Available online at www.sciencedirect.com

ScienceDirect

Journal of Economic Theory 169 (2017) 234–269

JOURNAL OF Economic Theory

www.elsevier.com/locate/jet

Learning to disagree in a game of experimentation ☆

Alessandro Bonatti [a,*], Johannes Hörner [b]

[a] *MIT Sloan School of Management, 100 Main Street, Cambridge, MA 02142, USA*
[b] *Yale University, 30 Hillhouse Avenue, New Haven, CT 06520, USA*

Received 22 September 2015; final version received 8 February 2017; accepted 11 February 2017
Available online 21 February 2017

Abstract

We analyze strategic experimentation in which information arrives through fully revealing, publicly observable "breakdowns." When actions are hidden, there exists a unique symmetric equilibrium that involves randomization over stopping times. With two players, this is the unique equilibrium. Randomization leads to dispersion in actions and to belief disagreement on the equilibrium path. The resulting lack of coordination has significant welfare consequences. In contrast, when actions are observable, the equilibrium is pure and welfare improves.

JEL classification: C73; D83; O33

Keywords: Experimentation; Free-riding; Mixed strategies; Monitoring; Delay

1. Introduction

This paper studies games of strategic experimentation in which information arrives through public breakdowns. Such *bad-news* learning processes naturally occur upon the introduction of a new technology that holds out hopes of cost savings but entails risks. Such risky technologies

☆ We would like to thank Francesc Dilmé, Daniel Gottlieb, Gabriele Gratton, Matti Liski, Sven Rady, Tavneet Suri, Juuso Toikka and seminar participants at Bonn, Boston College, Glasgow, MIT, Montréal, Oslo, Paris II, Pittsburgh, PSE, Rochester, Toronto, Toulouse, and SAET for their helpful comments, and we are grateful to Xiaosheng Mu for the excellent research assistance.
* Corresponding author.
E-mail addresses: bonatti@mit.edu (A. Bonatti), johannes.horner@yale.edu (J. Hörner).

http://dx.doi.org/10.1016/j.jet.2017.02.007
0022-0531/

include new drugs and medical devices, and innovative processes such as hydraulic fracturing for oil production. Some technologies that are socially undesirable (perhaps because they impose negative externalities on other sectors) also fit in this broad class. Consider financial fraud or tax evasion when agents have incomplete information about the effectiveness of the detection technology. In all these cases, there also exist significant barriers to information, making unobservable actions a good starting point. For example, the decision to evade taxes is private, but getting caught is typically a public event.

We show that unobservable actions and learning through bad news *must* lead to dispersion in actions (*e.g.*, in the timing of technology adoption). In turn, differences in unobserved actions lead to differences in beliefs, including higher-order beliefs, despite perfectly aligned fundamentals and outcomes that are common knowledge.

To explain how dispersion and disagreement arise, we rely on a strategic experimentation model with exponential bandits. Players choose whether to experiment in the face of purely aggregate uncertainty. Formally, they continuously choose how much weight to assign to a risky action. Externalities are informational only. Players observe only one another's outcomes, not their actions. We assume binary individual outcomes (a "breakdown" or not) and a common binary state of the world (good or bad). Occasional, publicly observable breakdowns occur when a player puts weight on the risky action and when the state is bad. Hence, whereas a breakdown reveals the state of the world to all players, the absence thereof causes objective and strategic uncertainty: inferences regarding the state interact with inferences regarding the actions of others. In the continuing absence of any breakdown, players grow increasingly optimistic about the state over time. As a result, they are tempted to delay their use of the risky arm to free-ride on the experimentation of others.

The game admits a unique symmetric mixed-strategy equilibrium. In particular, no pure-strategy Nash equilibrium exists and, with two players, no asymmetric mixed-strategy equilibrium exists either. In light of the literature, this is surprising because we do not assume discrete action sets: by definition, giving (say) equal weight to both the risky action and its safe alternative is a *pure* action in our framework. Because time is also continuous, mixing is caused not by discreteness but by the intrinsic nature of incentives. This relationship stands in contrast to the experimentation literature, discussed below, in which a "mixed strategy" is merely an interpretation of actions that are interior (*i.e.*, players assign positive weight to both arms) as opposed to extremal ("bang-bang").

In equilibrium, mixing involves each player choosing *at random* a time before which he exclusively plays safe and after which he only plays risky. The distribution of switching times is continuously increasing over an interval, with an atom at the upper end. Despite being indifferent over an entire interval of such random times, players are unwilling to play an interior action during that interval (pure strategies involving such actions are strictly worse). Another way to appreciate the difference is that players are uncertain of the aggregate amount of experimentation undertaken up to a given time.

Randomization over switching times drives the dispersion of beliefs. Not observing a breakdown can be explained in two ways: either the other players have not yet begun experimenting, or the state of the world is good. A player's own choice of action helps him sort through these competing explanations: the earlier he began experimenting himself, the more likely he is to believe that the lack of a breakdown can be attributed to the state of the world being good rather than to the other players waiting to experiment. In the absence of a breakdown, beliefs about the state remain private at all times: while there exists a finite time at which players commonly know that everyone is experimenting, they still do not know when everyone else began experimenting.

Why do players mix? Two forces are combined here. The first and familiar force is mentioned above: free-riding prevents players from adopting the same extremal pure strategy. If his opponent is switching to the risky arm at a given time, then a player's best reply can involve experimenting immediately to avoid wasting time, as nothing will be learned until then, or taking advantage of his experimentation by choosing to wait long enough to benefit from it. In existing experimentation models, this force drives the players' equilibrium choice of interior actions.[1] Here, a second force compels players to choose extremal actions. Experimentation breeds experimentation: a player who deviates from an interior action to an action that places greater emphasis on the risky arm will see his choice confirmed by the absence of a breakdown; this observation makes him more optimistic about the state of the world. If he were indifferent between risky and safe if he had not deviated, his deviation would have led him to strictly prefer experimentation in the future.

Methodological contribution Solving the game requires understanding the structure of the best-reply correspondence, and then exploiting this structure to identify the equilibrium. The main methodological contribution of the paper lies in the first step. We identify the certainty-equivalent problem, that is, a deterministic optimization program that is equivalent to the best-reply problem under uncertainty faced by a player whose opponents play arbitrary strategies. Such a reduction is systematic, if not necessarily explicit, in the literature on strategic bandits: the repeated game of incomplete information is reduced to a stochastic game of complete information via the (piecewise) deterministic process of the posterior probability that the arm is good. This approach is insufficient in our model, as there is also uncertainty regarding the (pure) strategy randomly selected by the other players. A second (deterministic, but arbitrary) process is introduced, which follows no simple recursive dynamics, but summarizes how the strategies of others affect the hazard rate of a breakdown, as evaluated by a given player.[2]

Equipped with these two processes, the certainty-equivalent problem of the best-reply problem can be formulated, and analyzed using standard methods from (deterministic) optimal control.[3] Up to this point, all that is required is that a player cares about the other players' actions only to the extent that they affect the arrival of breakdowns.

On the other hand, the solution of this deterministic problem does depend on the particular payoff structure. For instance, the convexity properties of the objective imply that any best-reply must involve the exclusive use of the safe (resp., risky) arm above (below) a given (but possibly random) threshold. Hence, a strategy (a distribution over functions) can be summarized by one function only, the c.d.f. of the switching time from safe to risky. Of course, solving for the fixed-points in this infinite-dimensional space remains challenging, and forms the bulk of the analysis.

Game-theoretic contribution In our game, breakdowns are common-knowledge events. Yet players receive private signals about the state through the absence of a breakdown, because the informativeness of this signal depends on the player's private experimentation decision. Thus, belief heterogeneity despite public signals bears resemblance to repeated games in which

[1] For example, Keller et al. (2005), Bonatti and Hörner (2011), and Keller and Rady (2015).

[2] Because of the lack of recursive structure, dynamic programming does not appear to be the best method to use, and we rely on Pontryagin's maximum principle instead.

[3] This step applies to the best-reply problem only, and we do not identify a stochastic game of complete information that is equivalent to the game of incomplete information.

players use private strategies. But in contrast to repeated games, equilibria in private strategies are not merely some of many possibilities: there are no other equilibria. Indeed, relative to the literature on strategic experimentation (Bolton and Harris, 1999; Keller et al., 2005; Bonatti and Hörner, 2011, or Keller and Rady, 2015) ours is the first game in which (i) the equilibrium is known to be unique (this is an open question in the case of Bolton and Harris, 1999), and (ii) any equilibrium is in mixed strategies.

Economic contribution and empirical evidence We find that dispersion in actions and dispersion in beliefs add to the cost of under-experimentation due to free-riding: given the overall amount of experimentation, the dispersion in timing is costly; holding his opponents' strategies fixed, a player would be strictly better off if he could determine when his opponents actually began experimenting.

This last statement extends to equilibrium: players are better off in the (symmetric) Markov equilibrium in the game in which they can observe one another's actions. Intuitively, when actions are observable, each player can accelerate the common learning by deviating to the risky arm. In the absence of a breakdown, this leads to greater optimism and more experimentation by others, which in turn alleviates the under-provision problem and improves payoffs. This is in contrast with good-news models, where observable experimentation leads to greater pessimism and depresses further experimentation. Finally, players also benefit from a mediator helping them to coordinate their play via private recommendations.

A growing body of empirical evidence is broadly consistent with our model's findings. Indeed, to the extent that they have been documented, there is substantial dispersion in practices and productivity across firms and industries.[4] In particular, Skinner and Staiger (2007) document U.S. state-level variation in the adoption rates for four technological innovations (hybrid corn, tractors, computers, and beta-blockers) and suggest informational barriers as a potential explanation. Consistent with the idea that barriers to information generate cross-sectional heterogeneity in the new technology adoption rate, Covert (2015) documents frictions in drilling companies' learning processes regarding the relationship between inputs and oil production.[5]

With monitoring used as a design variable, our results in Section 6 help explain the information sharing observed in several industries. Indeed, in health care, industry associations and government agencies promote the sharing of information on best practices. For instance, the U.S. Food and Drug Administration (FDA) recently launched the Unique Device Identification (UDI) system "to adequately identify medical devices through their distribution and use." The UDI system provides information on outcomes and on adoption rates through usage intensity.[6] In oil drilling, regulations encourage sharing information regarding input choices for fracking operations (see Covert, 2015).

[4] A long history of empirical literature has documented heterogeneity in the adoption rates for new technologies: Mansfield (1961) observes patterns of "slow imitation" for a small number of innovations; Coleman et al. (1966) show distinct differences across physicians in the adoption of new medical technology; and more recent studies (Bloom and Van Reenen, 2007; Syverson, 2011; Gibbons and Henderson, 2013) document the wide dispersion in managerial practices within an industry and relate it to persistent productivity differences.

[5] For an example of breakdowns in that context, see "The Downside of the Boom," *The New York Times*, November 22, 2014.

[6] See http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/UniqueDeviceIdentification/ for more details.

Related literature From a theoretical perspective, our work is closest to Keller and Rady (2015) and our earlier paper (Bonatti and Hörner, 2011). Our game differs from the former in that actions are not observed, and from the latter in that the news is bad rather than good. In addition, our earlier work features payoff externalities that are absent from the current model. These differences have significant implications: with good news, the equilibrium is not unique, and the symmetric equilibrium is in interior pure strategies, with experimentation dwindling but never ceasing. With good news, a deviation to greater experimentation leads to increased pessimism and hence less future experimentation; thus, behavior is "mean-reverting," and best replies are necessarily pure. This also explains why, with good news, the Markov equilibrium with observable actions is worse than the symmetric equilibrium with unobservable actions, contrary to what we find with bad news. A comparison with Keller and Rady is in Section 6.

From the point of the motivation, which emphasizes the interplay between experimentation and observational learning, Murto and Välimäki (2011) is close. In their model, players receive exogenous (random) private signals over time. In the present study, however, learning is endogenous, with players' actions influencing their private beliefs. Another major difference is that although signals are private in their model, actions (exit or not) are publicly observed. As a result, the dynamics that they identify are different from ours, with waves of exits alternating with what they term "flow modes" until a collapse ends the game. In our game, unless a breakdown occurs, players' unobserved behavior leads to smooth updating of their beliefs, except for the atom at the last switching time assigned a positive probability.

Mixed strategies are new to the literature on experimentation, as mentioned. As noted, the necessity of considering mixed strategies in our game should not be confused with the necessity of allowing pure actions that are not extremal appearing elsewhere. To restore existence in games of strategic experimentation without needing to confront the measure-theoretic difficulties raised by the modeling of independent randomization in continuous time, various authors (*e.g.*, Bolton and Harris, 1999; Keller et al., 2005; Keller and Rady, 2015) have redefined the space of actions available to a player at a given instant to be a convex set (that is, the set of pure strategies is sectionally convex). This redefinition is usually achieved by simple convexification, replacing the lotteries over $\{0, 1\}$ by the interval $[0, 1]$ (with the interpretation of players choosing how to allocate a unit resource), but is not always accomplished in this way: in Keller and Rady (2003), this redefinition involves players choosing two actions at every instant—a mean price and a mean variance. In these papers, this redefinition suffices to restore the existence of an equilibrium in pure (but not extremal) strategies.[7,8] In fact, there are special games for which privately randomizing over stopping times and using non-extremal pure strategies *are* equivalent. The question of whether to describe equilibrium strategies with a hazard rate or with a pure strategy taking value in a convexified set is then merely a matter of convenience.[9] The following is a manifestation of the difference between the equilibria of the games considered in these papers (whether strategic experimentation or games of timing) and the approach in our study: in these

[7] Pure but non-extremal optimal policies contrast with the solution of decision-theoretic versions of bandit problems, which admit optimal solutions within the class of extremal policies. See Yushkevich (1988) or Presman and Sonin (1990).

[8] One should not confuse such a convexification with some clever application of Kuhn's theorem that would obviate mathematical difficulties. Kuhn's theorem also applies to continuous-time games (see Weizsäcker, 1974, or Shmaya and Solan, 2014), but the set of behavioral strategies (properly defined) is much larger than the set of pure strategies, even when action sets are convex.

[9] Examples include wars of attrition (*e.g.*, Milgrom and Weber, 1985) or more recent versions of timing games allowing for additional learning (Murto and Välimäki, 2011; Rosenberg et al., 2013).

equilibria, a player is indifferent regarding *all* his strategies (over the relevant time interval). In the unique equilibrium of our game, a player is indifferent over stopping times (over some interval), but he strictly prefers any of these stopping times to a strategy that uses an interior action over a set of times of positive measure: he is willing to mix, but not to play the pure strategy that specifies the expected value of the mixture.

