

# Strategic Experimentation with Private Payoffs\*

Paul Heidhues<sup>†</sup>    Sven Rady<sup>‡</sup>    Philipp Strack<sup>§</sup>

September 16, 2010

*Preliminary and incomplete*

We consider the strategic choices of two players facing identical experimentation problems. Each player observes the opponent's behavior but only her own realized payoffs. Players may, however, communicate their results via cheap talk and hence potentially have three sources of information: their own payoffs, the opponent's actions, and the cheap-talk messages. If payoffs were public information, the players would have an incentive to free-ride on each other's experimentation, and thus perform an inefficiently low number of experiments. When payoffs are private information, however, we show that it is possible to overcome the free-riding problem. We provide conditions under which the socially optimal symmetric experimentation profile can be supported as a perfect Bayesian equilibrium.

JEL classification:

Keywords: Strategic Experimentation, Private Monitoring, Cheap Talk, Information Externality, Bandit Problem.

---

\*Our thanks for helpful discussions and comments are owed to Nicolas Klein and seminar participants at UCLA, UC Berkeley and Yale. 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.

<sup>†</sup>Department of Economics, University of Bonn, Adenauerallee 24-42, D-53113 Bonn, Germany and CEPR.

<sup>‡</sup>Department of Economics, University of Munich, Kaulbachstr. 45, D-80539 Munich, Germany.

<sup>§</sup>Bonn Graduate School of Economics, Adenauerallee 24-26, D-53113 Bonn, Germany.

# 1 Introduction

In many real-life situations economic agents face a tradeoff between exploring new options and exploiting their knowledge about which option is likely to be best. A stylized model capturing this feature is two-arm bandit problem in which a gambler repeatedly decides which of two different slot machines to play with the ultimate goal of maximizing his monetary reward. The consecutive payoffs of the arms are identical and independently distributed random variables whose underlying distribution is unknown. When playing a given arm, the agent learns more about its underlying distribution—knowledge that is useful for future choices. Starting with Rothschild (1974), variants of the multi-arm bandit problem have been applied to a wide variety of economic settings (see Bergemann and Välimäki 2008 for a recent overview).

It is, however, also natural to presume that the agent can learn not only from her own exploration but also from the experiences of others in many economic settings. Experimental consumption is a case in point. As a stylized example, suppose there are two agents with common tastes. The agents choose between two items on a given menu and both know the quality of one regular item that has been on the menu for a long time. There is also a new item of unknown quality on the menu whose quality critically depends on how well it is cooked. The restaurant's chef is believed to be either good or very good. A very good chef is able to prepare the item nearly perfectly most of the time, while a good chef is only able to do so some of the time. Each agent can now learn through three channels: he may experiment himself and try the new item, he may learn from observing what the other agent chooses, and finally he may ask the other agent about whether the item was well-prepared whenever she tried it.

This paper deals with the problem of strategic experimentation under private information. We consider an extension of the two-armed bandit problem in which multiple agents face an identical experimentation problem. Agents observe each others' behavior but not the realized payoffs. Furthermore we allow agents to communicate by cheap talk. In such a context, agents have three sources of information: their own signals, their observation of other agents, and the cheap talk messages. Because generating information is costly, agents may try to excessively learn from their fellow agents' behavior rather than their costly own experimentation.

The scenario where every player can observe all other players' actions and signals (without any communication) is well studied in the literature (Bolton and Harris 1999, 2000, Keller, Rady and Cripps 2005, Keller and Rady 2010). As a consequence of payoffs being observable, all players share a common belief about the state of the world. The literature shows that, if the agents can only condition on this common belief, i.e. use Markov-perfect strategies, it is impossible to realize the socially optimal outcome.

[To be completed.]

## 2 The Model

There is an infinite number of periods  $t = 0, 1, \dots$  and there are two players who in each period choose between a safe and a risky action. Before making this choice, they can costlessly communicate with each other.

