Learning, Experimentation and Information Design

Johannes Hörner and Andrzej Skrzypacz*

August 4, 2016

1 Introduction

The purpose of this paper is to survey recent developments in a literature that combines ideas from experimentation, learning, and strategic interactions. Because this literature is multifaceted, let us start by circumscribing our overview. First and foremost, all surveyed papers involve nontrivial dynamics. Second, we will restrict attention to models that deal with uncertainty. Models of pure moral hazard, in particular, will not be covered. Third, we exclude papers that focus on monetary transfers. Our goal is to understand incentives via other channels –information in particular, but also delegation. Fourth, we focus on strategic and agency problems, and so leave out papers whose scope is decision-theoretic. However, rules are there to be broken, and we will briefly discuss some papers that deal with one-player problems, to the extent that they are closely related to the issues at hand. Finally, we restrict attention to papers that are relatively recent (specifically, we have chosen to start with Bolton and Harris, 1999).

Our survey is divided as follows. First, we start with models of strategic experimentation. These are abstract models with few direct economic applications, but they develop ideas and techniques that percolate through the literature. In these models, players are (usually) symmetric and externalities are (mostly) informational.

Moving beyond the exploitation/exploration trade-off, we then turn to agency models that introduce a third dimension: motivation. Experimentation must be incentivized. The first way this can be done (Section 3) is via the information that is being disclosed to the agent performing the experimentation, by a principal who knows more or sees more. A second way this can be done is via control. The nascent literature on delegation in dynamic environments is the subject of Section 4.

Section 5 turns to models in which information disclosure is not simply about inducing experimentation, but manipulating the agent's action in broader contexts. To abstract from experimentation altogether, we assume that the principal knows all there is to know, so that only the agent faces uncertainty.

Finally, Section 6 discusses experimentation with more than two arms (Callander, 2011).

*This survey was prepared for the Econometric Summer Meetings in Montréal, 2015. We thank Kostas Bimpikis, Alessandro Bonatti, Daria Khromenkova, Nicolas Klein, Erik Madsen and Chiara Margaria for detailed comments.

2 Equilibrium Interactions

2.1 Strategic Bandits

Strategic bandit models are game-theoretic versions of standard bandit models. While the standard "multi-armed bandit" describes a hypothetical experiment in which a player faces several slot machines ("one-armed bandits") with potentially different expected payouts, a strategic bandit involves several players facing (usually, identical) copies of the same slot machine. Players want to stick with the slot machine if and only if the best payout rate makes it worth their time, and learn not only from their own outcomes but also from their neighbors.

Equilibrium strategies are not characterized by simple cut-offs in terms of the common belief. As a result, solutions to strategic bandits are only known for a limited class of distributions, involving two states of the world only. In Bolton and Harris (1999, BH), the observation process (of payoffs) follows a Brownian motion, whose drift depends on the state. In Keller, Rady and Cripps (2005, KRC), it follows a simple Poisson process, with positive lump-sums ("breakthroughs") occurring at random (exponentially distributed) times if and only if the arm is good. Keller and Rady (2015, KR15) solve the polar opposite case in which costly lump-sums ("breakdowns") occur at random times if and only if the arm is bad. Keller and Rady (2010, KR10) consider the case in which breakthroughs need not be conclusive.

These models share in common, in addition to the binary state framework, their focus on symmetric Markov perfect equilibria (MPE).[1] Throughout, players are Bayesian. Given that they observe all actions and outcomes, they share a common belief about the state, which serves as the state variable. They are also impatient and share a common discount rate.

BH and KR10 are the most ambitious models and offer no closed-form solutions for the equilibrium. Remarkably, however, they are able to prove uniqueness and tease out not only their structure, but their dependence on parameters. While it is BH that first develops both the ideas (including the concepts of free-riding and encouragement effects) as well as the methods used throughout this literature, the most interesting insights can already be gleaned from the simple exponential bandits in KRC and KR15. What makes these two models tractable is that these models can be viewed as deterministic: unless a breakdown or breakthrough ("news") occurs, the (conditional) posterior belief follows a known path. If news arrives, the game is over, since if it is commonly known that the state is good (or bad), informational externalities cease to matter, and each player knows the strictly dominant action to take.