Of course, mixed strategies arise in many games of economic interest. In static games with convex action sets, they either arise because the payoff function fails to be continuous, or fails to be quasiconcave. Most well-known examples involve discontinuous payoffs.[10] Here, the problem is not continuity (payoffs are continuous in the weak topology), but quasiconcavity, as the analysis in Section 4.4 makes plain. In a dynamic, continuous-time setting, we are not aware of another paper with a clear economic interpretation in which mixed strategies must be considered to ensure existence, despite convex action sets.[11] Our paper shows that such phenomena are both relevant to economic applications and amenable to mathematical analysis. (See Akcigit and Liu, 2015; Board and Meyer-ter Vehn, 2014, for models in which mixed-strategy equilibria *might* exist.) Note that such equilibria might also arise when outcomes rather than actions are private, as in Rosenberg et al. (2007). Non-Markovian equilibria might also be required for games with incomplete (rather than imperfect) information and payoff externalities, see Décamps and Mariotti (2004).

2. The model

2.1. Setup

Time is continuous, and the horizon is infinite. Players $i = 1, \dots, I$ ($I \geq 2$) choose an action $u^i \in [0, 1]$ at all times.[12]

There is a binary state of the world $\omega \in \{B, G\}$. Players assign a common prior probability $p^0 \in (0, 1)$ to the event $\{\omega = B\}$. Conditional on ω, player i's action controls the instantaneous intensity of a conditionally independent Poisson process $\{N_t^i : t \geq 0\}$. The process N_t^i is interpreted as the number of lump-sum payoffs observed up to time t. That is, the action paths $u^i = (u_t^i)_{t=0}^{\infty}$, alongside ω, define the instantaneous intensity of an inhomogeneous Poisson process with intensity $\lambda(t) := \lambda \mathbf{1}_{\{\omega=B\}}(1 - u_t^i)$, where $\lambda > 0$ and $\mathbf{1}_A$ is the indicator function of an event A. Note that this intensity is zero if $\omega = G$, independent of the actions chosen. We interpret the action u_t^i as the amount of risk-reducing effort exerted by the agent. Thus, when $u^i = 1$, player i pulls the *safe arm* exclusively, which prevents the occurrence of (costly) lump sums. Conversely, when $u^i = 0$, player i pulls the *risky arm* exclusively. Unless a player pulls the safe

[10] See Dasgupta and Maskin (1986), who argue, however, that lack of quasiconcavity is more fundamental, as it arises in simple variations of those games that satisfy payoff continuity. See Blume (2003) for an explicit mixed-strategy equilibrium in a class of Bertrand games. The problem also arises in zero-sum games, see Karlin (1959) and references therein.

[11] Of course, it is well known in optimization that sectional convexity is insufficient to guarantee the type of convexity in the policy space that is required for the existence of solutions of optimal control problems. *A fortiori*, the problem arises in games, and there are well-known examples of zero-sum games with sectionally convex action spaces for which the optimal policies cannot be found within the class of pure policies (see Karlin, 1959, and references therein).

[12] Where possible, we adopt standard notation from optimization theory: controls are u_t, optimal choices of controls are *policies*, adoption times are stopping times (though experimentation starts then), etc.

arm exclusively, he might learn about the state.[13] Hence, we state that player i *experiments* when $u^i < 1$.

The safe arm has a flow cost $s > 0$ and each lump sum entails a cost $h > 0$. That is, given an integrable function $u^i = (u^i_t)$ and the realization of the process $\{N^i_t : t \geq 0\}$, the realized cost of player i is given by

$$\int_0^\infty re^{-rt}(h\mathrm{d}N^i_t + su^i_t\mathrm{d}t),$$

where $r > 0$ is the players' common discount rate. Note that this is a game of informational externalities only, as player $j \neq i$'s actions do not enter player i's cost.

Throughout this setup, we assume that player i observes the realization of the process $\{N^j_t : t \geq 0\}$ for all j, *i.e.*, he can condition his action on the *breakdowns*; however, he observes nothing else. In particular, player i does not observe past values of u^j_t, $j \neq i$. That is, players observe outcomes, but not actions.

We assume that $g := \lambda h > s$. Therefore, conditional on $\{\omega = B\}$, to minimize the expected cost, it is optimal to allocate the resource exclusively to the safe arm, that is, to set $u^i_t = 1$ for all t. Conditional on $\{\omega = G\}$, the risky arm is optimal, independent of other players' actions.

Hence, player i's problem reduces to a course of action up to the first arrival of a breakdown for *any* player, as it is strictly dominant to pull the safe arm thereafter. Let $\tau \in \mathbf{R}_+ \cup \{+\infty\}$ be the time of this first arrival. (Note that $\tau = +\infty$ if $\omega = G$.) Therefore, we can and do assume that the game ends at time τ.

A terminal history h^τ specifies the stopped action paths $\{(u^i_t)^\tau_{t=0} : i = 1, \ldots, I\}$ up to time τ. We can rewrite the cost for which we minimize the expectation as

$$C^i(u^i) = \int_0^\tau re^{-rt}\left(h\mathrm{d}N^i_t + su^i_t\mathrm{d}t\right) + e^{-r\tau}s, \tag{1}$$

where the last term is the "terminal" cost equal to the expected cost over an infinite horizon conditional on $\{\omega = B\}$ under $u^i_t = 1$ for all t.

Some of the parameters are relevant only in combination. In particular, we define the following variables:

$$\gamma := \frac{g - s}{s}, \quad \text{and} \quad \mu := \frac{r}{\lambda}.$$

Thus, up to normalization, g and s enter only through the cost-benefit ratio γ, and by a standard change in the variable, the discount rate r and intensity parameter λ appear via the ratio μ only.

[13] It might be desirable to allow for "background learning," which corresponds to the agent being unable to prevent a breakdown. Thus, learning is slowed when the safe arm is pulled but does not come to a halt. We may then assume that $\lambda(t) := \lambda \mathbf{1}_{\{\omega=B\}}(\bar{u}/I - u^i_t)$, where $\bar{u} > I$. Long-run beliefs are clearly very different in that case, but there is no discontinuity in payoffs or equilibrium policies.

2.2. Policies and equilibrium

A deterministic (or "pure") policy for player i is a measurable function $\pi^i : \mathbf{R}_+ \to [0,1]$ that specifies player i's action u^i at time t conditional on the event $\{t < \tau\}$.[14] We interpret u^i_t as the share of i's resources allocated to the safe arm. Let Π^i denote the set of all deterministic policies. Of special importance are *stopping policies*, which are defined as follows. Given $t \geq 0$, let π^i_t be the policy that sets $\pi^i_t(s) = 1$ for $s < t$ and $\pi^i_t(s) = 0$ for $s \geq t$. The set of stopping policies is denoted Π^i_S.

Ultimately, it is not sufficient to consider deterministic policies. Mixed policies must be introduced. We adopt the following definition of mixed policies based on Aumann (1964). A mixed policy is a measurable map $\phi^i : [0,1] \to \Pi^i$ such that for all $\beta^i \in [0,1]$, $\phi^i(\beta^i) \in \Pi^i$.[15] This definition can be interpreted as follows: player i privately flips a "coin" at the beginning of the game, and the realization β^i of a random variable uniformly distributed on $[0,1]$ determines the deterministic policy that he then follows. Let Φ^i denote the set of (mixed) policies of player i.

Given $\phi^{-i} \in \Phi^{-i} := \times_{j \neq i} \Phi^j$, player i minimizes

$$\mathcal{C}^{\phi^i} := \mathbf{E}^{\phi^i}_{p^0}\left[\mathcal{C}^i(u^i)\right]$$

over $\phi^i \in \Phi^i$.

Of particular interest are *stopping time* policies—"random" stopping policies. According to these policies, for some non-decreasing function $t^i : [0,1] \to \mathbf{R}_+$, $\phi^i(\beta^i) = \pi^i_{t^i(\beta^i)}$ (a.s.). Hence, in these policies, player i randomizes over the time that he stops pulling the safe arm. Let Φ^i_S denote the set of stopping time policies of player i (including Π^i_S). It is often more convenient to represent such policies using the distribution function $F^i : \mathbf{R}_+ \to [0,1]$, defined as $F^i(t) := \sup\{\beta^i \mid t^i(\beta^i) \leq t\}$; that is, t^i is the quantile function of F^i.

Given that players do not observe one another's actions, there is no loss in considering Nash equilibria, relative to refinements like perfection. Hence, an *equilibrium* is a vector $\phi^* \in \Phi := \times_i \Phi^i$ such that for all i and for all $\beta^i \in [0,1]$, $\phi^{*i}(\beta^i)$ minimizes $\mathcal{C}^{\phi^i}$ over $\phi^i \in \Phi^i$, given ϕ^{*-i}. Of particular interest are symmetric equilibria, which are equilibria in which $\phi^j = \phi^i$ for all i, j. However, our attention is not restricted to those equilibria.

3. Learning

Players face two sources of uncertainty. First, they do not know the state of the world. As time passes without breakdowns occurring, they learn about the state. Second, players do not know the specific deterministic policy selected by the other players—if indeed this policy was chosen at random. In this regard, time is also informative: because breakdowns are more likely if others pull the risky arm, the absence of breakdowns is indicative of safe play. Both sources of

[14] The policy does not define behavior after one's own deviation, an unnecessary specification given the information structure. In those rare instances in which we comment on behavior after such off-path histories, we use the word "strategy" instead.

[15] Let $\mathcal{B}_{[0,1]}$ (resp., $\mathcal{B}$) denote the σ-algebra of Borel sets of $[0,1]$ (resp., $\mathbf{R}_+$) with the Lebesgue measure. We endow the set of measurable functions from $(\mathbf{R}_+, \mathcal{B})$ to $([0,1], \mathcal{B}_{[0,1]})$ with the σ-algebra generated by sets of the form $\{f : f(s) \in A\}$ with $s \in \mathbf{R}_+$ and $A \in \mathcal{B}_{[0,1]}$. The notion that such a definition is equivalent to the use of "behavioral decision rules" follows from Weizsäcker (1974). See also Shmaya and Solan (2014) on the equivalence and Touzi and Vieille (2002) on mixed policies in timing games.

uncertainty affect the choice of optimal action: if player i knew that others were experimenting, then he might be tempted to "free-ride" on this experimentation and pull the safe arm unless he is very optimistic about the risky arm.

In addition, players' beliefs are private. A player who adopts a riskier policy becomes optimistic at a faster pace than if he had adopted a safer policy; indeed, if he pulled the safe arm exclusively, he would only learn from others. This statement implies that other players do not know player i's beliefs. Those players have a belief about his belief, as in equilibrium, they know the distribution over deterministic policies that player i is using.[16]

However, given player i's policy, all these beliefs (including higher-order beliefs) are derived from a common source of information: time. Because the game ends with the first breakdown, there is only one information set corresponding to a given time t (conditional on i's policy throughout). Player i faces no "uncertainty" regarding these conditional beliefs: he can perfectly forecast at time t what his beliefs will be at any time $t' > t$, conditional on no breakdown occurring in the meantime. In particular, he can forecast the instantaneous probability with which a breakdown will occur on that date. The *hazard rate* of a breakdown is all that matters for determining best replies, but each player's forecast reflects the two sources of uncertainty that he faces.

Formally, fix a player i throughout. Define $p_t^i := \mathbf{P}_{p^0}^{\phi}[\omega = B \mid (u_s^i)_{s=0}^t]$ for $t < \tau$. As the conditioning clearly indicates, it is player i's belief and his only, although we occasionally omit the superscript. Suppose for now that players $j \neq i$ use pure policies π^j. Given that breakdowns follow an exponential distribution, the probability of a breakdown not occurring by time t (denoted $\varnothing_t$), conditional on $\{\omega = B\}$, is given by

$$H_t := \mathbf{P}[\varnothing_t \mid \omega = B] = e^{-\int_0^t \lambda\left(I - \sum_{j=1}^I u_s^j\right)\mathrm{d}s}.$$

Under pure policies, Bayes' rule reduces to the ordinary differential equation

$$\dot{p}_t^i = -\lambda p_t^i (1 - p_t^i)\left(I - \Sigma_{j=1}^I u_t^j\right), \quad p_0^i = p^0, \tag{2}$$

where $\lambda\left(I - \Sigma_j u_t^j\right) = -\partial \ln H_t / \partial t$ is the hazard rate of a breakdown.

If instead players $j \neq i$ use mixed policies ϕ^j, we must derive the law of motion of the (deterministic) process p_t^i taking into account the uncertainty regarding the realized policies π^j. The probability of no breakdown occurring by time t, conditional on $\{\omega = B\}$ and $(u_s^i)_{s=0}^t$, is given by

$$\begin{aligned} H_t := \mathbf{P}_{p^0}^{\phi}[\varnothing_t \mid \omega = B] &= \mathbf{E}_{p^0}^{\phi}\left[e^{-\int_0^t \lambda\left(I - \sum_{j=1}^I u_s^j\right)\mathrm{d}s} \mid (u_s^i)_{s=0}^t\right] \\ &= e^{-\int_0^t \lambda(I - u_s^i)\mathrm{d}s} \prod_{j \neq i} \mathbf{E}_{p^0}^{\phi}\left[e^{\int_0^t \lambda u_s^j \mathrm{d}s}\right], \end{aligned}$$

where the last equality follows from the independence of the players' policies. Because the first term on the right-hand side is only a function of player i's own action, uncertainty appears only via the second term.

[16] However, this second-order belief is not common knowledge because player j's posterior belief regarding i's adopted policy depends on j's belief regarding the state of the world. In turn, this belief depends on his own policy, which is not observed by others.

Indeed, player i's belief is private, but his private information appears separately, as captured by the hazard rate

$$-\frac{\partial \ln H_t}{\partial t} = \lambda(I - u_t^i) - \sum_{j \neq i} \frac{\partial}{\partial t} \ln \mathbf{E}_{p^0}^{\phi}[e^{\int_0^t \lambda u_s^j \mathrm{d}s}]. \tag{3}$$

Thus, the contribution to *his* belief attributable to all other players' expected policies is common knowledge. We then define

$$\nu_t^{-i} := \sum_{j \neq i} \frac{1}{\lambda} \frac{\partial}{\partial t} \ln \mathbf{E}_{p^0}^{\phi}[e^{\int_0^t \lambda u_s^j \mathrm{d}s}]. \tag{4}$$

Note that $\nu_t^{-i} \in [0, I-1]$ because $u_s^j \in [0,1]$, all $s \leq t$, $j \neq i$. In particular, $\nu_t^{-i} = 0$ means player i is certain that all other players are playing risky at time t, and $\nu_t^{-i} = I - 1$ means he is certain they are playing safe.

Because the function ν^{-i} plays an important role in the analysis, it is important to develop some intuition for it. Because breakdowns follow an exponential distribution, the function ν^{-i} is, in general, different from the expected action of the other players $\mathbf{E}_{p^0}^{\phi}[\sum_{j \neq i} u_t^j]$. Instead, ν^{-i} measures the expected contribution of the other players' experimentation to the hazard rate of the first breakdown.[17]

Therefore, the experimentation of player $j \neq i$ affects player i's belief revision at time t, and it is not simply a matter of whether player j is playing safe at that time. The entire path of player j's actions affects player i's belief regarding the state of the world at time t and, hence, how much this belief must be revised if no breakdown occurs in the next instant. It follows from (3) that p^i is also differentiable and that it solves the differential equation

$$\dot{p}_t^i = -\lambda p_t^i (1 - p_t^i)(I - u_t^i - \nu_t^{-i}), \quad p_0^i = p^0. \tag{5}$$

Thus, if $\nu_t^{-i} = I - 1$ and $u_t^i = 1$ then $\dot{p}_t^i = 0$. If, in addition, all players use stopping time policies, this implies that $p_t^i = p^0$.