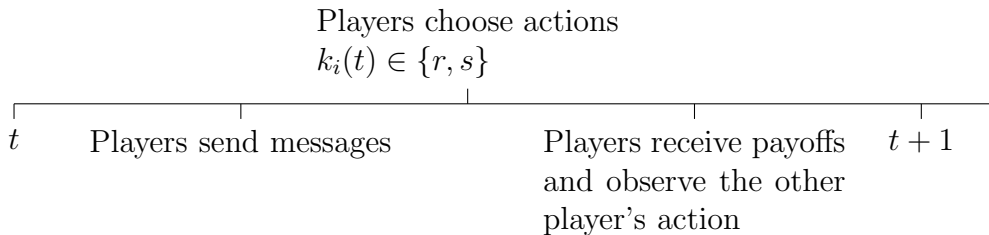


Figure 1: Timeline of the game.

More precisely, at the beginning of period  $t$ , each player  $i$  chooses a cheap-talk message  $m_i(t) \in [0, 1]$ . Upon having observed the other player's cheap-talk message, each player  $i$  then chooses an investment  $k_i(t) \in \{r, s\}$ . If  $k_i(t) = s$ , the player receives a safe payoff normalized to 0; if  $k_i(t) = r$ , the player receives a risky payoff  $X_i(t)$  which 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 the state of the world, payoffs are drawn independently across players and periods. In the good state of the world, the probability of the high payoff  $\mathbb{P}(X_H|\theta = 1) = p_1$ ; in the bad state, it is  $\mathbb{P}(X_H|\theta = 0) = p_0$ . We write  $E_\theta$  for the conditional expectation  $\mathbb{E}[X_i(t)|\theta]$  of the risky payoff in any given period, and assume that  $E_0 < 0 < E_1$ .

For analytical tractability, we focus on the special case of *fully revealing payoffs*  $p_1 > 0$  and  $p_0 = 0$  of this general setup. Our model is the discrete analog of the "good news" Poisson case analyzed in Keller, Rady and Cripps (2005). With fully revealing payoffs, a single successful experiment proves that the state of the world is good.

Our primary interest below is in analyzing the game in which the realizations of the payoffs  $X_i(t)$  are private information but the choice whether or not to experiment is observable. When considering this game, we partition the set of private histories for each player into those after which he has to send a message and those after which he has to choose an action. Formally, the set of *private*

message histories of player  $i$  at time  $t$  is

$$H_{i,t}^m = \left( \underbrace{[0, 1]^2}_{\text{messages}} \times \underbrace{\{X_L, 0, X_H\}}_{\text{own payoff}} \times \underbrace{\{r, s\}^2}_{\text{observed actions}} \right)^t.$$

Define the set of all private message histories  $H_i^m = \bigcup_{t=0}^{\infty} H_{i,t}^m$ . Similarly, the set of all private action histories at time  $t$  is defined by

$$H_{i,t}^a = \left( \underbrace{[0, 1]^2}_{\text{messages}} \right)^{t+1} \times \left( \underbrace{\{X_L, 0, X_H\}}_{\text{own payoff}} \times \underbrace{\{r, s\}^2}_{\text{observed actions}} \right)^t,$$

and the set of all private action histories by  $H_i^a = \bigcup_{t=0}^{\infty} H_{i,t}^a$ .

A *pure strategy* is thus a mapping that assigns to each private message history  $h_i^m \in H_i^m$  a message  $m_i(h_i^m) \in [0, 1]$  and to each private action history  $h_i^a \in H_i^a$  an action  $k_i(h_i^a) \in \{r, s\}$ . Mixed strategies are defined in the usual way.

Given a probability  $\pi_i(0) = \pi$  that player  $i$  assigns to the good state of the world, his expected payoff from a pure-strategy profile is

$$(1 - \delta) \mathbb{E}_{\pi} \left[ \sum_{t=0}^{\infty} \delta^t k_i(h(t)) X_i(t) \right],$$

where the factor  $1 - \delta$  serves to express the overall payoff in per-period units. Note that player  $j$ 's strategy only enters through the expectation operator—there is just an informational externality at play here.

We will assume that players start with a common prior  $\pi = \pi_1(0) = \pi_2(0)$  and solve for *perfect Bayesian equilibria* (PBE) of the game. We call an equilibrium perfect Bayesian if all players act optimally after every history given their beliefs and the other players' strategies, and if the players' beliefs are updated according to Bayes' rule whenever possible.

