Q. Liu (2013): "Optimal Contracts for Experimentation," working paper, Columbia University and University of Warwick.

Halac, M., N. Kartik and Q. Liu (2015): "Contests for Experimentation," working paper, Columbia University and University of Warwick.

Heidhues, P., S. Rady and P. Strack (2012): "Strategic Experimentation with Private Payoffs," SFB/TR 15 Discussion Paper No. 387, available at `http://www.sfbtr15.de/uploads/media/387.pdf`.

Hopenhayn, H.A. and F. Squintani (2011): "Preemption Games with Private Information," *Review of Economic Studies*, 78, 667–692.

Hörner, J. and L. Samuelson (2013): "Incentives for Experimenting Agents," *RAND Journal of Economics*, 44, 632-663.

Hörner, J., N. Klein and S. Rady (2014): "Strongly Symmetric Equilibria in Bandit Games," SFB/TR 15 Discussion Paper No. 469, available at `http://www.sfbtr15.de/uploads/media/469.pdf`.

Kamien M., Y. Tauman and S. Zamir (1990): "On the Value of Information in Strategic Conflict," *Games and Economic Behavior*, 2, 129–153.

Kandori, M. and H. Matsushima (1998): "Private Observation, Communication and Collusion," *Econometrica*, 66, 627–652.

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

Keller, G. and S. Rady (2015): "Breakdowns," *Theoretical Economics*, 10, 175–202.

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

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

Kreps, D.M. and R. Wilson (1982): "Sequential Equilibrium," *Econometrica*, 50, 863–894.

Malueg, D.A. and S.O. Tsutsui (1997): "Dynamic R&D Competition with Learning," *RAND Journal of Economics*, 28, 751–772.

Murto, P. and J. Välimäki (2011): "Learning and Information Aggregation in an Exit Game," *Review of Economic Studies*, 78, 1426–1461.

Myerson, R.B. (1986): "Multistage Games with Communication," *Econometrica*, 54, 323–358.

Rosenberg, D., E. Solan and N. Vieille (2007): "Social Learning in One-Armed Bandit Problems," *Econometrica*, 75, 1591–1611.

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

Sugaya, T. and A. Wolitzky (2014): "Perfect Versus Imperfect Monitoring in Repeated Games," working paper, Stanford University.

Thomas, C. (2014): "Strategic Experimentation with Congestion," working paper, University of Texas at Austin.

Wuggenig, M. (2014): "Learning Faster or More Precisely? Strategic Experimentation in Networks," SFB/TR 15 Discussion Paper No. 485, available at `http://www.sfbtr15.de/uploads/media/485.pdf`.