Now suppose that players use stopping time policies such that $\phi \in \Phi_S$. Hence, players switch from the safe arm to the risky arm at time t according to some distribution function $F^j : \mathbf{R}_+ \to [0,1]$. Write $\bar{F}^j = 1 - F^j$ for the complementary distribution function. In that case, we obtain an alternative, perhaps more expressive, formula for ν^{-i}. By definition,

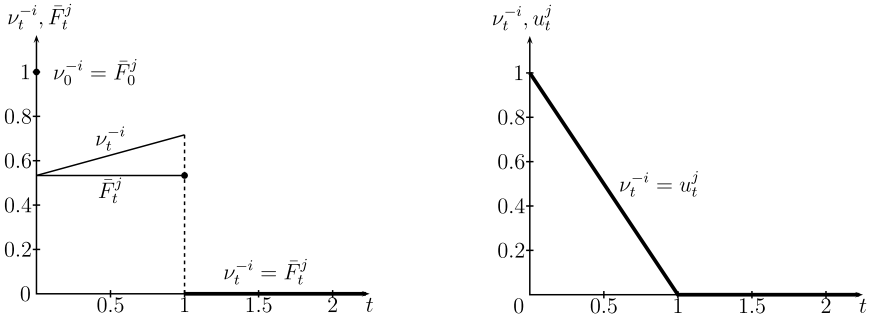
$$\nu_t^{-i} = \frac{1}{\lambda} \sum_{j \neq i} \frac{\partial}{\partial t} \ln \mathbf{E}_{p^0}^{\phi}[e^{\int_0^t \lambda \mathbf{1}_{\{s \leq t^i(\beta^i)\}} \mathrm{d}s}] = \frac{1}{\lambda} \sum_{j \neq i} \frac{\partial}{\partial t} \ln \left[e^{\lambda t}(1 - F_t^j) + \int_0^t e^{\lambda s} \mathrm{d}F_s^j \right],$$

such that, explicitly,

$$\nu_t^{-i} = \sum_{j \neq i} \frac{e^{\lambda t} \bar{F}_t^j}{\bar{F}_0^j + \int_0^t \lambda e^{\lambda s} \bar{F}_s^j \mathrm{d}s} \geq \sum_{j \neq i} \frac{e^{\lambda t} \bar{F}_t^j}{1 + \int_0^t \lambda e^{\lambda s} \mathrm{d}s} = \sum_{j \neq i} \bar{F}_t^j. \tag{6}$$

It follows that ν_t^{-i} is a function that begins at $I - 1$, remains there as long as $F^j(t) = 0$ for all $j \neq i$, discontinuously decreases when F^j discontinuously increases for some $j \neq i$, and

[17] Given a mixed strategy profile ϕ, one can also construct agent i's time-t posterior beliefs over $\Sigma_{j \neq i} u_t^j$. However, the other players' actions affect player i's payoff through the hazard rate of a breakdown only.



(a) Mixed policy (b) Pure policy

Fig. 1. Hazard rate ν^{-i} compared with $\bar{F}^j$ and u^j, for $I = 2$.

continuously increases when $t \notin \cup_{j \neq i} \operatorname{supp} F^j$, strictly, unless it is equal to 0, which occurs when $F^j(t) = 1$ for all $j \neq i$.[18] It always exceeds the total probability of others not having switched to their risky arm, as the likelihood of a breakdown is lowest when they are still pulling the safe arm. Finally, ν_t^{-i} increases when t is not in the support of any F^j because in the absence of a breakdown player i assigns growing weight to subsequent realizations of his opponent's switching time.

Fig. 1 illustrates this scenario for the case of two players from the perspective of player i: in the case of some (arbitrary) pure policy (when player j selects a deterministic policy u^j), the hazard rate ν^{-i} coincides with it; in the case of a mixed policy where player j randomizes between switching to $u = 0$ at $t = 0$ and at $t = 1$, ν^{-i} and $\bar{F}^j$ coincide (at least) at the initial instant and once they reach 0.[19] For $t \in [0, 1]$, ν^{-i} increases over time and is given by $\nu_t^{-i} = e^{\lambda t}/(1 + e^{\lambda t})$.

4. Best replies

This section elucidates the structure of the best-reply function of a fixed player i, taking the behavior of others as given, as summarized by the hazard rate ν^{-i}. From the definition of ν_t^{-i}, it is clear that player $j \neq i$'s past actions (at dates $s < t$) impact future values of $\{\nu_s^{-i}\}_{s>t}$. Furthermore, the current value of ν_t^{-i} (or of any finite-dimensional vector) cannot summarize this impact in general: because ν_t^{-i} involves the expectation of a nonlinear function, knowledge of the interdependence between past and future actions becomes necessary to compute future values of ν_t^{-i}. We cannot get a recursive structure for the law of motion of ν^{-i}, and so must take this function as given. Therefore, we can find a certainty-equivalent problem for the best-reply problem, but cannot find such a deterministic representation for the game (unlike in the usual bandit problems with observable actions, which can be reformulated as stochastic games with the belief as state variable).

Our approach is indirect: by analyzing the (deterministic) best-reply problem, we obtain that any candidate equilibrium strategy can be summarized by a distribution function. *Given* this structure, the game can be reduced to a deterministic one, and its equilibria can be found. Our procedure involves five steps. First, we explain why this hazard rate is indeed a summary statistic

[18] Given a distribution G, we write $\operatorname{supp} G$ for the set of points of increases in G.

[19] To be clear, ν_t^{-i} isn't linear in the left panel, despite visual deception.

for the best-reply problem. Second, we show that any best reply is necessarily a stopping time policy. Third, we derive the unique cooperative solution as an immediate by-product, in which $v^{-i} = I - 1$. Fourth, in the case of two players, we solve for the best-reply function and show how its structure—first increasing and then decreasing to 0—eliminates the possibility of the existence of an equilibrium in pure policies. Fifth, we explain why this non-existence extends to the case of more than two players.

4.1. The certainty-equivalent problem

The optimization problem faced by player i satisfies certainty equivalence (in the sense of filtering theory): *the optimal action u_t^i is exactly the same as it would be if all unknowns were known and if their values equaled their best estimates (the conditional expectations), given by* (p^i, v^{-i}). Furthermore, a separation principle (again, in the sense of filtering theory) holds: optimal estimation and optimal control can be decoupled. That is, as is apparent from the definition of (p^i, v^{-i}), the choice of u^i does not affect this estimate.

We may now rewrite the problem of minimizing $\mathcal{C}^i$. First, we can express the probability of the event that no breakdown has occurred by time t in terms of p_t^i; by the martingale property of beliefs,

$$\mathbf{P}^{\phi}_{p^0}[\varnothing_t] \cdot p_t^i + (1 - \mathbf{P}^{\phi}_{p^0}[\varnothing_t]) \cdot 1 = p^0,$$

so that

$$\mathbf{P}^{\phi}_{p^0}[\varnothing_t] = \frac{1 - p^0}{1 - p_t^i}.$$

We then obtain the following certainty-equivalent minimization problem

$$\min \int_{t \geq 0} e^{-rt} \left(r p_t^i g\left(1 - u_t^i\right) + r u_t^i s + \lambda p_t^i (I - u_t^i - v_t^{-i}) s \right) \frac{1 - p^0}{1 - p_t^i} \mathrm{d}t \tag{7}$$

over measurable policies $\pi^i : \mathbf{R}_+ \to [0, 1]$, subject to (5), *i.e.*,

$$\dot{p}_t^i = -\lambda p_t^i (1 - p_t^i)(I - u_t^i - v_t^{-i}), \quad p_0^i = p^0.$$

This is the program $\mathcal{P}$.[20] Here, the function $v^{-i} : \mathbf{R}_+ \to [0, I - 1]$ is treated as an exogenous (measurable) function. Yet, we omit it as an explicit argument of $\mathcal{P}$. By the Filippov–Cesari theorem (see Cesari, 1983), a solution exists, that is to say, the infimum is achieved. We will examine the necessary conditions given by Pontryagin's maximum principle.

The interpretation of the objective is as follows. As explained above, $(1 - p^0)/(1 - p_t^i)$ is the probability of reaching time t without a breakdown. At that time, if player i invests u_t^i in the safe arm, then the rate at which he suffers a breakdown is $(1 - u_t^i)\lambda p_t^i$, with the expected cost rh. If any of the players has a breakdown (which occurs at rate $\lambda p_t^i (I - u_t^i - v_t^{-i})$), then player i switches to the safe arm, yielding the net present cost s. As was the case for learning dynamics, the pair (p^i, v^{-i}) also summarizes all the information required to compute payoffs.

[20] With a slight abuse: the program $\mathcal{P}$ examined in the Appendix is a minor modification of it.

4.2. Stopping time policies

Here, we show that any best reply must be within the class of stopping time policies.

Lemma 1. *If $\pi^i \in \Pi^i$ solves $\mathcal{P}$, then $\pi^i \in \Pi^i_S$.*

Informally, Lemma 1 states that if a player begins experimenting, he should do so indefinitely (*i.e.*, until a breakdown occurs), and conversely, if he plays safe, he must have played safe at all earlier times.

To gain further intuition, we examine the optimal *timing* of a fixed amount of experimentation from player i's perspective. Thus, we consider the *arbitrage equation* of player i, which describes the trade-off between *backloading* and *frontloading* experimentation, *i.e.*, between shifting an amount of experimentation $\mathrm{d}u$ across the time intervals $[t, t+\mathrm{d}t)$ and $[t+\mathrm{d}t, t+2\mathrm{d}t)$. (See the proof of Lemma 1 for the formal argument.) We will show that player i prefers to backload experimentation over the relevant range of beliefs. The marginal value of backloading experimentation is given by

$$r \cdot \underbrace{\left(p_t^i g - s\right)}_{\text{flow cost}} + \underbrace{\lambda p_t^i \left(I - u_t^i - v_t^{-i}\right)}_{\text{breakdown rate}} \cdot (g-s) - \lambda p_t^i \cdot \underbrace{\left(1 - u_t^i\right)}_{\text{change in action}} \cdot (g-s). \tag{8}$$

The first term is the time-preference effect of delaying the expected flow cost $p_t g$ and anticipating the cost s. The second term pertains to the event of a breakdown (at rate $\lambda p_t^i \left(I - u_t^i - v_t^{-i}\right)$): if so, safe play would occur at $t+\mathrm{d}t$ regardless of the player's earlier action; in that event, pulling the safe arm more at t yields marginal savings of $g-s$. Finally, the third term considers the effect of the player's action on the likelihood of a breakdown: by frontloading safe play, the player reduces (at a rate λp_t^i) the arrival of a breakdown, in which case he would switch from the current action u_t^i to $u^i = 1$; because this scenario can occur only in the bad state, this action yields a loss $g-s$.

Note that the sum of the last two terms is non-negative. Hence, equation (8) implies that backloading is profitable when p is sufficiently large. Conversely, if a player were certain that the state is good, discounting would suggest frontloading the risky action. Lemma 1 then establishes that over the relevant range of beliefs (*i.e.*, for $p^i \geq p^*$; see Lemma 2), the marginal value of backloading experimentation is positive.

The stopping-time property of best replies is not only a feature of bad-news learning. In a good-news model with no payoff externalities and unobservable actions, every best reply involves frontloading the risky action.[21] In other words, the pure interior action paths described by Keller et al. (2005), Keller and Rady (2015), and Bonatti and Hörner (2011) rely on either observable actions (the former two) or payoff externalities (the latter). Instead, the welfare comparisons drawn in these papers do not rely on the presence of payoff externalities.

Finally, note that Lemma 1 does not imply that the solution to $\mathcal{P}$ is unique; rather, it implies that all deterministic solutions are in Π^i_S. Furthermore, one can determine the bounds on how early or late a player is willing to switch to the risky arm. Next, we provide such bounds in terms of player i's beliefs.

[21] This result can be obtained by adapting Theorem 1 in Bonatti and Hörner (2011) to the case of pure informational externalities.

We define

$$\frac{p^*}{1-p^*} := \frac{\mu+I}{\mu+I-1}\frac{1}{\gamma},$$

as well as

$$\frac{p^{**}}{1-p^{**}} := \frac{\mu+1}{\mu}\frac{1}{\gamma},$$

where we recall that $\gamma = (g-s)/s$ and $\mu = r/\lambda$.

As follows from these definitions, $p^{**} > p^*$ for $I \geq 2$. The next result establishes that once a player becomes sufficiently optimistic (specifically, when $p_t^i < p^*$), he allocates his entire resource to the risky arm.

Lemma 2. *If π^i solves $\mathcal{P}$, then $u_t^i = 0$ for all t such that $p_t^i < p^*$. Conversely, if $p_t^i > p^*$, then $u_t^i = 0$ implies $v_t^{-i} > 0$.*

Lemma 2 establishes a lower bound on experimentation: the belief p^* is the threshold value at which a player is willing to experiment even if all other players are pulling the risky arm thereafter. The lower bound p^* depends on the number of players because of the amount of information generated when $I-1$ players pull the risky arm exclusively.

Note that, if $p^0 < p^*$, Lemma 2 implies there exists a unique equilibrium: all players choose $u_t^i = 0$ at all times. In what follows, we assume that $p^0 \geq p^*$. The next result establishes a tight upper bound on the amount of experimentation.

Lemma 3. *If π^i solves $\mathcal{P}$, then $u_t^i = 1$ for all t such that $p_t^i > p^{**}$. Conversely, if $p^0 \leq p^{**}$, then $u_0^j = 0$ for some j, that is, some player starts experimenting immediately.*

The upper bound p^{**} coincides with the threshold belief for the single-agent problem. This result is familiar in exponential-bandit models with good news (Keller et al., 2005, and Bonatti and Hörner, 2011) in which players are not willing to experiment more than in the single-player case. However, it contrasts with the result of Keller and Rady (2015) for a bad news model with observable actions (see Section 6 below). Lemma 3 immediately implies that there exists a unique equilibrium for $p^0 > p^{**}$: all players choose $u_t^i = 1$ at all times, and beliefs are "frozen" at their initial level.

4.3. Cooperative solution

We briefly mention the cooperative solution, which is implied by Lemma 3. Assume that players perfectly observe one another's actions (an innocuous assumption because the optimum involves pure policies) and act to minimize the sum of their costs. We define

$$\frac{p^{FB}}{1-p^{FB}} := \frac{\mu+I}{\mu}\frac{1}{\gamma}.$$

Note that p^{FB} is larger than p^{**}, given $I \geq 2$. In fact, it coincides with p^{**} and also p^* when $I = 1$ is inserted into the formulas.

Given a pair (p,u) such that p is the belief path generated by $u := \sum_{i=1}^I u^i$, given p^0, along the history with no breakdown, the action path $(u_t)_t$ is *measurable with respect to the belief path* $(p_t)_t$ if $p_t = p_{t'} \Rightarrow u_t = u_{t'}$ for all t, t'. We write $u(p)$ for the value of u at belief $p \leq p^0$, which

is then well defined. The cooperative solution given in the next lemma is measurable with respect to its belief path.