As a benchmark, we will also study the game with observable payoffs and no communication. Starting from a common prior, this is a stochastic game in which the posterior belief  $\pi(t)$  evolves according to Bayes' rule from one period to the next. For this game, we shall use the solution concept of *subgame perfect equilibrium* (SPE).

Whether payoffs are public information or not, we shall say that a player *experiments* if she chooses the risky action while still being uncertain about the true state of the world. For future reference, we define

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

This is the belief at which the expected current payoff from the risky option just equals zero, i.e. the safe payoff. A myopic player chooses the risky arm if and only if his posterior belief exceeds  $\pi^m$ . We therefore call  $\pi^m$  the myopic cutoff belief.

### 3 The Planner's Problem

In this section, we discuss the problem of a social planner who chooses a strategy profile to maximize the average of the two players' objective functions. The optimal outcome in this situation will serve as a benchmark for the case that individual players pursue their goals independently.

Let  $k = (k_1, k_2)$  denote a pure strategy profile. Then the expected average payoff, expressed in per-period units, is

$$U(\pi, k) = (1 - \delta) \mathbb{E}_\pi \left[ \frac{1}{2} \sum_{i=1}^2 \sum_{t=0}^{\infty} \delta^t k_i(h(t)) X_i(t) \right],$$

where  $\pi$  denotes the probability that the planner initially assigns to the good state.

For the social planner, it can never be strictly beneficial to have players hide information from each other because he can always choose a strategy profile that ignores unwanted information. Hence, when discussing the planner's problem, we will focus on strategy profiles in which all players truthfully communicate their past payoffs via their cheap-talk messages.

The planner's problem then becomes a Markovian decision problem with a single state variable, the posterior belief  $\pi(t)$ . We will now consider the two cases of fully revealing and equally revealing payoffs.

A single success in the past fully reveals that the state of the world is good ( $\theta = 1$ ). It is then a dominant choice for the planner to have both players take the risky action in all following periods. Using this fact and restricting attention to symmetric strategy profiles, we can think of the planner as choosing a period  $\tau$  after which all experimentation stops in case every prior experiment has been unsuccessful. For any such choice of  $\tau$ , expected average payoffs are

$$\begin{aligned} U(\pi, \tau) &= (1 - \delta) \left\{ (1 - \pi) \sum_{t=0}^{\tau-1} \delta^t E_0 + \pi \sum_{t=0}^{\tau-1} \delta^t E_1 + \pi [1 - (1 - p)^{2\tau}] \sum_{t=\tau}^{\infty} \delta^t E_1 \right\} \\ &= (1 - \delta^\tau) [\pi E_1 + (1 - \pi) E_0] + \delta^\tau \pi [1 - (1 - p)^{2\tau}] E_1 \\ &= E_\pi + \delta^\tau \{ (1 - \pi) |E_0| - (1 - p)^{2\tau} \pi E_1 \}, \end{aligned}$$

where we define  $E_\pi = \pi E_1 + (1 - \pi) E_0$  for the sake of brevity. The social planner prefers  $\tau + 1$  experiments over  $\tau$  experiments if and only if  $U(\pi, \tau + 1)$  is greater than  $U(\pi, \tau)$  which gives us

$$\begin{aligned} 0 &\leq U(\pi, \tau + 1) - U(\pi, \tau) \\ &= \delta^\tau \{ (\delta - 1)(1 - \pi) |E_0| - (1 - p)^{2\tau} \pi E_1 (\delta(1 - p)^2 - 1) \} \\ \Leftrightarrow 0 &\leq \frac{\delta - 1}{\delta(1 - p)^2 - 1} \frac{1 - \pi}{\pi} \frac{|E_0|}{E_1} - (1 - p)^{2\tau} \\ \Leftrightarrow \tau &\leq \frac{1}{2 \ln(1 - p)} \left[ \ln \frac{\delta - 1}{\delta(1 - p)^2 - 1} + \ln \frac{1 - \pi}{\pi} + \ln \frac{|E_0|}{E_1} \right]. \end{aligned}$$

Consequently the value function of the social planner is

$$v^{sc}(\pi) = (1 - \delta^{\tau^*})E_\pi + \pi(1 - (1 - p)^{2\tau^*})\delta^{\tau^*}E_1,$$

with

$$\tau^* = \left\lfloor \frac{1}{2 \ln(1 - p)} \left[ \ln \frac{\delta - 1}{\delta(1 - p)^2 - 1} + \ln \frac{1 - \pi}{\pi} + \ln \frac{|E_0|}{E_1} \right] \right\rfloor.$$