Let us consider here a simplified version combining the good and bad news models. The state is $\omega \in \{G, B\}$. Each player $i = 1, \ldots, I$ controls the variable $u_t^i \in [0, 1]$, which is the fraction allocated to the risky arm at time $t \geq 0$ (the complementary fraction being allocated to the safe arm). The horizon is infinite. This leads to a total realized payoff of

$$\int_0^\infty e^{-rt}(h\mathrm{d}N_{G,t}^i - \ell \mathrm{d}N_{B,t}^i),$$

where $r > 0$ is the discount rate, $h, \ell > 0$ are the value and cost of a breakthrough or breakdown, and $N_{G,t}^i, N_{B,t}^i$ are Poisson processes with intensities

$$u_t^i \lambda_G \mathbf{1}_{\{\omega=G\}}, \text{ and } u_t^i \lambda_B \mathbf{1}_{\{\omega=B\}},$$

which are conditionally independent across players. Here, $\lambda_G, \lambda_B \geq 0$ are parameters. News is conclusive: if any player experiences a breakthrough (breakdown), all players immediately learn

[1] KRC and KR15 discuss asymmetric MPE as well, although this remains an open and challenging problem for the Brownian case.

that the state is good (bad) and allocate weight 0 (1) to the risky arm, the alternative having value normalized to 0.

The main distinction in these models hinges on the direction in which updating occurs in the absence of any news. The (conditional) belief $p_t = \mathbf{P}[\omega = G]$ that the state is good evolves according to

$$\dot{p}_t = -\Delta p_t(1-p_t)\sum_{i=1}^{I} u_t^i, \text{ where } \Delta := \lambda_G - \lambda_B.$$

Hence, if $\lambda_G > \lambda_B$, then this belief drifts down over time. We call such a model a good news model (because a special case is $\lambda_B = 0$, in which case the only news that might arrive is good). If $\lambda_G < \lambda_B$, the belief drifts up, unless news arrives. This is the bad news model.[2]

The first best, or *cooperative*, outcome is simple. Players should follow a common cut-off (Markov) policy, using the risky arm if and only if the belief p_t is above some threshold (say, $\bar{p}_I$ in the bad news case, and $\underline{p}_I$ in the good news one). Such behavior results in very different dynamics according to the news scenario. In the bad news case, experimentation goes on forever unless (bad) news occurs. In the good news model, experimentation stops unless (good) news occurs. This drastically changes equilibrium predictions.

To understand incentives, let us fix players $-i$'s strategies to what they would do, each on their own (that is, to be the the optimal single-agent policy). Let us start with the bad news case, in which this involves a threshold $\bar{p}_1$ above which they experiment (set $u^i = 1$). This threshold is below the myopic threshold p^m at which the expected reward from taking the risky action is precisely 0, namely,

$$\lambda_G p^m h = \lambda_B(1-p^m)\ell.$$

This is because the option value from learning via experimentation makes it worthwhile for a lone player to use the risky arm at beliefs slightly below p^m, when the flow loss from doing so is small enough.

Consider now the best-reply of a player to such a strategy by the others at beliefs in the neighborhood of this threshold $\bar{p}_1$. Right above $\bar{p}_1$, player i need no longer carry out costly experimentation on his own to learn about the state, because other players are experimenting. Hence, there is a range of beliefs right above $p > \bar{p}_1$ in which his best-reply is to use the safe arm only, depressing overall experimentation at such beliefs. This is the *free-riding effect*. See Figure 1. As a result, player i's payoff is boundedly higher than what it would be if he were on his own, given that the value of information is strictly positive.

Consider now beliefs right below $\bar{p}_1$. If player i does not experiment, nothing will happen, given that nobody else does. Yet, getting the belief "over the edge" $\bar{p}_1$ is very valuable, as it will kickstart experimentation by players $-i$. Because player i would be indifferent at $\bar{p}_1$ if he were on his own, without the added benefit provided by others' experimentation, he strictly prefers to do so if this is the prerequisite to get others started. Hence, there is a range of beliefs below $\bar{p}_1$ at which player i's best-reply is to experiment. This the *encouragement effect*. Absent free-riding, the encouragement effect is responsible for the monotonicity of the cooperative amount of experimentation to increase with the number of players (*e.g.*, $\bar{p}_I$ decreasing in I). Of course, for very low beliefs, playing safe is the best-reply.