Unsurprisingly, the optimal policy involves all players employing the risky arm if beliefs are below a threshold p^{FB} and the safe arm otherwise. Thus, if $p^0 \geq p^{FB}$, then all the players play the safe action forever, and if $p^0 < p^{FB}$, all the players immediately use the risky action from the start and keep doing so until the first breakdown. This is formally stated below (see also Keller and Rady, 2015, Proposition 1). In addition, Lemma 4 establishes that total costs decrease in the intensity with which the risky arm is pulled, as long as the ranking of the total amount of experimentation holds pointwise in the beliefs.

Lemma 4. *The cooperative solution u^{FB} is given by $u_t^{FB} = I$ for all t such that $p_t \geq p^{FB}$ and $u_t^{FB} = 0$ otherwise. Furthermore, let $p', p'' : \mathbf{R}_+ \to \mathbf{R}$ be two feasible paths such that the corresponding action paths u', u'' are measurable with respect to their belief path, with $u^{FB}(p) \leq u'(p) \leq u''(p)$ for all $p \leq p^0$. The cost is then weakly lower under p' than under p'' and strictly lower when $u'(p'_t) < u''(p'_t)$ for a set of times t of positive measure.*

4.4. Best replies with two players

To understand why equilibrium is necessarily in mixed policies (unless $p^0 \notin (p^*, p^{**})$), it is useful to derive the best-reply correspondence in the special case of two players. Suppose that player $j \neq i$ switches (with probability one) to the risky arm at time t^j. We may distinguish player i's cost according to whether he switches to the risky arm first or second.

If player i decides to go second, he must do so when his private belief reaches the threshold p^* (or immediately if this belief has been reached by the time j switches). Hence, if going second is best, then player i's best reply must be

$$t^i = t^j + \lambda^{-1} \ln\left(\frac{p^0}{1-p^0} \Big/ \frac{p^*}{1-p^*}\right) =: t^j + t^*.$$

The fixed delay t^* defined above is equal to the time required for beliefs to reach the threshold p^* based on player j's experimentation alone.

If i decides to go first and preempt player j, he begins experimenting immediately. In other words, if moving first is the preferred course of action, then moving immediately is best. Intuitively, player i will not learn before time t^j unless he experiments. If player i is not willing to wait until then, he should begin immediately. Conversely, if delaying experimentation is not too costly, then player i will choose to "freeze" beliefs until t^j.

What remains to be determined is when player i prefers to go first or second. This preference depends on when player j switches. Unsurprisingly, the larger t^j is, the more tempting it is to go first. Intuitively, if t^j is very high, then the cost of waiting until player j's actions take beliefs to the threshold causes an overly costly delay in learning. Conversely, when player j is expected to switch to the risky arm soon, the benefits of free-riding on his experimentation when beliefs are most pessimistic outweigh the cost of delay. This behavior is summarized in the following lemma.

Lemma 5. *The best-reply correspondence $t^i : \mathbf{R}_+ \rightrightarrows \mathbf{R}_+$ is given by, for some $\hat{t} \in \mathbf{R}_+$,*

$$t^i(t^j) = \begin{cases} t^j + t^* & \text{if } t^j < \hat{t}, \\ \{0, \hat{t} + t^*\} & \text{if } t^j = \hat{t}, \\ 0 & \text{if } t^j > \hat{t}. \end{cases}$$

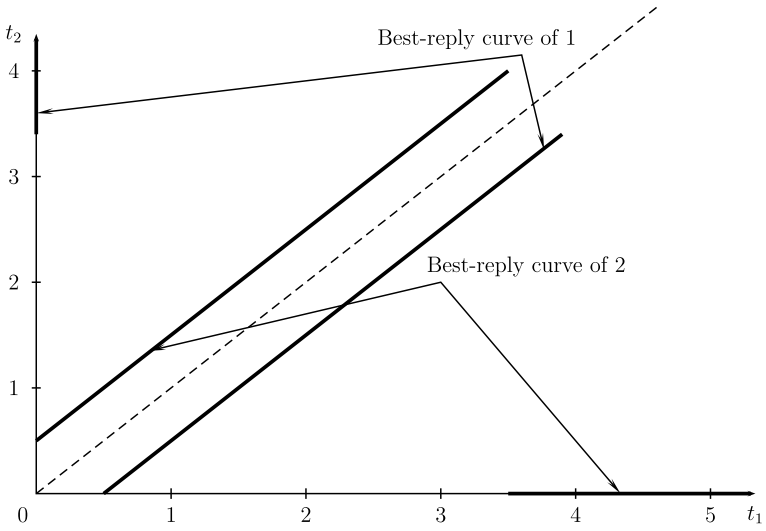


Fig. 2. Best-reply curves with two players for $(r, \lambda, \gamma, p^0) = (1/10, 1, 4, 1/2)$.

Consequently, player i's best-reply curve jumps downward at $\hat{t}$. Fig. 2 provides an illustration. Because the two best-reply curves do not cross, no pure-policy equilibrium exists.

To obtain some intuition, suppose there exists an asymmetric equilibrium in which player i switches immediately to the risky arm and player j waits. Then player j must optimally wait until time t^* when his belief is $p^j_{t^*} = p^*$. However, one can show that waiting, say, until $2t^*$ is a profitable deviation for player i. At time t^*, player j switches, and at time $2t^*$, player i's belief reaches p^*. This is true for all parameter values. For example, consider the discount rate r. Intuitively, for low r, waiting is not very costly; and for high r, the risk of a quick breakdown looms large on player i.

More generally, for the case of two players, let $\mathcal{C}(t_1, t_2)$ denote the cost of player i if he switches at time t_1 and player j switches at t_2. It is not hard to show that the cost function $\mathcal{C}$ is subadditive, *i.e.*,

$$\mathcal{C}(t_1, t_2) + \mathcal{C}(t_2, t_1) \geq \mathcal{C}(t_1, t_1) + \mathcal{C}(t_2, t_2),$$

with strict inequality whenever t_1 and t_2 are distinct. This observation rules out asymmetric pure equilibria since, if 1 does not want to mimic 2, $\mathcal{C}(t_1, t_2) \leq \mathcal{C}(t_2, t_2)$, and similarly, if 2 does not want to mimic 1, $\mathcal{C}(t_2, t_1) \leq \mathcal{C}(t_1, t_1)$. Subadditivity of the cost function indicates the potential cost savings (at the aggregate level) associated with coordinating stopping times between agents. We revisit this property when studying correlated equilibria in Section 6.

Finally, symmetric equilibria in which both players use pure strategies are also ruled out by the characterization of best-responses in Lemma 5 (*i.e.*, there is no stopping time t with $t_i(t) = t$). Therefore, the only candidate equilibrium is one in which one player uses a pure strategy and the other player use a mixed strategy. This requires one player to choose exactly time $\hat{t}$. If player j uses $\hat{t}$, then player i has two best replies. However, consider Fig. 2: for each of his choices, player j's best reply would be much smaller than $\hat{t}$ (indeed, it would be 0 if i used the larger best reply and t^* if he uses 0). Unsurprisingly, regardless of how player i randomizes between these two choices, player j's best reply is strictly lower than $\hat{t}$. We immediately obtain the following result.

Lemma 6. *Suppose that $I = 2$ and $p^0 \in (p^*, p^{**})$. There exists no equilibrium in which either player uses a pure policy.*

As discussed after Lemma 2, all players choose the risky arm when $p_t^i \leq p^*$. Thus, for $p^0 \leq p^*$, there exists a unique equilibrium, which is pure. Conversely, suppose $p^0 \geq p^{**}$. Because p^{**} is the threshold belief for a single agent, there exists a pure strategy symmetric equilibrium in which all players choose the safe arm forever. Theorem 1 below shows this is the unique symmetric equilibrium of the game for those parameters.

4.5. More than two players

Can an equilibrium in pure policies exist when $I > 2$? Deriving best-reply curves is no longer an easy task. However, a pure-policy equilibrium cannot exist based on the following simple argument. Suppose that such an equilibrium exists, and let t^i denote the time at which player i switches to the risky arm. Without loss of generality, suppose that $t^1 \geq t^2 \geq \cdots$. Suppose first that $p(t^3) > p^*$. Consider the game starting at time t^3 and the corresponding initial belief $p(t^3)$. This game involves only two players, players 1 and 2 (assuming indeed that t^3 is optimal for player 3). A necessary condition for the policy profile to be an equilibrium is that players 1 and 2 play mutual best replies in this game. Yet, the two-player game admits no pure-policy equilibrium. If instead $p(t^3) = p^*$, then given Lemma 3, because $p(t^I) \geq p^{**}$, there exists j such that $p(t^j) > p^* = p(t^{j-1}) = \cdots = p(t^1)$. As in the two-player case, past time t^j, any player $i = 1, \ldots, j-1$ would gain from deviating to the risky arm immediately.[22]

5. Main results

5.1. Symmetric equilibrium

We now turn to the equilibrium analysis. Recall that we assume throughout that $p^0 > p^*$. Given F^{-i} and, hence, given ν^{-i}, each time $t \in \operatorname{supp} F^i$ is such that the stopping policy π_t^i is a solution to $\mathcal{P}$. Furthermore, it holds that, given any $t \in \operatorname{supp} F^i$, $p_t \geq p^*$. We let $\bar{t}^i := \max\{t \in \mathbf{R}_+ : t \in \operatorname{supp} F^i\}$.

First, we focus on symmetric equilibria and accordingly write F, $\bar{t}$ for F^i, $\bar{t}^i$, unless we emphasize a given player's perspective. The next result derives the unique symmetric equilibrium of the game.

Theorem 1. *There exists a unique symmetric equilibrium. If $p^0 \geq p^{**}$, then the equilibrium is pure and involves $F^i(t) = 0$ at all times.*

If $p^0 \in (p^*, p^{**})$, the equilibrium involves mixed policies. Specifically, player i chooses a stopping policy π_t among the set $[0, \bar{t}]$, with $\bar{t} > 0$ and $p_{\bar{t}} = p^*$; this distribution is positive and continuous over $(0, \bar{t})$ and has an atom at times $t = 0, \bar{t}$.

[22] For any number of players I, the proof of Lemma 5 establishes that if players $-i$ switch at some time $t = T$, then player i wants to switch at a different time.

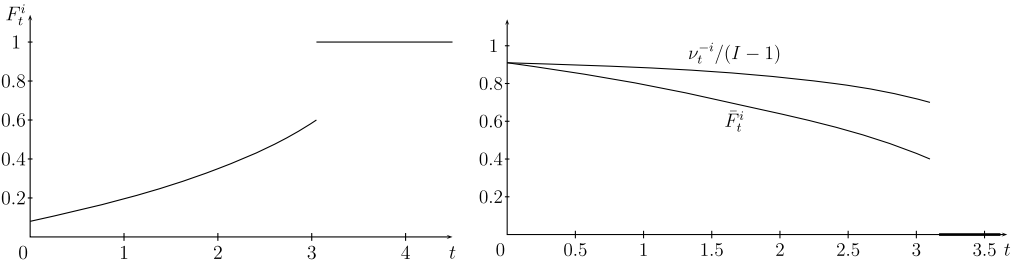


Fig. 3. Equilibrium distributions F_t^i (left), $\bar{F}_t^i$ and hazard rate ν_t^{-i} (right) for $(r, \lambda, \gamma, I, p^0) = (1/10, 1, 4, 3, 1/2)$.

The equilibrium distribution function can be solved in closed form as

$$\bar{F}(t) = \left(\frac{A - e^{\mu t}}{A-1}\right)^{\frac{1}{I-1}} \left(1 - \frac{\mu}{(I-1)(Ae^{-\mu t} - 1)}\right),$$

where we normalized λ to 1 and

$$A := \left(1 - \gamma \frac{\mu}{1+\mu} \frac{p^0}{1-p^0}\right)^{-1},$$

$$\bar{t} = \frac{1}{\mu} \ln \left(\frac{I-1}{(I+\mu-1)\left(1+\mu-\gamma\mu\frac{p^0}{1-p^0}\right)}\right).$$

The initial atom is given by $F_0^i = 1 - \nu_0^{-i}/(I-1)$. In equilibrium, this is a decreasing function of p^0 that vanishes as $p^0 \to p^{**}$ and converges to $(I+\mu)^{-1}$ as $p^0 \to p^*$. The final atom is also a decreasing function of p^0, which is bounded away from 1 even as $p^0 \to p^{**}$, and converges to $1 - (I+\mu)^{-1}$ as $p^0 \to p^*$, consistent with our earlier result that players choose $u_t^i = 0$ at all times when $p^0 \leq p^*$.

Fig. 3 illustrates the equilibrium distribution (left panel) and compares the complementary distribution function $\bar{F}_t^i$ with the hazard rate ν_t^{-i} (right panel). Over time, players learn from their own experience and from that of others. In particular, as time passes, a player assigns growing weight to the event in which his opponent has already switched to the risky arm, conditional on which, learning occurs faster. Moreover, the contribution of any other player's experimentation to player i's learning $(1 - \nu^{-i}/(I-1))$ is always smaller than that player's distribution function F_t^i because, as time passes and no breakdown occurs, player i also assigns growing weight to subsequent realizations of his opponent's switching time, which slows this learning process.

The maximum range of stopping times in the symmetric equilibrium has a natural interpretation: the "earliest" that a player may switch to the risky arm is when his belief is p^{**}: this is the belief for which he would switch if he were on his own (cf. Section 4.3). The latest he might switch is when his belief reaches p^*: this would be his uniquely optimal belief if all others were always experimenting. Because his opponents' behavior lies somewhere between these two extremes, so does his set of best replies. Indeed, the second atom occurs at the time t when the belief of a player who has not yet switched reaches p^*.

In contrast to typical mixed-strategy equilibria of normal-form games, player j does not *need* to randomize over stopping policies to make his opponents indifferent over all stopping times in the relevant time interval. Player j could achieve this through the deterministic policy $u^j = \nu^{-i}/(I-1)$. However, player j is not willing to play such a deterministic but interior policy.

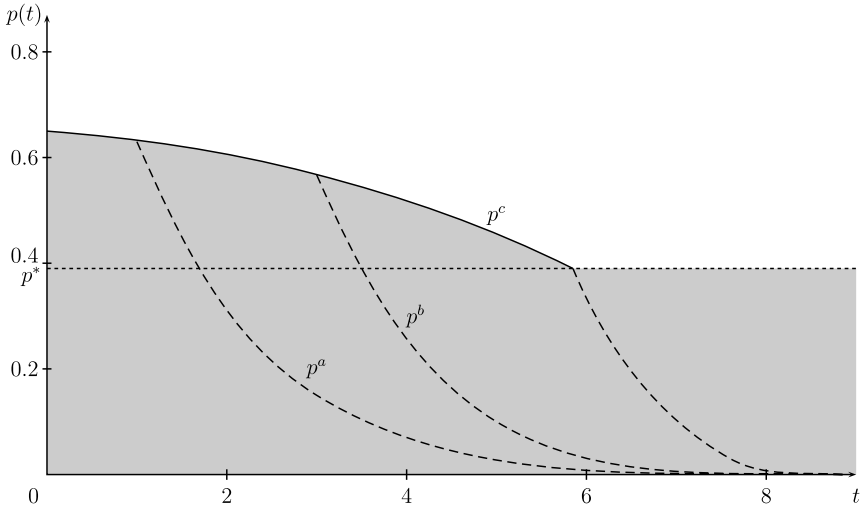


Fig. 4. Belief trajectories for $(\gamma, p^0, I, \mu) = (3, 13/20, 2, 1/8)$.