The planner prefers experimenting above the belief  $\pi^{sc}$  at which  $U(\pi^{sc}, 1) = 0$ , that is, above

$$\pi^{sc} = \frac{(1 - \delta)|E_0|}{(1 - \delta)(E_1 + |E_0|) + \delta(2 - p)pE_1}, \quad (1)$$

If the social planner can use asymmetric strategy profiles, it is optimal for him to experiment beyond the belief  $\pi^{sc}$ . In fact, the expected discounted payoff from letting one player experiment and stopping all experimentation thereafter,

$$(1 - \delta)^{\frac{1}{2}} [\pi E_1 + (1 - \pi)E_0] + \delta \pi p E_1,$$

is positive above the cutoff

$$\pi^c = \frac{(1 - \delta)|E_0|}{(1 - \delta)(E_1 + |E_0|) + \delta 2pE_1} < \pi^{sc}.$$

For the full description of the social optimum, we will also need the following cutoff:

$$\tilde{\pi}^c = \frac{(1 - \delta)|E_0|}{(1 - \delta)(E_1 + |E_0|) + \delta 2p(1 - p)E_1}.$$

It is straightforward to see that  $\tilde{\pi}^c > \pi^{sc}$  and that starting at  $\pi = \tilde{\pi}^c$ , one failed experiment takes the posterior belief below  $\pi^c$ .

**Proposition 1 (Socially optimal strategy profile).** *There exists a socially optimal strategy profile in which both players always communicate their payoffs truthfully, both choose the risky option at all beliefs  $\pi > \tilde{\pi}^c$ , and one player chooses the risky option at all beliefs  $\pi$  such that  $\tilde{\pi}^c \geq \pi > \pi^c$ . Furthermore, there is at most one period in which a single agent experiments.*

*Proof.* Because  $\pi^c \leq \pi^{sc}$  it follows that below  $\pi^c$  the social planner prefers not experimenting over one or two experiments. We calculate the payoff of doing one experiment and then stopping as

$$(1 - \delta)^{\frac{1}{2}} E_\pi + \delta \pi p E_1$$

end the payoff of doing two experiments and then stopping as

$$(1 - \delta)E_\pi + \delta \pi [1 - (1 - p)^2]E_1.$$

Subtracting the latter payoff from the former, we get

$$-(1 - \delta)\frac{1}{2}E_\pi + \delta\pi p(1 - p)E_1$$

which is negative above  $\tilde{\pi}^c$  and positive below. Because it is suboptimal to do more than two experiments below  $\tilde{\pi}^c$  it follows that between  $\pi^c$  and  $\tilde{\pi}^c$ , the planner prefers one experiment to two experiments.

Next, we establish that above  $\tilde{\pi}^c$ , the planner prefers the sequence 2-1 to 1-2. Suppose it is optimal for the social planner to play a strategy in which he experiments one player experiments in one round and in the subsequent round both player experiment. We will prove that the social player can make himself strictly better off by first letting two players experiment, then one player and afterwards using the same strategy as before. In the two considered rounds the expected payoff of the social planner using 2-1 is

$$(1 - \delta) \left\{ E_\pi + \delta \left( \frac{1}{2}E_\pi + \pi[1 - (1 - p)^2]\frac{1}{2}E_1 \right) \right\}.$$

The expected payoff of 1-2 is

$$(1 - \delta)\left(\frac{1}{2}E_\pi + \delta E_\pi\right).$$

Subtracting the latter payoff from the former, we get

$$(1 - \delta)\frac{1}{2} \left\{ \left( E_\pi + \delta\pi[1 - (1 - p)^2]E_1 \right) - \delta E_\pi \right\}.$$

The part in parentheses is positive above  $\pi^{sc}$  by definition, and  $-\delta E_\pi$  is positive above  $\pi^{sc}$  because  $\pi^{sc} \leq \pi^m$ , so the sequence 1-2 will never be used by the social planner above  $\pi^{sc}$ .