Hence, the best-reply to the optimal (single-player) cut-off policy is not a cut-off policy. In fact, it is not monotone. With some work, however, it can be shown that a symmetric monotone (pure-strategy) equilibrium exists, but it involves interior levels of experimentation ($u^i \in (0,1)$) for some beliefs.

[2]The special case $\lambda_G = \lambda_B$ can be nested in either case.

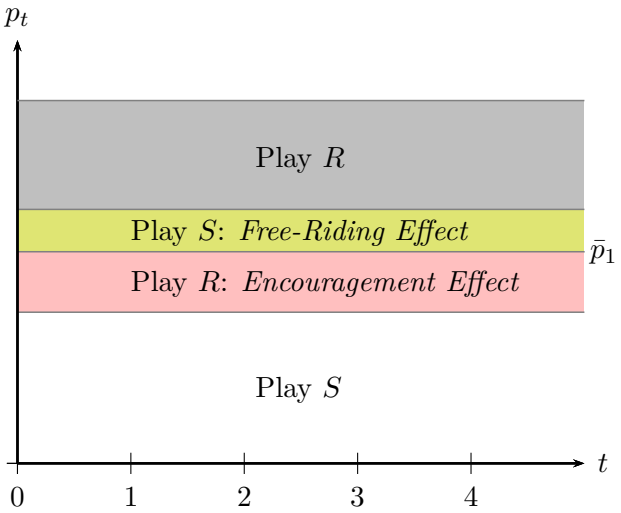


Figure 1: Best-reply to the one-player optimal policy under bad news.

In contrast, consider now the case of good news. Figure 2 illustrates the best-reply to the one-player optimal policy, which specifies play of the risky arm if and only if $p \geq \underline{p}_1$. Beliefs are now partitioned into only three intervals. There is no encouragement region for beliefs below $\underline{p}_1$. This is because of the dynamics of beliefs. Here, experimentation at such beliefs does not move the common belief closer to the region at which other players would experiment (unless a breakthrough occurs, obviously). Because the conditional belief absent a breakthrough only drifts down further, a player is on his own, for beliefs below $\underline{p}_1$. Hence, by the definition of this threshold, it is optimal to play safe. This does not imply that the one-player optimal policy is an equilibrium, however, because at beliefs immediately above $\underline{p}_1$, an agent still has incentives to free-ride on others' experimentation. Because $\underline{p}_1$ is strictly below the myopic threshold at which playing risky already pays off in the next instant, playing risky at such beliefs is motivated by the option value from learning. Because other players perform this experimentation, it is then best to play safe oneself, at least for beliefs that are sufficiently close to (and above) $\underline{p}_1$.

Hence, here as well, the symmetric equilibrium involves interior levels of experimentation. In this sense, it looks similar to the symmetric equilibrium in the bad news case. But the absence of the encouragement region for beliefs below $\underline{p}_1$ implies that experimentation levels are lower than the socially efficient level, with experimentation ceasing altogether at this threshold independent of the number of players. In particular, independent of I, the total amount of experimentation performed by players over the infinite horizon (which the asymptotic belief precisely measures) is the same as if there was only one player. Delay is higher, as one can show that, unlike with one player, this belief is only asymptotically achieved, as experimentation rates dwindle down when the belief approaches the threshold. Players are better off, of course (they can always ignore the existence of other players, and replicate the one-player policy and payoff), because the cost of experimentation is shared among them. Nonetheless, the outcome is inefficient.

We now provide some details on how these symmetric equilibria can be solved for. Consider the Hamilton-Jacobi-Bellman (HJB) equation for player i's continuation value v, taking as given the aggregate experimentation by other players, $u^{-i}(p) = \sum_{j \neq i} u^j(p)$. It holds that

$$rv(p) = \max_{u^i} \{u^i \pi(p) + (u^i + u^{-i}(p))[\underbrace{(p\lambda_G(\lambda_G h/r - v(p)) - (1-p)\lambda_B v(p)))}_{\text{jump in value if news arrives}} - \underbrace{\Delta p(1-p)v'(p)}_{\substack{\text{drift in value} \\ \text{if no news arrives}}}]\},$$