Indeed, randomizing over stopping policies is the unique *cost-minimizing* way to make the other players indifferent over their own stopping times.

Randomized stopping times have rich implications for the dispersion of equilibrium beliefs. The shaded area in Fig. 4 indicates (time, belief) pairs for which an agent pulls the risky arm. Fig. 4 also illustrates belief paths as a function of time and behavior. For instance, p^c is the belief path of player i who chooses the latest possible equilibrium stopping time, $\bar{t}$. The solid line indicates when he pulls the safe arm; the dashed line indicates when he has already switched to the risky arm. Trajectories p^a and p^b correspond to earlier switching times. Once the player begins pulling the risky arm, his belief decreases faster, reinforcing his preference for the risky arm (absent any breakdown). This elucidates the "off-path" behavior of player i. After an arbitrary history $(u_s^i)_{s=0}^t$ (along which he might have deviated from the prescribed behavior), Lemmas 1–2 remain valid: player i's optimal policy is a stopping policy (from time t onward) that prescribes stopping no later than the first time his belief reaches p^*.

As discussed, an agent's decision whether to start experimentation depends on his higher-order beliefs. Depending on his own stopping time, he entertains different beliefs about the state of the world, whether his opponent has started experimented himself, etc. Fortunately, alongside his own stopping time, the function ν^{-i}, which is common knowledge, summarizes all relevant information for this hierarchy, so that the optimal policy can be described as a function of this pair only.

Finally, the necessity to randomize is a robust phenomenon: as Fig. 2 clearly indicates, given that best-reply curves vary continuously with the parameters, the non-existence of pure-policy equilibrium is robust to perturbations in parameters, regardless of whether symmetry is preserved. Furthermore, it is not necessary to consider that when the safe arm is pulled, no learning occurs. Our results generalize to the case containing background learning. In that case, even if the initial belief is above p^{**}, players use stopping time policies in the unique symmetric equilibrium, which is mixed. However, the earliest stopping time within the support of the equilibrium policy corresponds to the time when the players' beliefs reach p^{**}.

Derivation of the equilibrium distribution Fix the other players' behavior in terms of ν_t^{-i}, and consider player i's stopping time $t^i = T$. The first-order effect of playing safe longer (differentiating (7)) is given by

$$\frac{e^{-\mu T}}{1-p_T}(\mu s - p_T(\mu g + s)) + \int_T^\infty e^{-\mu t}\left(\mu g + (I - \nu_t^{-i})s\right)\frac{p_t}{1-p_t}\mathrm{d}t.$$

The first term (which is negative) captures the myopic benefit (cost reduction) of playing safe longer. The second term is instead the added cost of slower learning, which is captured by a higher hazard rate of a breakdown at all future times.

Pointwise indifference requires the marginal cost of playing safe longer to be nil over the entire support. Thus, we turn to the second-order effect of playing safe, which is given by the sum of the following four terms,

$$-\frac{e^{-\mu T}\mu}{1-p_T}(\mu s - p_T(\mu g + s)) - (I - \nu_T^{-i} - 1)\frac{e^{-\mu T}p_T}{1-p_T}(\mu s - (\mu g + s))$$
$$-\frac{e^{-\mu T}p_T}{1-p_T}\left(\mu g + (I - \nu_T^{-i})s\right) + \int_T^\infty \frac{e^{-\mu t}p_t}{1-p_t}\left(\mu g + (I - \nu_t^{-i})s\right)\mathrm{d}t.$$

The second-order effect is given by (a) the delayed myopic benefit, (b) the lower myopic benefit (note that $(I - \nu_T^{-i} - 1)p_T/(1-p_T)$ is the derivative of the hazard rate), (c) the postponed cost of diminished learning, and (d) the higher marginal cost of diminished learning (because the hazard rate is exponential in u^i). Because the first-order condition must hold pointwise on the support, the last term is equal to the myopic (first-order) benefit of delaying switching.

These four terms can be combined into an expression characterizing the equilibrium ν^{-i} as a function of the belief p,

$$\frac{p}{1-p}(g-s)\left(I + \mu - \nu^{-i} - 1\right) - s(\mu + 1). \tag{9}$$

Note that these beliefs are those of the most pessimistic type of player i, *i.e.*, the player who has not yet switched to the risky arm.

Next, we use the law of motion of beliefs to derive ν_t^{-i} as a function of time alone. We then derive the equilibrium distribution from the definition of ν^{-i} as

$$\frac{\nu_t^{-i}}{I-1} = \frac{(1-F_t)e^{\lambda t}}{(1-F_t)e^{\lambda t} + \int_0^t e^{\lambda s}\mathrm{d}F_s},$$

which yields a differential equation for F_t, resulting in

$$F_t = 1 - \frac{\nu_t^{-i}}{I-1}e^{\int_0^t\left(\frac{\nu_s^{-i}}{I-1} - 1\right)\mathrm{d}s},$$

and we then plug the formula for ν_t^{-i} from equation (19) in the Appendix.

5.2. Uniqueness

Lemma 6 above rules out asymmetric equilibria in pure (deterministic) policies, but is silent about asymmetric equilibria in mixed policies. As Fig. 2 clarifies, our game is not supermodular: in particular, best-reply curves are not monotone, which implies standard methods to prove

uniqueness fail.[23] Moreover, the different methods and tricks described in Karlin (1959) do not appear to be effective. In the case of two players, it can be shown that no other equilibrium exists.[24] Our proof carries no philosophical charm and is based on particular features of the payoff function.

Theorem 2. *Assume that $I = 2$. The equilibrium is then unique (and thus equal to the mixed equilibrium of Theorem 1).*

Uniqueness contrasts with the multiplicity that is prevalent in games with strategic experimentation, not only when actions are observable (Keller et al., 2005; Keller and Rady, 2015) but also when they are not (Bonatti and Hörner, 2011). Because of the pervasive free-riding incentives, asymmetric equilibria typically exist when players alternate (finitely or infinitely often) between experimenting and taking advantage of the opponent's experimentation—leading to the existence of additional asymmetric equilibria. By contrast, in our game, free-riding finds its expression in how early a player is willing to begin experimenting; the earlier the opponent begins experimenting, the later the player finds it optimal to do so. However, the *ordering* of actions is unambiguous: for a given total amount of experimentation, it is always optimal to use a stopping policy, pulling the safe arm if and only if a threshold time has not yet been reached.[25] In other words, no player can ever have an incentive to use a policy that involves pulling the risky arm before the safe arm, precluding any type of alternation in the experimentation that players conduct.

5.3. Comparative statics

As the number of players increases, the free-rider problem worsens in terms of both the timing and the amount of experimentation.

In computing beliefs, we encounter a difficulty: the belief that player i holds at a given time is not uniquely determined in the mixed equilibrium; the earlier a player stops, the lower is his belief at a given time t, provided that no breakdown has occurred. We are thus led to adopt the perspective of an *outside observer* who observes no actions at all: conditional on a given time t being reached without a breakdown under either informational assumption, what probability does he attach to the event $\{\omega = B\}$? In the observable case, this belief coincides with that of any player, at least on path. In the unobservable case, it is some weighted average of a player's belief where the weight reflects the probability attached by this observer to a player switching to risky play at a given time, suitably updated given that time t is reached without a breakdown. Formally, we compute

$$p_t = \mathbf{P}^{\phi}_{p_0}[\omega = B \mid \varnothing_t],$$

[23] See Vives (1999) for an excellent discussion. The best-reply function is not a contraction either (otherwise, the equilibrium would be pure), and the fact that the equilibrium is mixed implies that the Gale–Nikaido theorem or the Poincaré–Hopf theorem cannot work either; or rather, that one should work with the mixed-strategy space directly and possibly use an infinite-dimensional extension of those.

[24] For more than two players, uniqueness is an open problem.

[25] This is also the key reason why the equilibrium must be in mixed policies instead of pure policies with interior actions: for a given amount of experimentation, players have strict incentives to backload experimentation, eliminating the possibility of pulling arms with interior intensity over any interval of time.

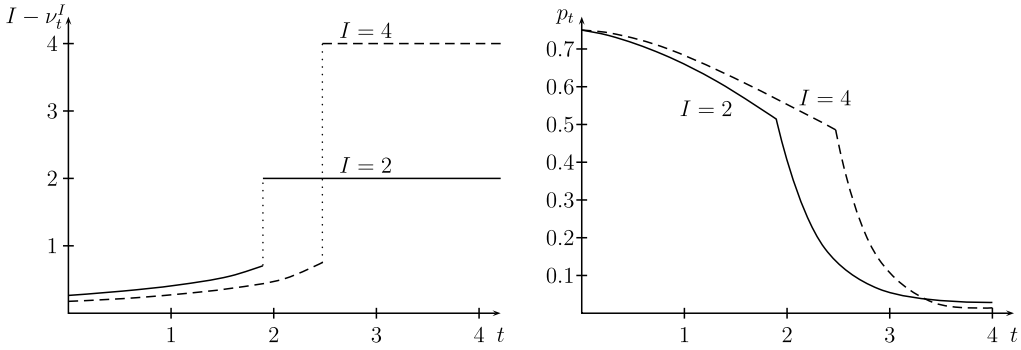


Fig. 5. Hazard rate and belief paths for $(\mu, \gamma, p^0) = (1/4, 1, 3/4)$.

where, unlike in Section 3, we do not condition on any particular player's action path. It follows that the outside observer's belief satisfies

$$\dot{p}_t = -p_t(1-p_t)(I - \nu_t^I), \quad p_0 = p^0,$$

where ν_t^I captures the expected hazard rate of a breakdown from his perspective.[26] For the purpose of the next proposition, we index distributions and stopping times, among others, by the number of players $I \geq 1$.

Proposition 1.

1. *The distributions F^I are ranked by stochastic dominance: F_t^I decreases in I, for all t.*
2. *For an outside observer, $\nu_t^{I'} \geq \nu_t^I$ for all $I' > I$, with strict inequality for all $t < \bar{t}_{I'}$.*
3. *For $I' > I$, the belief path $p_t^{I'}$ crosses p_t^I exactly once (from above).*
4. *For all $I > 1$, total individual costs in the symmetric equilibrium are given by*

$$p^0(g\mu + s) - \mu s. \tag{10}$$

In summary, as the number of players increases, the "mixing phase" lasts longer (until $\bar{t}_I$) and drives beliefs to a lower threshold p_I^*. Free-riding incentives are sufficiently strong that adding more players leads to slower learning, and no improvement in each player's welfare. Moreover, adding one player does not modify the upper bound on experimentation p^{**}. Thus, although the social planner would like to experiment under more pessimistic prior beliefs as I increases, there can be no experimentation in equilibrium if the prior is above a constant threshold.

The distributions of stopping times F^I are ranked by first-order stochastic dominance: a larger number of players increases the likelihood of later stopping times. Furthermore, the expected hazard rate from the outside observer's perspective $I - \nu_t^I$ is decreasing in I as long as p^* has not been reached. The outside observer's belief facing $I' > I$ players eventually overtake the belief that he would hold with I players. In Fig. 5, we illustrate the hazard rate $I - \nu_t^I$ and the belief paths for $I = 2, 4$.

[26] Adjusting equation (4), we define $\nu_t^I := \sum_{j=1}^I \frac{1}{\lambda} \frac{\partial}{\partial t} \ln \mathbf{E}_{p^0}^{\phi}[e^{\int_0^t \lambda u_s^j \, \mathrm{d}s}]$.

6. The role of information

We begin by recalling Keller and Rady's result regarding symmetric Markov equilibria in the game with observable actions. Players are restricted to Markov policies $u^i : [0,1] \to [0,1]$ with the left limit p_{t-} of the common posterior belief as the state variable. Policies are required to be left-continuous and piecewise Lipschitz. We define $u^o : [0,1] \to [0,1]$ as

$$u^o(p) = \begin{cases} 1 & \text{if } p \geq \bar{p}, \\ \frac{I+\mu-1}{I-1} - \frac{\mu\left(\ln(p/(1-p)) - \ln(p^*/(1-p^*))\right)+1}{(I-1)(\gamma p/(1-p)-1)} & \text{if } p \in [p^*, \bar{p}), \\ 0 & \text{if } p < p^*, \end{cases}$$

where $\bar{p} > p^*$ is an "upper threshold" belief.[27]

Theorem 3 *(Keller and Rady, 2015). In the game with observable actions, the unique symmetric Markov equilibrium is given by u^o.*

It is worth emphasizing that this is not the unique Markov equilibrium: asymmetric Markov equilibria exist, and the ranking in terms of welfare can go either way. (See Section 3.3 of Keller and Rady, 2015.) Theorem 3 is established by Keller and Rady (2015), but the closed-form expression for the policy is a contribution of this paper.

Under observable actions, players benefit from a larger group, although not at the same rate as the social planner. Furthermore, as the number of players (and hence the value of information) grows, the first-best policy eventually involves immediate full experimentation. Each player's cost then converges to its level under complete information: the arrival rate of a breakdown grows, conditional on the bad state, and the probability of suffering a breakdown is inversely proportional to I. This relationship cannot be found when actions are not observable because experimentation does not even begin unless $p^0 < p^{**}$. Even under observable actions, the threshold belief $\bar{p}$ for experimentation to begin is increasing in I but converges to a finite value. Furthermore, the duration of the mixing phase does not decline as the number of players grows. Therefore, the cost under complete information is not attainable if $p^0 > p^*$.

Next, we compare the total amount of experimentation up to some t under both observable and unobservable actions. We write p_t^o, p_t^n and p_t^{FB} for these beliefs, depending on whether we consider the observable, unobservable or cooperative case, respectively. We can show stronger results than the ranking of the belief paths. For the unobservable case, let $v(p) \coloneqq v^I_{t(p)}$, where $t(p)$ denotes the time at which the outside observer's belief reaches a value of p. We may rank $v(p)$ across the three cases.

Proposition 2. *The following inequalities hold for all p:*

$$v^n(p) \geq v^o(p) \geq v^{FB}(p).$$

[27] Specifically, we define $\bar{p}$ as (the unique solution of)

$$\frac{\bar{p}}{1-\bar{p}} \coloneqq \frac{p^*}{1-p^*} \exp\left(-\frac{1+\mu}{\mu} - W_{-1}\left(-\gamma \frac{p^*}{1-p^*} e^{-\frac{1+\mu}{\mu}}\right)\right),$$

where W_{-1} is the (negative branch of the) Lambert function.