This means that if the planner ever switched to 1 above  $\tilde{\pi}^c$ , he would have to continue with 1 until  $\pi^c$  is reached. In a last step, we rule this out by showing that the planner engages in at most one round of experimentation by a single agent before stopping.

If the planner finds it optimal at some belief  $\pi$  to engage in two rounds of experimentation by a single agent and then stop, his expected discounted payoff is

$$(1 - \delta)\frac{1}{2}E_\pi + \delta\pi pE_1 + \delta(1 - \pi p) \left\{ (1 - \delta)\frac{1}{2} \left[ \frac{\pi(1 - p)}{1 - \pi p}E_1 + \frac{1 - \pi}{1 - \pi p}E_0 \right] + \delta\frac{\pi(1 - p)}{1 - \pi p}pE_1 \right\},$$

which must be non-negative. The expression in braces must be non-negative as well or else it would not be optimal to perform one final experiment after a failure. The expected discounted payoff from performing two experiments at once and then stopping is

$$(1 - \delta)E_\pi + \delta\pi[1 - (1 - p)^2]E_1.$$

Subtracting the former payoff from the latter, we obtain

$$\begin{aligned}
& (1 - \delta)^{\frac{1}{2}} [\pi(1 - p)E_1 + (1 - \pi)E_0] + (1 - \delta)^{\frac{1}{2}} \pi p E_1 + \delta \pi(1 - p)p E_1 \\
& \quad - \delta \left\{ (1 - \delta)^{\frac{1}{2}} [\pi(1 - p)E_1 + (1 - \pi)E_0] + \delta \pi(1 - p)p E_1 \right\} \\
& = (1 - \delta) \left\{ (1 - \delta)^{\frac{1}{2}} [\pi(1 - p)E_1 + (1 - \pi)E_0] + \delta \pi(1 - p)p E_1 \right\} + (1 - \delta)^{\frac{1}{2}} \pi p E_1,
\end{aligned}$$

which is strictly positive since the term in braces is nonnegative and  $E_1 > 0$ . So the planner engages in at most one round of experimentation by a single agent before stopping.  $\square$

## 4 Strategic Experimentation

While the case of fully revealing payoffs is special, it is analytically easy and allows us to derive strong results when comparing equilibria with publicly observable payoffs to those when only action choices but not payoffs are observable. With some minor qualifications due to the discrete nature of the experimentation problem, we show that with observable payoffs the amount of experimentation in any PBE equals that of the single-agent problem. We therefore provide a discrete-time foundation for the results in Keller, Rady and Cripps (2005) and show that these results do not depend on the assumption of Markov perfect play. We then show that with privately observable payoffs, more efficient equilibria can be played. Hence, the ability to transmit information allows players to mitigate the free-rider problem.

### 4.1 Strategic Experimentation with Public Payoffs

As both players choose the risky action after any history  $h(t)$  in which a success has been observed, we can restrict our attention to histories with no prior success. We begin our analysis of equilibrium behavior with the observation that in every SPE of the game with public and fully revealing payoffs, both players choose the safe arm after any history that takes their common belief below the social planner's cutoff  $\pi^c$ . To see this, suppose to the contrary that there exists an SPE in which a player experiments at some belief  $\pi < \pi^c$ . From the analysis of the planner's solution, we know that the average of the players' objective functions at  $\pi$  is negative. Consequently there needs to be at least one player who receives a negative expected payoff. By deviating and always choosing the safe arm this player can increase her payoffs.

To get a first intuition for why the total amount of experimentation is limited by the single-agent amount (plus one), consider pure-strategy SPE first. In a pure-strategy equilibrium, in every period  $t$  in which there was no prior success, a player either experiments or not on the path of play. Since in equilibrium players do not experiment below  $\pi^c$ , players can therefore only engage in finitely

many experiments prior to finding a success. There are thus only finitely many periods in which a pure-strategy profile can require players to experiment if there has been no prior success. Now consider the last period in which a player is meant to experiment. In this last period of experimentation, each player knows that if she fails, no player will experiment in future. Hence, she will only be willing to experiment if this is individually optimal. Whenever a single agent is meant to experiment in this last period, therefore, the belief  $\pi$  must be above the single-agent cutoff  $\pi^a$ . When both players are meant to experiment, the belief must also be above  $\pi^a$  because the value of experimenting in this last period is lower if one's fellow player also experiments. Hence, in any pure-strategy equilibrium there can be at most one more experiment than in the single agent case.