where

$$\pi(p) = p\lambda_G h - (1-p)\lambda_B \ell$$

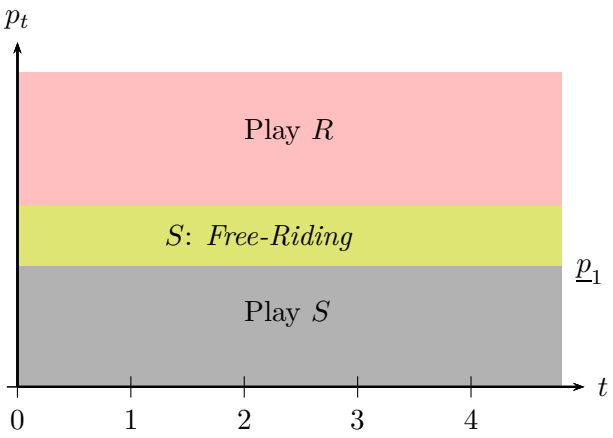


Figure 2: Best-reply to the one-player optimal policy under good news.

is the expected flow payoff from the risky arm. Over the next instant, $\pi(p)$ is collected by player i at rate u^i. Either one of the players experiences a breakthrough or breakdown, in which case the value jumps from $v(p)$ to either $\lambda_G h/r$ if a breakthrough occurs, or 0 if it is a breakdown. In the absence of news, the value changes at a rate proportional to the belief change $\dot{p}$. We see that the right-hand side is linear in u^i, implying that over a range of beliefs $[p_1, p_2]$ over which player i is indifferent, the coefficient on u^i must be zero, that is,

$$\pi(p) + (p\lambda_G(\lambda_G h/r - v(p)) - (1-p)\lambda_B v(p))) - \Delta p(1-p)v'(p) = 0, \tag{1}$$

a differential equation that can be solved explicitly, given the relevant boundary conditions. Adding and subtracting $u^{-i}(p)\pi(p)$ from the HJB equation (and using (1) to eliminate almost all terms) gives

$$rv(p) = -u^{-i}(p)\pi(p), \tag{2}$$

over such a range.[3]

In a symmetric equilibrium, $u^{-i}(p) = (I-1)u^i(p)$, and the optimal experimentation level follows, as explained next. Equations (1)–(2) are equally valid for good and bad news. The difference appears in the boundary conditions. In the case of bad news, the value drifts up. It is readily verified that u^i increases in p, and so p_2 is determined by $u^i(p_2) = 1$ and $v(p_2) = \bar{v}(p_2)$, which together with (2) imply $r\bar{v}(p_2) = -(I-1)\pi(p_2)$, where $\bar{v}$ is the per-player value from the cooperative solution discussed above. Once this threshold is reached, players achieve the first best. Because of the upward drift, this suffices to determine the candidate value $v(p)$ for all $p < p_2$. Because playing safe is always an option, $v(p) \geq 0$, and finding the solution to $v(p_1) = 0$ then determines p_1, below which indifference no longer holds and the safe arm is exclusively played.

[3]Because $v(p) \geq 0$, we see that such a range must be a subset of the beliefs over which π is negative. This should not come as a surprise because above the myopic threshold players certainly use the risky arm.

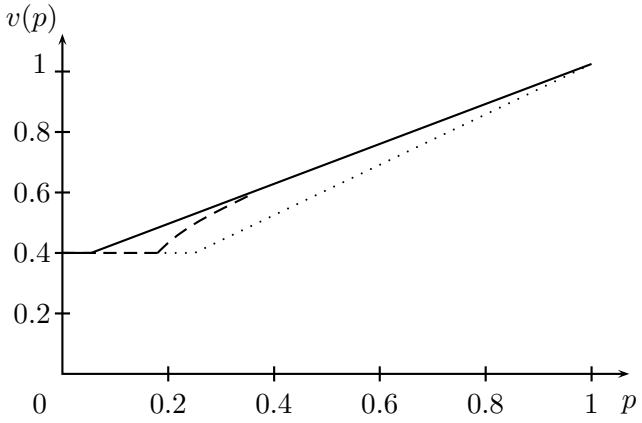


Figure 3: Value function with bad news.

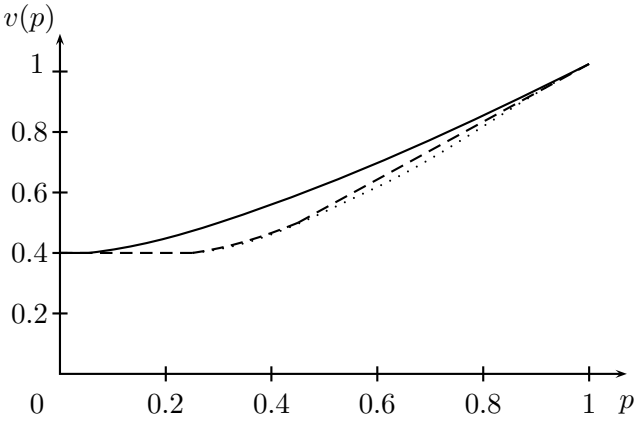


Figure 4: Value function with good news.