The second inequality is strict when $p < p^{FB}$*, and the first is strict when* $p^o < \bar{p}$*. In particular,* $\bar{p} > p^{**}$.

An immediate consequence of Proposition 2 is that, for all t,

$$p_t^n \geq p_t^o \geq p_t^{FB},$$

with strict inequalities as described in the previous proposition. Furthermore, Lemma 4 implies the ranking of the symmetric equilibrium costs $\mathcal{C}(p^0)$, given prior belief p^0.

Corollary 1. *The following inequalities hold for all* p*:*

$$\mathcal{C}^n(p) \geq \mathcal{C}^o(p) \geq \mathcal{C}^{FB}(p).$$

Both inequalities are strict when $p > p^*$.

Hence, monitoring is helpful in our context, although it is not helpful with good news. The basic intuition is easy to grasp: when actions are observable, a player's incentive to deviate is related not only to the direct cost or benefit from this deviation but also to the indirect cost or benefit in terms of the change in actions by other players. By deviating to the risky arm, a player accelerates the common learning that, in the absence of news, leads to greater optimism and more experimentation by others; this outcome is good because players do not experiment enough. By contrast, with good news, experimentation by a player leads to greater pessimism in the absence of news and hence depresses experimentation provision.[28]

Coordination Forcing players to disclose their actions might not be easy to achieve in practice. A less drastic intervention might involve introducing a disinterested intermediary who makes private but correlated recommendations to each player regarding when they should begin experimenting. Mediation is a particularly weak form of intervention; it is self-enforcing and costless. Its formal implementation requires no more than a private correlation device, but in practice, this mediation is undertaken by trade associations, political representatives, or any institution commonly involved in the social dialogue.

Clearly, the optimal correlation scheme should be in the extensive form: there is no benefit in telling a player when to switch before the mediator intends for him to do so, as telling him the specific stopping time in advance only makes it more difficult to induce compliance with the recommendation, giving him more information than needed. However, as we explain, *even* in the normal form (by telling each player privately at the beginning of the game when he should stop playing safe), such correlation is helpful.[29]

[28] Observable actions discourage experimentation with good news regardless of whether payoff externalities or pure informational externalities are present. For the case of payoff externalities, see Theorem 2 in Bonatti and Hörner (2011). For the case of pure informational externalities and unobservable actions, all best replies are stopping-time policies as in Lemma 1 (see the earlier discussion in Section 4.2). Thus, the model of Keller et al. (2005) with unobservable actions admits a unique symmetric equilibrium, in which players pull the risky arm until their beliefs reach the single-agent threshold. In contrast, the unique symmetric Markov equilibrium under observable actions features the same amount of experimentation but unbounded delay.

[29] We are unable to solve for the optimal correlation scheme in the extensive form. In fact, even in the normal form, we are able to solve for it only in the special case of a particular parametrized family of correlation schemes, as described below. But this case suffices to show that independence is not optimal.

Specifically, consider the case of two players. With slight abuse of notation, we denote the joint distribution over switching times in our symmetric equilibrium as $F(t_1, t_2) = F(t_1)F(t_2)$. We construct a new distribution by slightly perturbing the independent randomization according to a bivariate FGM copula.[30] Let ρ denote the correlation parameter of the joint distribution F.

We modify our equilibrium marginal distribution to introduce a small amount of correlation and preserve incentives. At an abstract level, the incentive-compatibility constraint for obeying the recommendation to switch at time t is a functional equation that is linear in the distribution F. We can then write this constraint as the combination of two *linear* operators K_0 and K_1. In particular, we have

$$K_0(F) + \rho K_1(F) = 0.$$

We can use this constraint to capture the restriction that incentives (under a small amount of correlation) impose on the marginal distribution. In particular, we identify a distribution that we use to (locally) modify our equilibrium distribution while preserving incentives. We denote this distribution by $F_1(t; \rho)$.

Clearly, regardless of the degree of correlation ρ, no player can begin experimenting before p^{**} or after p^*. The design variable is the degree of correlation but requires adjusting the support of the marginal distribution to match p^* of the most pessimistic type. In particular, the mass point at time $\bar{t}$ is now a function of ρ. We then differentiate total costs under the distribution $F_1(t; \rho)$ in a neighborhood of $\rho = 0$.

For any value of the parameters, the derivative of the cost is negative, *i.e.*, positive correlation is beneficial. We conclude that some (possibly small) amount of positive correlation of switching times (subject to incentive compatibility) improves upon independence.[31]

However, the role of positive correlation (across switching times) must not be confused with that of pure vs. mixed policies. Recall that only stopping time policies are optimal for any player. It is then important and immediately clear that the symmetric ("coordinated") pure-policy profile $\{u_t^i\}_{i=1}^I$, where $u_t^i = \nu_t^{-i}/(I-1)$, yields strictly higher costs than our mixed equilibrium. Again, from each player's perspective, it is irrelevant whether others randomize (holding ν^{-i} fixed). However, the best-reply problem admits only switching-policy solutions—it is costlier for a player to use the pure (non-extremal) policy u_t^i defined above than to use the distribution F^i over stopping times that is equivalent to ν_t^i, from the perspective of the other players.

7. Conclusions

Our results rely on a number of assumptions. Here, we briefly discuss how we expect them to extend in two important dimensions.

Inconclusive bad news A complete analysis under a scenario of inconclusive bad news (that is, when a breakdown does not reveal the state) seems out of reach. However, we believe that

[30] For a marginal distribution $G(t)$, the Farlie–Gumbel–Morgenstern (FGM) copula is given by $G(t_1, t_2) = G(t_1)G(t_2)(1 + \rho(1 - G(t_1))(1 - G(t_2)))$, with parameter $\rho \in [-1, 1]$. See Nelsen (2006). Throughout this case, we assume symmetry of this distribution, and we introduce an (arbitrarily small amount of) background learning, *i.e.*, $\dot{p}_t^i = -p_t^i(1 - p_t^i)(\bar{u} - u_t^i - \nu_t^j)$, with $\bar{u} > 2$.

[31] The details of the calculations leading to this comparative statics result are in the annotated Mathematica file `correlated.nb` on the authors' websites.

the belief-disagreement result would become more pronounced. First, if all agents stop experimenting upon observing a breakdown, then learning stops and beliefs freeze at different levels depending on the agents' prior actions. Such endogenous belief heterogeneity has an effect on policy effectiveness, *e.g.*, if an external agent (the government) were to attempt to subsidize the risky arm to resume experimentation. Second, behavior after the first breakdown can potentially differ across players. In particular, those agents who were experimenting earlier may revert to the risky arm for some time, whereas others may not. Conditional on the true state, this dispersion in actions causes more persistent performance differences, relative to the case of conclusive news.

Monitoring and payoff externalities in games of strategic experimentation More radical policy interventions that modify the payoffs of the game can be further helpful. In the working paper (Bonatti and Hörner, 2015) we show that risk-sharing has several advantages over other types of interventions. By risk-sharing, we refer to a well-calibrated group-insurance scheme whereby a player who suffers a breakdown obtains partial compensation from the other players. First, such a scheme restores the first-best outcome in contrast to, for instance, externally funded subsidies that improve the amount of experimentation without solving the coordination problem. Second, the optimal scheme is robust to the specific monitoring structure: whether players observe one another's actions is irrelevant to the calibration of this scheme.

We have already remarked on the different welfare implications of observable actions in the models of strategic experimentation with good and bad news. Observability improves welfare under bad news, whereas it is detrimental under good news. This result has differing implications for outcomes and provides a clear, if stylized, criterion to guide policy interventions depending on the nature of the technology. In particular, in the good-news case, unobservable actions eliminate inefficient delay but preserve the suboptimal *amount* of experimentation. Thus, subsidies can be used to augment the amount of experimentation. Under bad-news learning, we highlighted three sources of inefficiency: players experiment too little, with excessive dispersion in both actions and beliefs. Subsidies are able to address the first source only, and group insurance may be more appropriate.

Appendix

Throughout the appendix, proofs are facilitated by working with log-likelihood ratios. We define $\ell_t := \ln(p_t/(1-p_t))$, as well as $\ell^* := \ln(p^*/(1-p^*))$, $\ell^{**} := \ln(p^{**}/(1-p^{**}))$.

Appendix A. Reformulation of the objective

Here we reformulate each player's objective, and we keep track of additional cost terms that will be necessary for comparative statics. Each player minimizes

$$\int_{t\geq 0} e^{-rt}\left(rp_t g\left(1-u_t^i\right) + ru_t^i s + \lambda p_t (I - u_t^i - v_t^{-i})s\right)\frac{1-p^0}{1-p_t}\mathrm{d}t, \tag{11}$$

subject to

$$\dot{p} = -\lambda p_t(1-p_t)(I - u_t^i - v_t^{-i}), \quad p_0 = p^0.$$

Let us do the transformations one by one, first rewriting the objective in terms of the log-likelihood ratio. The objective becomes

$$\int_{t\geq 0} e^{-rt}\left(re^{\ell_t}g(1-u_t^i)+ru_t^i s(1+e^{\ell_t})+\lambda e^{\ell_t}(I-u_t^i-v_t^{-i})s\right)(1+e^{\ell^0})^{-1}\mathrm{d}t.$$

Next, we make the change of variable $t \mapsto t/\lambda$, and we define $\gamma := (g-s)/s$ and $\mu := r/\lambda$. Finally, we factor out $(1+e^{\ell^0})^{-1}$ to get

$$\int_{t\geq 0} e^{-\mu t}\left(\mu e^{\ell_t}g+\mu(s(1+e^{\ell_t})-ge^{\ell_t})(\dot{\ell}_t+I-v_t^{-i})-\dot{\ell}_t e^{\ell_t}s\right)\mathrm{d}t.$$

Integrating the last term yields

$$e^{\ell^0}s+\int_{t\geq 0} e^{-\mu t}\left(e^{\ell_t}(\mu g+\mu(s-g)(\dot{\ell}_t+I-v_t^{-i}))+\mu s(\dot{\ell}_t+I-v_t^{-i})-\mu s e^{\ell_t}\right)\mathrm{d}t.$$

Integrating the first two terms by parts, and factoring out s, we obtain the following expression for the expected cost:

$$W(\ell^0) := \frac{s(1+\mu\gamma)}{1+e^{-\ell^0}} + \frac{\mu s}{1+e^{\ell^0}}\int_{t\geq 0} e^{-\mu t}\left(\mu(\ell_t-\ell^0)-\gamma(I-v_t^{-i}-1+\mu)e^{\ell_t}+I-v_t^{-i}\right)\mathrm{d}t. \quad (12)$$

Therefore, ignoring constant terms, player i minimizes

$$\int_{t\geq 0} e^{-\mu t}\left(\mu\ell_t-\gamma(I-v_t^{-i}-1+\mu)e^{\ell_t}\right)\mathrm{d}t,$$

subject to

$$\dot{\ell}_t = u_t^i + v_t^{-i} - I.$$

Appendix B. Proofs for Section 4

Proof of Lemma 1 The proof of this lemma relies on the proof of Lemma 2, proved next and independently (except for the last sentence of that next proof, which is not used here).

We apply the maximum principle to $\mathcal{P}$. It is easy to see that the program $\mathcal{P}$ is not abnormal (see Seierstad and Sydsæter, 1987, Ch. 2.4, Note 5).[32] The maximum principle implies that there exists an absolutely continuous $\psi : \mathbf{R}_+ \to \mathbf{R}$ such that (i) $\psi_t > 0 \Rightarrow u_t^i = 0$, (ii) $\psi_t < 0 \Rightarrow u_t^i = 1$, and (iii) almost everywhere

$$\dot{\psi}_t e^{\mu t} = \gamma(I-v_t^{-i}-1+\mu)e^{\ell_t}-\mu.$$

Because $v_t^{-i} \leq I-1$, a sufficient condition for $\dot{\psi}_t > 0$ at any time t such that $\ell_t \geq \ell^*$ is that

$$\gamma\mu e^{\ell^*} > \mu.$$

Using the definition of ℓ^*, this is equivalent to

[32] The argument given Seierstad and Sydsæter (1987) must be slightly modified, as it applies to a fixed horizon. The adjustment is straightforward.

$$(\mu + I)\mu \geq \mu(\mu + I - 1),$$

which is true.

It follows that ψ is strictly increasing at all times t such that $\ell_t \in [\ell^*, \ell^0]$; hence, given (i), there exists $\bar{t} \geq 0$ such that any solution must specify $u_t^i = 1$ for all $t < \bar{t}$ and $u_t^i = 0$ for $t \geq \bar{t}$ (recall that $u_t^i = 0$ when $\ell_t < \ell^*$).

Proof of Lemma 2 Consider the continuation cost corresponding to the objective (12), defined as the value from setting $u^i = 0$ (identically), given ℓ and t,

$$\mathcal{C}(\ell, t) := \int_{s \geq t} e^{-\mu s} \left(\mu(\ell + \chi_s) + \gamma(\nu_s^{-i} - I - \mu + 1) e^{\ell + \chi_s} \right) \mathrm{d}s, \tag{13}$$

where $\chi_s := \int_{\tau=t}^{s} (\nu_\tau^{-i} - I) \mathrm{d}\tau$. Note that, integrating by parts,

$$\mathcal{C}(\ell, t) = e^{-\mu t}(\ell - \gamma e^{\ell}) + \int_{s \geq t} e^{-\mu s} \left(\mu \chi_s + \gamma e^{\ell + \chi_s} \right) \mathrm{d}s,$$

which is differentiable with respect to ℓ, with

$$\frac{\partial \mathcal{C}(\ell, t)}{\partial \ell} = e^{-\mu t} \left(1 - \gamma e^{\ell} + \gamma e^{\ell} \int_{s \geq t} e^{\chi_s - \mu(s-t)} \mathrm{d}s \right).$$

This derivative is minimized by setting $\nu_\tau^{-i} = 0$ for all $\tau \geq t$. In that case, the right-hand side is equal to

$$e^{-\mu t} \left(1 - \gamma e^{\ell} + \frac{\gamma e^{\ell}}{I + \mu} \right),$$

which is positive if *and only if* $\ell < \ell^*$. Hence, independently of ν^{-i}, $\mathcal{C}(\ell, t)$ is strictly increasing in ℓ whenever $\ell < \ell^*$. It follows that, for $\ell < \ell^*$, $\mathcal{C}$ solves the Hamilton–Jacobi–Bellman ("HJB") equation

$$\frac{\partial \mathcal{C}(\ell, t)}{\partial t} + \min_{u^i} \left\{ \frac{\partial \mathcal{C}(\ell, t)}{\partial \ell} (u_t^i + \nu_t^{-i} - I) + e^{-\mu t} \left(\mu \ell_t - \gamma(I - \nu_t^{-i} - 1 + \mu) e^{\ell_t} \right) \right\} = 0,$$

so that setting $u_t^i = 0$ is optimal. Because of the "if and only if" above, if $\nu_s^{-i} = 0$ for all $s \geq t$ (for which it suffices that $\nu_t^{-i} = 0$), yet $\ell_t = \ell > \ell^*$, it cannot be that $u_s^i = 0$ for all $s \geq t$ (and so it must be that $u_t^i > 0$).