Conversely, it cannot be the case that both players stop experimenting at a belief strictly above the single-agent cutoff. The reason is simply that each player—believing that the other player stopped experimenting—would then face the single-agent tradeoff. The following proposition exploits this logic and extends it to mixed-strategy equilibria.

We call the number of times a player chooses the risky arm on the path of play when every experiment is unsuccessful the *amount of experimentation* performed by that player. The *total amount of experimentation* by both players is simply the sum of the individual amounts. The total amount will typically depend on the initial belief and, with mixed strategies, may be a random variable.

**Proposition 2.** *Given an initial belief, let the optimal amount of experimentation in the single-agent problem be  $K$ . In any SPE of the experimentation game with public and fully revealing payoffs, the total amount of experimentation is  $K$  or  $K + 1$ .*

*Proof.* First, consider any history of length  $t$  for which  $\pi(t) < 1$  and  $\pi(t) > \pi^a$ . Since  $\pi(t) < 1$ , no prior experiment has been successful, and because  $\pi(t) > \pi^a$ , the total amount of experiments by both players is less than  $K$ . We now argue that players experiment with probability one at least one more time following any such history. Let  $v^a(\pi(t)) > 0$  be the value of the single-agent problem at the belief  $\pi(t)$ . Let  $\tau$  be the smallest number of periods such that  $\delta^\tau E_1 < v^a(\pi(t))$ . If the probability that player  $i$  experiments at least once in the next  $\tau$  periods is too small, player  $j \neq i$  is strictly better off experimenting in period  $t$ . So the probability that at least one player experiments at least once in the next  $\tau$  periods is bounded away from zero. On the path of play there will thus be another experiment with probability 1.

Next, consider a history of length  $t$  for which  $\pi(t) = \pi < \pi^a$ . Let  $\phi \geq 0$  be the probability with which player  $j$  experiments at time  $t$ . Suppose that the SPE requires player  $i$  to experiment with positive probability. Then she can do no better by switching to the strategy of playing safe now and, in case player  $j$

experiments and is unsuccessful, continuing to play safe forever. This implies

$$\begin{aligned} \delta\phi\pi p E_1 \leq & (1 - \delta) [\pi E_1 + (1 - \pi) E_0] \\ & + \delta \{ \pi [p + \phi p - \phi p^2] E_1 + (1 - \pi [p + \phi p - \phi p^2]) v \}, \end{aligned}$$

where  $p + \phi p - \phi p^2$  is the probability of at least one success, and  $v$  player  $i$ 's continuation value after a double failure, that is, a payoff realization  $X_1(t) = X_2(t) = X_L$ . As  $0 \leq v \leq E_1$ , this in turn requires that

$$0 \leq (1 - \delta) [\pi E_1 + (1 - \pi) E_0] + \delta \{ \pi p E_1 + (1 - \pi p) v \}.$$

As  $\pi < \pi^a$ , we have  $(1 - \delta) [\pi E_1 + (1 - \pi) E_0] + \delta \pi p E_1 < 0$ , and hence  $v > 0$ . So some player must experiment with positive probability in round  $t + 1$  or later. Repeating this step until a time  $t + \tau$  at which  $\pi(t + \tau) < \pi^c$  in the absence of a success, we obtain a contradiction because no player can experiment below  $\pi^c$  in equilibrium.  $\square$

Thus, with a minor qualification due to discrete rather than continuous time, we replicate the finding of Keller, Rady and Cripps (2005) that in the fully revealing case the amount of experimentation is limited by the single-agent amount. For future reference, we note

**Corollary 1.** *Whenever the total amount of experimentation in the planner's optimal (or optimal symmetric) strategy profile exceeds the single-agent amount by more than 1, it cannot be implemented in a subgame perfect equilibrium of the experimentation game with observable payoffs.*

Above, we have fully characterized subgame perfect equilibria in terms of the total amount of experimentation that is carried out on the path of play. These, equilibria, however, may differ in other dimensions such as when agents experiment. For example, just above  $\pi^a$  agents may engage in a war of attrition as to who has to carry out the final experiment. Thus, there may be periods in which no agent experiments. Furthermore, agents may use their communication to coordinate on whether a given agent is meant to experiment in a given period. Below, we nevertheless show that the entire set of equilibrium paths of experimentation with observable payoffs can be replicated in perfect Bayesian equilibria of the game with unobservable payoffs. Moreover, we show that under certain conditions, higher amounts of experimentation can be supported when payoffs are not observable.

## 4.2 Private Payoffs

We begin by noting that truthful communication can easily be sustained with privately observed, fully revealing payoffs. Suppose that after every period of

experimentation players announce a first success by sending the message  $m_i(t) = 1$  and randomize uniformly over all other messages otherwise. Furthermore, after the first success has been announced and both players know that the state of the world is good, suppose there are no meaningful messages any more, that is, both players always randomize uniformly over the interval  $[0, 1]$ . Intuitively, we are then back in the case of public payoffs.<sup>1</sup>

The key observation now is that if players anticipate this communication strategy, truthful communication is 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 player  $j$  believes, communicates and does, so truthfully announcing a success is optimal. After such an announcement, player  $j$  believes with certainty that the state of world is good, and hence will play risky in all future periods independent of what player  $i$  does after the announcement. So if player  $i$  incorrectly announces a success, she cannot infer anything from player  $j$ 's future behavior, so she is weakly better off telling the truth. We thus have the following result.

**Proposition 3.** *For every SPE of the game with public and fully revealing payoffs, there exists a PBE of the game with private payoffs in which the same path of experimentation is followed, that is, the mapping from payoff histories to action profiles is the same in the two equilibria.*

So far, we have established that private information does not hurt players who play subgame-perfect equilibria. Our next, somewhat striking result, establishes that players can often do better. We require players to perform the optimal symmetric amount of experimentation whereafter, on the path of play, they once communicate and announce whether they had a prior success. If so, both players keep experimenting forever; otherwise, both stop experimenting. We punish early deviations (after the initial period) through beliefs: if a player refrains from experimenting at a time when the socially optimal symmetric strategy profile requires her to experiment, then the other player reacts to this out-of-equilibrium event by assigning probability 1 to the good state of the world. Formally, our equilibrium concept would allow us to assign the same optimistic beliefs to a player who observes a deviation at  $t = 0$ . Such beliefs, however, are clearly implausible: a player deviating in  $t = 0$  cannot have seen a prior success, and so we will—in the spirit of sequential equilibrium—require the other player not to update her beliefs in response to the deviation.<sup>2</sup>

---

<sup>1</sup> One caveat here is that players may use a controlled joint lottery to coordinate continuation play. But then, we can replicate the controlled lottery by reserving the first digit for sending a message as to whether the experiment was successful and use all following digits for the controlled lottery.

<sup>2</sup> We could avoid this problem by letting the players observe one draw from the distribution of risky payoffs before the game starts. The following proposition would then hold with  $\pi^m$  replaced by  $\pi^{sc}$ .

**Proposition 4.** *Suppose the players' initial belief is more optimistic than the myopic cutoff  $\pi^m$ . Then there exists a PBE of the game with private and fully revealing payoffs whose experimentation path coincides with that of the planner's optimal symmetric strategy profile.*

*Proof.* (Sketch) Let  $t = 0, 1, \dots, \tau^c$  be the periods in which the planner's optimal symmetric strategy would require both players to experiment if all experiments remained unsuccessful. Observe that any player who had a success by  $t = \tau^{sc}$  is willing to truthfully announce it in round  $\tau^{sc} + 1$ ; expecting this announcement to be truthful, the other player will then choose the risky action forever. If a player stops experimenting in any period  $t \in \{1, \dots, \tau^{sc}\}$ , the other player will assume that the state of the world is good and experiment in all subsequent periods, so the deviating player ceases to learn anything from observing this behavior. Since her belief is strictly above  $\pi^{sc}$ , she thus prefers to experiment and learn the other player's experimentation results.