Proof of Lemma 3 Ignoring some irrelevant constants, the continuation cost (13) can be rewritten as (abusing notation for $\mathcal{C}$)

$$\mathcal{C}_t^i := \frac{e^{-\mu t}}{\mu} \left(\mu \gamma e^{\ell_t^i} \int_t^{\infty} e^{\int_t^s (\nu_\tau^{-i} - (\mu + I)) \mathrm{d}\tau} \mathrm{d}s - 1 \right). \tag{14}$$

We first establish the upper bound ℓ^{**} on the amount of experimentation. Differentiating (14) with respect to t we obtain

$$\mathcal{C}_t^{i'} e^{\mu t} = 1 - \gamma e^{\ell_t^i} + \gamma e^{\ell_t^i} \int_t^\infty e^{\int_t^s (\nu_\tau^{-i} - (\mu+I))\mathrm{d}\tau} \mathrm{d}s.$$

This expression is increasing in $\int_t^s \nu_\tau^{-i} \mathrm{d}\tau$ and equal to

$$1 - \gamma e^{\ell_t^i} + \frac{\gamma e^{\ell_t^i}}{1+\mu},$$

for $\nu^{-i} \equiv I - 1$. By the definition of ℓ^{**} then, $\mathcal{C}_t^i$ is strictly decreasing in t whenever $\ell_t^i > \ell^{**}$.

We now show that in any equilibrium, at least one player must switch immediately if $\ell^0 < \ell^{**}$. Toward a contradiction, suppose that the first player i to switch to the risky arm does so at $t > 0$. Player i's cost must therefore have a local minimum at t. Because the time derivative $\mathcal{C}_t^{i'} = 0$, and all other players are setting $u^j = 1$, we have $\nu_t^{-i} = I - 1$ and hence

$$\mathcal{C}_t^{i''} e^{\mu t} = \gamma \mu e^{\ell_t^i} - (\mu + 1) < 0,$$

by definition of ℓ^{**}. It follows that for small enough $\varepsilon > 0$, $\mathcal{C}_{t-\varepsilon}^i < \mathcal{C}_t^i$, a contradiction.

Proof of Lemma 4 See Keller and Rady (2015, Proposition 1) for the cooperative solution u^{FB}. Note that if $\ell > \ell^{FB}$, $u^{FB}(\ell) \le u'(\ell) \le u''(\ell)$ implies that $u^{FB}(\ell) = u'(\ell) = u''(\ell) = I$ and costs are the same under all three policies. Hence, without loss, we assume $\ell^0 < \ell^{FB}$. Given some measurable $\underline{U}, \bar{U} : (-\infty, \ell^0] \to [0, I)$, with $0 \le \underline{U}(\ell) < \bar{U}(\ell)$ and $\bar{U}$ bounded away from I, consider the program $\mathcal{P}^{FB}(\underline{u})$:

$$\min \int_t e^{-\mu t} \left(\mu \ell_t - \gamma (I - 1 + \mu) e^{\ell_t} \right) \mathrm{d}t$$

over all $\pi : \mathbf{R}_+ \to [0, I]$, measurable, subject to

$$\dot{\ell}_t = u_t - I, \quad \ell_0 = \ell^0,$$

with, for all $t \ge 0$ and $\ell_t \le \ell^0$, $u_t \in [\underline{U}(\ell_t), \bar{U}(\ell_t)]$. By standard arguments, the optimal u is measurable with respect to the belief ℓ, and is the solution to the program

$$\min \int_\ell e^{-\mu t(\ell)} \left(\mu \ell - \gamma (I - 1 + \mu) e^{\ell} \right) \mathrm{d}\ell,$$

over all measurable $u : (-\infty, \ell^0] \to [0, I]$ such that $u(\ell) \in [\underline{U}(\ell), \bar{U}(\ell)]$, where $t(\ell)$ solves $t(\ell^0) = 0$ and

$$t'(\ell) = (u(\ell) - I)^{-1},$$

which is well defined because $u(\ell) < \bar{U}(\ell) < I$. A routine application of the maximum principle (Theorem 4.2, Cesari, 1983) yields that the optimal policy solves $u(\ell) = \underline{U}(\ell)$ a.e. Given u', u'' as stated in the lemma, the result follows if $u'' < I$ by setting $\underline{U} = u'$, $\bar{U} = u''$ and noting that u'' does not satisfy the necessary conditions. The same argument applies with trivial modifications if $\bar{U} = I$.

Proof of Lemma 5 Suppose that players $j \neq i$ stop at some fixed time $T \in \mathbf{R}_+$. For clarity, we use I rather 2 for the number of players, as the arguments do not depend on it (though the statement of Lemma 5 is specialized to that case). Throughout, we assume that $\ell^0 \in [\ell^*, \ell^{**}]$, as the result is trivial otherwise. Then, inserting into the objective of player i, he chooses τ to minimize

$$\int_{t\leq T} e^{-rt}\left(\mu(\ell^0+\lambda(t\wedge\tau-t))-\gamma\mu e^{\ell^0+l(t\wedge\tau-t)}\right)\mathrm{d}t$$
$$+\int_{t\geq T} e^{-rt}\left(\mu(\ell^0+\lambda(t\wedge\tau+(I-1)T-It))-\gamma(I-1+\mu)e^{\ell^0+\lambda(t\wedge\tau+(I-1)T-It)}\right)\mathrm{d}t.$$

This gives two expressions for the cost depending on $\tau \gtrless T$. Let us write $\mathcal{C}^1$ for the cost when $\tau \leq T$, and $\mathcal{C}^2$ for $\tau \geq T$ (the costs coincide when $\tau = T$). It is useful to use $x = \lambda\tau$ and $y = \lambda T$, instead of (τ, T). We explicitly compute both costs, which gives

$$\mathcal{C}^1(x) = -\frac{\gamma(I-1)e^{\ell^0+x-(\mu+1)y}}{(\mu+1)(I+\mu)} - \frac{(I-1)e^{-\mu y}}{\mu} - \gamma e^{\ell^0} + \frac{\gamma e^{\ell^0-\mu x}}{\mu+1} + l - \frac{e^{-\mu x}}{\mu},$$

and

$$\mathcal{C}^2(x) = \frac{e^{-(I+\mu)x-(\mu+1)y}\left(\gamma\mu e^{\ell^0+\mu y}\left(e^{Iy+x}-(I+\mu)e^{x(I+\mu)+y}\right)-(I+\mu)e^{Ix+y}\left((I-1)e^{\mu x}-\mu\ell^0 e^{\mu(x+y)}+e^{\mu y}\right)\right)}{\mu(I+\mu)}.$$

It is readily checked that $\mathcal{C}^1$ is concave, and so minimized either at $x = 0$ or $x = y$, while $\mathcal{C}^2$ is convex, and minimized at

$$x^* := y + \frac{\ell^0 - \ell^*}{I-1}.$$

Hence, we have only two candidates as global minimizer of the total cost, namely 0 and x^*. Note that (as shown in Fig. 2) the candidate minimizer x^* (resp., τ) is affine in y (resp., T). We compute the difference $\Delta := \mathcal{C}^2(x^*) - \mathcal{C}^1(0)$. Computing,

$$\Delta(y) := \gamma e^{\ell^0}\left(\frac{(I-1)e^{-(\mu+1)y}}{(\mu+1)(I+\mu)} - 1\right) + \frac{1 - \frac{(I-1)\left(\frac{\gamma(I+\mu-1)}{I+\mu}\right)^{\frac{\mu}{1-I}} e^{\mu\left(-\frac{\ell^0}{I-1}-y\right)}}{I+\mu-1}}{\mu} + \frac{\gamma\mu e^{\ell^0}}{\mu+1}.$$

We claim that $\Delta(y) < 0$ if and only if $y \leq \hat{y}$, for some $\hat{y} \geq 0$, and this will establish the result. First,

$$\lim_{y\to\infty}\Delta(y) = \frac{1}{\mu} - \gamma\frac{e^{\ell^0}}{1+\mu} > 0,$$

as $\ell^0 < \ell^{**}$. Second, $\Delta(0)$, viewed as a function of e^{ℓ^0}, is concave, zero at ℓ^*, with zero derivative at ℓ^*. Hence, $\Delta(0) \leq 0$ for all $\ell^0 \in [\ell^*, \ell^{**}]$ (the inequality being strict for $\ell^0 > \ell^*$). Finally, with the change of variable $Y = e^{-(1+\mu)y}$, we get that Δ is convex in Y, and hence admits at most one root Y, hence y.

Proof of Lemma 6 If any player j uses a pure policy in equilibrium, it must be $t^j = \hat{t}$ so that, by the best-reply analysis in Lemma 5, player i is indifferent between $t^i = 0$ and $t^i = \hat{t} + t^*$, where $t^* := \ell^0 - \ell^*$.

Lemma 5 further establishes that the best reply to $t^i = \hat{t} + t^*$ is $t^j = 0$ and that the best reply to $t^i = 0$ is $t^j = t^*$. We shall show that $t^* < \hat{t}$, so that $t^j = \hat{t}$ cannot be a best reply to any randomization over $t^i \in \{0, \hat{t} + t^*\}$.

It suffices to establish that the best reply to $t = t^*$ is, in fact, $\tau = 2t^*$. To do so, consider player i's marginal cost $\partial \mathcal{C}^i / \partial t^i$ evaluated $t^i = 0$ when player j uses $t^j = t^*$. This is proportional to

$$\left(\frac{\mu+2}{\gamma\mu+\gamma}\right)^{\mu}\left(-e^{-\mu\ell^0}\right) + \mu\left(\mu - \gamma(\mu+1)e^{\ell^0} + 2\right) + 1. \tag{15}$$

We want to show this expression is negative, so that switching to the risky arm later than $t = 0$ yields strict cost savings (hence that the best reply must be $2t^*$). Consider the derivative of the marginal cost with respect to ℓ^0. This is given by

$$\mu\left(\frac{\mu+2}{\gamma\mu+\gamma}\right)^{\mu} - \gamma\mu(\mu+1)e^{\ell^0(1+\mu)}.$$

This expression is strictly decreasing in ℓ^0 and negative (it is equal to $-\mu(1+\mu)$) when evaluated at $\ell^0 = \ell^*$. Therefore $\partial^2 \mathcal{C}^i / \partial t^i \partial \ell^0 < 0$ for all ℓ^0. To sign the marginal cost $\partial \mathcal{C}^i / \partial t^i$ evaluated at $t^i = 0$, it is sufficient to show that it is non-positive when $\ell^0 = \ell^*$. This is indeed the case, as the expression in (15) can be easily verified to be nil for $\ell^0 = \ell^*$.

Appendix C. Proofs for Section 5

Proof of Theorem 1 We first argue that in every symmetric equilibrium the support of the distribution is an interval: for all i, $\operatorname{supp} F^i = [\underline{t}, \bar{t}]$, for some $\underline{t} \leq \bar{t}$, with $\ell_{\bar{t}} = \ell^*$.

Using the same notation as in the proof of Lemma 2, let $\chi_t = \int_{s=0}^{t} (\nu_s^{-i} - I)\mathrm{d}s$. By stopping at time t, starting at time 0 with a "belief" ℓ, player i's cost is equal to (integrating (14) by parts)

$$\ell - \gamma e^{\ell} + \int_0^{\infty} e^{-\mu s}\mu\chi_s \mathrm{d}s + \frac{1 - e^{-\mu t}}{\mu} + \gamma \int_t^{\infty} e^{\ell + \chi_s - \mu s + t}\mathrm{d}s, \tag{16}$$

which is differentiable in t. If $t \in \operatorname{supp} F^i$, it must be that the derivative with respect to t be zero, that is,

$$e^{-\mu t}\left(1 - \gamma e^{\ell + \chi_t + t}\right) + \gamma \int_t^{\infty} e^{\ell + \chi_s - \mu s + t}\mathrm{d}s = 0. \tag{17}$$

Furthermore, this expression being itself differentiable in t, the second derivative must be non-negative, which is equivalent to (differentiating and using the first-order condition)

$$\gamma(I - 1 - \nu_t^{-i} + \mu) - (1+\mu)e^{-\ell_t} \geq 0. \tag{18}$$

Note that the left-hand side of (18) is decreasing in t if $t \notin \cup_{j \neq i} \operatorname{supp} F^j$. Hence, if $t_1, t_2 \in \operatorname{supp} F^i$, with $t_1 < t_2$, it must be that $(t_1, t_2) \cap \operatorname{supp} F^j \neq \emptyset$ for at least one $j \neq i$. Otherwise, (16) must admit a local maximum at some $t \in (t_1, t_2)$, at which value the inequality of (18) is reversed. This is inconsistent with the monotonicity of the left-hand side of (18) over (t_1, t_2), and

the fact that it is positive as either $t \downarrow t_1$ or $t \uparrow t_2$. Because we focus on symmetric equilibria, this implies that, for any $t_1, t_2 \in \operatorname{supp} F^i$, $t_1 < t_2$, there exists $t \in (t_1, t_2)$ such that $t \in \operatorname{supp} F^i$. Hence, the support of F^i (a closed set by definition) must be an interval, and by Lemma 2, we must have $\ell_{\bar{t}} = \ell^*$.

Because no pure-policy equilibrium exists, we know $\bar{t} > \underline{t}$. Assume for the time being that $\underline{t} = 0$ (we show later that $\underline{t} > 0$ cannot occur). Because the cost from stopping must be constant over $[0, \bar{t}]$, the second derivative given by (18) must be identically zero over $(0, \bar{t})$. Inequality (18) immediately gives ν_t^{-i} as a function of ℓ_t. Because ℓ is differentiable, so must ν^{-i} be. Hence, defining $\xi_t^{-i} = (I - 1 - \nu_t^{-i})/\mu$ and differentiating (18) (eliminating e^{ℓ_t} by using (18)) gives that ξ^{-i} obeys the differential equation

$$\dot{\xi}_t^{-i} = \mu \xi_t^{-i}(1 + \xi_t^{-i}),$$

and so $\xi_t^{-i} = (A_1 e^{-\mu t} - 1)^{-1}$ for some $A_1 > 0$ (because $\xi_t^{-i} > 0$), yielding

$$\nu_t^{-i} = I - 1 + \frac{\mu}{1 - A_1 e^{-\mu t}}, \tag{19}$$

for all $t \in (0, \bar{t})$. Hence,

$$\ln \mathbf{E}_{tj}[e^{\int_0^t u_s^j \mathrm{d}s}] = \frac{1}{I-1} \int \left(I - 1 + \frac{\mu}{1 - A_1 e^{-\mu s}} \right) \mathrm{d}s = \frac{\ln(A_1 - e^{\mu t})}{I-1} + t + A_2,$$

for some $A_2 \in \mathbf{R}$. That is,

$$\int_{s=0}^{t} e^s \mathrm{d}F(s) + (1 - F(t))e^t = e^{A_2} e^{\frac{1}{I-1}(\ln(A_1 - e^{\mu t}) + (I-1)t)}$$

$$= e^{A_2}(A_1 - e^{\mu t})^{\frac{1}{I-1}} e^t.$$

Differentiating both sides gives finally

$$1 - F(t) = \frac{e^{A_2}}{I-1}(A_1 - e^{\mu t})^{\frac{1}{I-1}} e^t \left(I - 1 - \frac{\mu}{A_1 e^{-\mu t} - 1} \right). \tag{20}$$

It remains to determine the constants A_1, A_2.

If $\ell^0 < \ell^{**}$, combine (18) (with equality) at $t = 0$ with (19) to get

$$A_1 = \left(1 - \frac{\mu}{1+\mu} \gamma e^{\ell^0} \right)^{-1}.$$

Moreover, note from (6) that $1 - F(0) = \nu_0^{-i}/(I-1)$. Plugging in (20) for $t = 0$ using (19) gives $A_2 = (A_1 - 1)^{-\frac{1}{I-1}}$. The resulting distribution is given by

$$\bar{F}(t) = \left(\frac{A_1 - e^{\mu t}}{A_1 - 1} \right)^{\frac{1}{I-1}} \left(1 - \frac{\mu}{(I-1)(A_1 e^{-\mu t} - 1)} \right). \tag{21}$$

Let us make a few final remarks. First, note that the complementary distribution function is equal to 1 at $\ell^0 = \ell^{**}$. That is, if the game starts with this belief, it never changes and the safe arm is pulled throughout. We must now rule out that $\underline{t} > 0$ for this special case. If $\ell^0 = \ell^{**}$, there is nothing to show (as the safe arm is pulled forever anyhow). If $\ell^0 > \ell^{**}$, the safe arm must be pulled throughout (the support of the distribution of stopping beliefs must be convex, yet the

cost is strictly quasi-convex in t for $\ell^0 > \ell^{**}$, yielding a contradiction if this region included a stopping time). Now suppose $\ell^0 < \ell^{**}$ and $\underline{t} > 0$. Given Lemma 1, the only potentially profitable deviations are stopping policies π_t^i with $t < \underline{t}$. Note that, given that players $j \neq i$ use the stopping policy F^j, it holds that $\nu_t^{-i} = I - 1$ for all $t < \underline{t}$. Hence, a necessary condition for player i to follow the equilibrium policy is that his cost be convex at $t = \bar{t}$. Note that the value of (18) at $t = \underline{t}$ is

$$\gamma(I - 1 + \mu - \nu_{\underline{t}}^{-i}) - (1 + \mu)e^{-\ell_{\underline{t}}} = \gamma\mu - (1 + \mu)e^{-\ell^0}, \tag{22}$$

which, using the definition of ℓ^{**}, is negative. Because player i's cost is constant over $(\underline{t}, \bar{t})$, we conclude that deviating to pulling the risky arm at time $\underline{t} - \varepsilon$ would be a profitable deviation for $\varepsilon > 0$ small enough.

Proof of Theorem 2 As mentioned, the proof of this theorem is rather tedious, and the interested reader might want to consult both the supplementary materials file and a Mathematica file with some of the omitted algebraic operations, available on the authors' websites (entitled `supplementary.pdf` and `theorem2proof.nb`).

The logic of the argument is as follows. Suppose another equilibrium exists. Because on any interval over which a player's opponent does not switch with positive probability, a player's cost is convex, there is at most one time during such an interval at which he is willing to switch. Because of Lemma 6, we know that each player's equilibrium policy must include in its support at least two switching times. If the support of a player's policy is a dense subset of some interval, then so must be his opponent's (because of convexity, as explained), and continuity of the cost function then implies that this support is precisely $[0, \bar{\tau}]$, as defined in Theorem 1, and the equilibrium is the one described there. Hence, we might assume that there exists at least two times t_1, t_3, with $0 < t_1 < t_3$, such that, say, player 1's policy assigns positive probability of switching at times t_1 and t_3, and at no time in between. This however implies (convexity again) that there is some time $t_2 \in (t_1, t_3)$ and some time $t_0 < t_1$ such that player 2 is willing to switch at time t_0 and t_2, but at no time in between (and 1 does not switch at any time in (t_0, t_1) either).[33] We then derive a contradiction, showing that independently of how players behave at times not in $[t_0, t_4]$, the necessary (first- and second-order) conditions cannot hold simultaneously at those four dates. See `supplementary.pdf` for the details.

Proof of Proposition 1 (1.) Fix the number of players I and use equation (6) to write the stopping-time distribution in terms of the function ν^{-i}. The symmetric equilibrium stopping-time distribution F_t^I is then given by

$$F_t^I = 1 - \frac{\nu_t^{-i}}{I - 1} e^{\int_0^t \frac{\nu_s^{-i}}{I-1} \mathrm{d}s - t},$$

where ν_t^{-i} solves equation (19). Note that, for a given t, the term $\nu_t^{-i}/(I - 1)$ is increasing in I. The exponential term is equal to

$$\left(\frac{e^{-\ell^0} \left(1 + \mu + e^{\mu t} \left(\mu \left(\gamma e^{\ell^0} - 1\right) - 1\right)\right)}{\gamma\mu} \right)^{\frac{1}{I-1}},$$

[33] More precisely, either there is such a $t_0 < t_1$, or a $t_4 > t_3$ in the support of 2's policy, but relabeling the players if necessary, we may as well assume it is $t_0 < t_1$.

hence it is smaller than one and increasing in I. Therefore, the partial derivative of F_t^I with respect to I is positive for all $t < \bar{\tau}$. In addition, $1 - F_0^I = \nu_0^{-i}/(I-1)$ which is increasing in I. Therefore the distributions F_t^I are ranked by first-order dominance.
(2.) From the outside observer's perspective,

$$\nu_t^I = \frac{I}{I-1}\left(\mu - \frac{\mu(\mu+1)}{\mu + e^{\mu t}\left(\mu\left(\gamma e^{\ell^0}-1\right)-1\right)+1}\right) + I.$$

Notice that the first term is negative (as $\nu_t^I \leq I$). This implies ν_t^I is increasing in I.
(3.) The speed of learning of the outside observer is

$$-\dot{\ell}_t^I = I - \nu_t^I,$$

which is decreasing by inspection of ν_t^I. Therefore, during the mixing phase, beliefs decrease faster with a *lower* number of players. Furthermore, as I increases, the length of the mixing phase increases. However, for $t > \bar{t}$, beliefs decrease at rate I, which implies faster learning for a higher number of players. Therefore, the outside observer's belief trajectories for $I' > I$ cross once at a time $t > \bar{t}_{I'}$.
(4.) Straightforward computations of the total cost yield expression (10) in the text. This cost is constant for any $I \geq 2$ and (because of positive informational externalities) strictly lower than the single-agent cost.

Appendix D. Proofs for Section 6

Proof of Proposition 2 The second inequality of the proposition ($\nu^o(p) \geq \nu^{fb}(p)$) being immediate given that $\bar{p} < p^{FB}$, it is the first inequality that must be established. Given ℓ^0 and $\ell < \ell^0$, we let $t(\ell)$ denote the time at which the belief of the outside observer reaches belief ℓ. The outsider's belief at time t satisfies

$$\dot{\ell}_t = -(I - \nu_t^I), \quad \ell_0 = \ell^0.$$

Now suppose towards a contradiction that there exists a "belief" $\hat{\ell}$ such that the outside observer's hazard rate in the unobservable case $\nu^n(\hat{\ell})$ is equal to the hazard rate in the observable case $\nu^o(\hat{\ell})$. We derive an ordinary differential equation for $\nu^{-i}(\ell) := (I-1)\nu_{t(\ell)}^I/I$ in both cases.

In the unobservable case, we know from the proof of Theorem 1 that

$$\nu_t^{-i} = -1 + I + \frac{\mu}{1 + \frac{e^{-\mu t}(1+\mu)}{e^{\ell^0}\gamma\mu - 1 - \mu}}.$$

Differentiating ν_t^{-i} with respect to t, we obtain

$$\frac{\mathrm{d}\nu_t^{-i}}{\mathrm{d}t} = \frac{\mu^2(\mu+1)e^{\mu t}\left(\mu\left(\gamma e^{\ell^0}-1\right)-1\right)}{\left(e^{\mu t}\left(\mu\left(\gamma e^{\ell^0}-1\right)-1\right)+\mu+1\right)^2}.$$

Solving for $e^{\mu t}$ from the definition of ν_t^{-i} and plugging back into the derivative, we obtain

$$\frac{\mathrm{d}\nu^{-i}(\ell)}{\mathrm{d}\ell} = -(-1+I-\nu^{-i}(\ell))(-1+\mu+I-\nu^{-i}(\ell))t'(\ell),$$

where

$$t'(\ell) = \frac{1}{\frac{I}{I-1}\nu^{-i}(\ell) - I}.$$

Finally, we obtain the derivative

$$\frac{\mathrm{d}\nu^{-i}(\ell)}{\mathrm{d}\ell} = (\mu + I - \nu^{-i}(\ell) - 1)\frac{I - \nu^{-i}(\ell) - 1}{I - \nu^{-i}(\ell) - \frac{\nu^{-i}(\ell)}{I-1}}. \tag{23}$$

Note that $\nu^{-i}(\ell)$ is increasing in ℓ, as expected. Also notice that the second term in (23) is smaller than one, because $\nu^{-i} \leq I - 1$.

In the observable case, we already have the expression for the hazard rate

$$\nu^{-i}(\ell) = \mu + I - 1 - \frac{1 + (\ell - \ell^*)\mu}{e^{\ell}\gamma - 1}.$$

Differentiating with respect to ℓ and replacing e^{ℓ} with the solution to the previous equation, we obtain the following differential equation

$$\frac{\mathrm{d}\nu^{-i}(\ell)}{\mathrm{d}\ell} = (\mu + I - \nu^{-i}(\ell) - 1)\frac{I - \nu^{-i}(\ell) + \mu(\ell - \ell^*)}{1 + \mu(\ell - \ell^*)}. \tag{24}$$

Notice that $I - \nu^{-i} > 1$, and therefore the ratio in (24) is larger than one. Furthermore, the first term $(\mu + I - \nu^{-i} - 1)$ is identical in the two expressions (23) and (24). Thus, if the two paths $\nu^o(\ell)$ and $\nu^n(\ell)$ cross, the observable path must be steeper. This yields a contradiction, because

$$\nu^o(\ell^{**}) < \nu^n(\ell^{**}) = I - 1,$$

and therefore if the paths $\nu(\ell)$ cross, the unobservable path must be steeper at the crossing point closest to ℓ^{**}.

References

Akcigit, U., Liu, Q., 2015. The role of information in innovation and competition. J. Eur. Econ. Assoc. 14, 828–870.

Aumann, R.J., 1964. Mixed and behavior policies in infinite extensive games. In: Dresher, M., Shapley, L.S., Tucker, A.W. (Eds.), Advances in Game Theory. In: Annals of Mathematics Studies, vol. 52. Princeton University Press, Princeton, pp. 627–650.

Bloom, N., Van Reenen, J., 2007. Measuring and explaining management practices across firms and countries. Q. J. Econ. 122, 1351–1408.

Blume, A., 2003. Bertrand without fudge. Econ. Lett. 78, 167–168.

Board, S., Meyer-ter-Vehn, M., 2014. A reputational theory of firm dynamics. working paper UCLA.

Bolton, P., Harris, C., 1999. Strategic experimentation. Econometrica 67, 349–374.

Bonatti, A., Hörner, J., 2011. Collaborating. Am. Econ. Rev. 101, 632–663.

Bonatti, A., Hörner, J., 2015. Learning to disagree in a game of experimentation. working paper Cowles Foundation for Research in Economics.

Cesari, L., 1983. Optimization: Theory and Applications. Problems with Ordinary Differential Equations. Applications of Mathematics, vol. 17. Springer-Verlag, Heidelberg.

Coleman, J.S., Katz, E., Menzel, H., 1966. Medical Innovation: A Diffusion Study. Bobbs–Merrill Company, Indianapolis.

Covert, T., 2015. Experiential and social learning in firms: the case of hydraulic fracturing in the Bakken Shale. working paper University of Chicago.

Dasgupta, P., Maskin, E., 1986. The existence of equilibrium in discontinuous economic games, I: theory. Rev. Econ. Stud. 53, 1–26.

Décamps, J.-P., Mariotti, T., 2004. Investment timing and learning externalities. J. Econ. Theory 118, 80–102.

Gibbons, R., Henderson, R., 2013. What do managers do? Exploring persistent performance differences among seemingly similar enterprises. In: Handbook of Organizational Economics. Princeton University Press.
Karlin, S., 1959. Mathematical Methods and Theory in Games, Programming and Economics. Addison–Wesley, Reading, Mass., London.
Keller, G., Rady, S., 2003. Price dispersion and learning in a dynamic differentiated-goods duopoly. Rand J. Econ. 34, 138–165.
Keller, G., Rady, S., 2015. Breakdowns. Theor. Econ. 10, 175–202.
Keller, G., Rady, S., Cripps, M., 2005. Strategic experimentation with exponential bandits. Econometrica 73, 39–68.
Mansfield, E., 1961. Technical change and the rate of imitation. Econometrica 29, 741–766.
Milgrom, P.R., Weber, R.J., 1985. Distributional policies for games with incomplete information. Math. Oper. Res. 10, 619–632.
Murto, P., Välimäki, J., 2011. Learning and information aggregation in an exit game. Rev. Econ. Stud. 78, 1426–1461.
Nelsen, R.B., 2006. An Introduction to Copulas, 2nd ed.. Springer Series in Statistics. Springer-Verlag, Heidelberg.
Presman, E.L., Sonin, I.N., 1990. Sequential control with incomplete information. In: Economic Theory, Econometrics, and Mathematical Economics. Academic Press, San Diego, CA.
Rosenberg, D., Salomon, A., Vieille, N., 2013. On games of strategic experimentation. Games Econ. Behav. 82, 31–51.
Rosenberg, D., Solan, E., Vieille, N., 2007. Social learning in one-arm bandit problems. Econometrica 75, 1591–1611.
Seierstad, A., Sydsæter, K., 1987. Optimal Control Theory with Economic Applications. North-Holland, Amsterdam.
Shmaya, E., Solan, E., 2014. Equivalence between random stopping times in continuous time. working paper Kellogg School of Management.
Skinner, J., Staiger, D., 2007. Technology adoption from hybrid corn to beta-blockers. In: Hard-to-Measure Goods and Services: Essays in Honor of Zvi Griliches. University of Chicago Press.
Syverson, C., 2011. What determines productivity? J. Econ. Lit. 49, 326–365.
Touzi, N., Vieille, N., 2002. Continuous-time Dynkin games with mixed policies. SIAM J. Control Optim. 41, 1073–1088.
Vives, X., 1999. Oligopoly Pricing, Old Ideas and New Tools. MIT Press, Cambridge, MA.
Weizsäcker, H., 1974. Zur Gleichwertigkeit zweier Arten der Randomisierung. Manuscr. Math. 11, 91–94.
Yushkevich, A.A., 1988. On the two-armed bandit problem with continuous time parameter and discounted rewards. Stochastics 23, 299–310.