It remains to construct a continuation equilibrium that deters deviations at  $t = 0$ . Let  $B(-1, \pi)$  be the posterior belief after one failed experiment. Fix a subgame perfect equilibrium of the game with observable payoffs starting with the common prior  $B(-1, \pi)$ . Let  $v$  be the lower of the two equilibrium values in this SPE, and note that  $v \leq v^{sc}(B(-1, \pi))$ . Following a deviation by player  $i$  only, we require player  $j$  to communicate whether he had a success. If he failed, both players' continuation play corresponds to the selected SPE, with player  $i$  obtaining the value  $v$ . So, if player  $i$  experiments at  $t = 0$ , her expected overall payoff is

$$(1 - \delta)[\pi E_1 + (1 - \pi)E_0] + \delta \left\{ \pi[1 - (1 - p)^2]E_1 + (1 - \pi[1 - (1 - p)^2])v^{sc}(B(-2, \pi)) \right\};$$

if she deviates, this payoff is no more than

$$\delta \left\{ \pi p E_1 + (1 - \pi p)v^{sc}(B(-1, \pi)) \right\}.$$

Since the social planner strictly prefers both players to experiment to one player experimenting above  $\pi^m$ , the former payoff exceeds the latter.  $\square$

The belief following a deviation in an early (but not the initial) round may seem somewhat unusual. Intuitively, upon observing such a deviation, a player  $i$  reasons as follows. Clearly,  $j$  was not careful and made a mistake. She must already know that the state of the world is good to be so careless. While this reasoning is compatible with the logic of sequential equilibrium, the equilibrium construction hinges crucially on this particular choice of off-equilibrium beliefs. Our next aim is therefore to show that private payoffs can lead to a more efficient outcome even under the stringent requirement that whenever a player observes a deviation to the safe action, her beliefs become as pessimistic as possible.

**Definition 1** (Pessimistic Beliefs). *We say that a perfect Bayesian equilibrium has pessimistic beliefs if a player who observes a deviation to the safe action believes that the past experiments of the deviating player were all failures.*

**Proposition 5.** *Suppose that the total amount of experimentation in the planner's optimal symmetric strategy profile exceeds the single-agent amount by at least 2. For sufficiently high discount factors, there exists a PBE with pessimistic beliefs in which the total amount of experimentation exceeds the single-agent amount.*

*Proof.* (Rough sketch) Choose a pure-strategy experimentation equilibrium and a player  $i$  who experiments in the last period. Now we do not let  $i$  communicate at the beginning of period  $\tau + 1$  but let  $j$  communicate. Let  $\pi_j(\tau + 1)$  be  $j$ 's belief following truthful past communication. Now let  $i$  experiment with probability  $\phi_i(\tau + 1)$  in period  $\tau + 1$  so that player  $j$ 's belief in period  $\tau + 2$  is such that  $j$  is indifferent between experimenting and not experimenting given that he learns whether or not  $i$ 's prior experiments have been successful. Player  $i$  experiments in period  $\tau + 2$  only if he had a prior success and not otherwise—implying that  $j$  indeed learns whether  $i$ 's prior experiments have been successful. Hence,  $j$  is indifferent between experimenting and not in period  $\tau + 2$ . We choose  $j$ 's probability of experimentation in such a way that  $i$  is indeed indifferent between experimenting and thereby triggering the possibility of another experiment by  $j$  and not experimenting in period  $\tau + 1$ .  $\square$

## References

- BERGEMANN, D. and J. VÄLIMÄKI (2008): “Bandit Problems,” in: *The New Palgrave Dictionary of Economics*, 2nd edition, ed. by S. Durlauf and L. Blume. Basingstoke and New York, Palgrave Macmillan Ltd.
- BOLTON, P. and C. HARRIS (1999): “Strategic Experimentation,” *Econometrica*, 67, 349–374.
- BOLTON, P. and C. HARRIS (2000): “Strategic Experimentation: the Undiscounted Case,” in: *Incentives, Organizations and Public Economics – Papers in Honour of Sir James Mirrlees*, ed. by P.J. Hammond and G.D. Myles. Oxford: Oxford University Press, 53–68.
- BONATTI, A. and J. HÖRNER (2010): “Collaborating,” *American Economic Review*, forthcoming.
- KELLER, G. and S. RADY (2010): “Strategic Experimentation with Poisson Bandits,” *Theoretical Economics*, 5, 275–311.
- KELLER, G., S. RADY and M. CRIPPS (2005): “Strategic Experimentation with Exponential Bandits,” *Econometrica*, 73, 39–68.
- 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.