In contrast, with good news, the candidate value function v over the to-be-determined interval of beliefs $[p_1, p_2]$ over which experimentation levels are interior must be solved "bottom-up," starting from p_1. As argued, $p_1 = \underline{p}_1$, the level at which a lone player ceases experimentation. In addition, $v'(\underline{p}_1) = 0$ (smooth-pasting) must hold.[4] This suffices to solve (1) for the candidate value v, and p_2 is determined by $u^{-i}(p_2) = I - 1$. Hence, solving $v(p_2) = -r(I-1)\pi(p_2)$ gives p_2, above which indifference no longer holds and risky is played exclusively.

See Figure 3 and Figure 4 for an illustration of the value functions. The intermediate (dashed) curve is the symmetric equilibrium value function (as a function of the belief), and the lower (dotted) and higher (solid) curves are the single-player and cooperative value ($\bar{v}$) functions. Considering the left figure (bad news), the encouragement effect drives the larger range of beliefs at which players experiment (as can be shown), relative to the single-player case, yet the free-riding effect leads to an amount of experimentation that remains too low nonetheless. In the right figure (good news), the impact of free-riding has a more dramatic impact, as experimentation stops at the same threshold, as in the single-player case.

To what extent is the solution concept (symmetric MPE) driving these results? As KRC show, symmetry is partly responsible for the inefficiency, and equilibria with higher payoffs can be achieved when players take turns experimenting in a Markov equilibrium, for instance. More generally, Hörner, Klein and Rady (2015) show that the Markov restriction plays an important role in the inefficiency result, as well as the equilibrium structure. Strongly symmetric equilibrium (SSE), in which all players choose the same continuation strategy after a given history (independent of the identity of a potential deviator, for instance), is a restrictive solution concept in general, yet it is still weaker than symmetric MPE which are automatically SSEs. In the experimentation context, SSE turn out to allow for exactly efficient equilibria for a range of (but not all) parameters. This is the case, in particular, whenever news is bad (whether it is conclusive or not), or in the Brownian case. Furthermore, as far as the total surplus (or the players' average payoff) is concerned, it can be shown that, in all cases, no perfect Bayesian equilibrium outcome outperforms SSE.

Several extensions are worth mentioning, although space does not allow us to discuss them further. Klein and Rady (2011) solve the case of negatively correlated arms (a disease might be either bacterial or viral, but not both). Players might be able to communicate publicly, in the absence of direct monitoring of each others' experimentation, a case considered by Heidhues, Rady

[4]This derivative cannot be negative, since, for all $\epsilon > 0$, $v(\underline{p}_1 + \epsilon) \geq 0 = v(\underline{p}_1)$. It cannot be positive either. Otherwise, there is a solution to (1) with $v(\underline{p}_1 - \epsilon) = 0$ as a boundary condition, giving $v(\underline{p}_1) > 0$, a contradiction.

and Strack (2015). Players might be asymmetric (Das, 2015), etc. Nonetheless, open questions remain even for the benchmark models. While Cohen and Solan (2013) solve the problem of optimal experimentation *for a single player* for a reasonably general class of Lévy processes (not including the exponential bad news case, however), a similar analysis that would encompass both the Brownian and the exponential model remains to be done. Also, whereas in the exponential case, KRC (and KR15) were able to solve for the asymmetric MPE, it is unknown whether asymmetric MPE exist in the Brownian case.[5]

2.2 On the Role of Information

What if players do not observe the outcomes of the other players, but only what they do? And what if players don't observe the actions of the other players, but only the outcomes? It is easy to think of examples of either scenario: firms might observe each others' profits, but not techniques; co-authors see each others' output, not input. On the other hand, consumers see each others' choices, not derived utilities; and firms might see each others' supply chain without being able to evaluate their satisfaction with specific suppliers.

The first problem –observed actions, unobserved outcomes– remains largely unsolved. Rosenberg, Solan and Vieille (2007) provide a partial answer to it, under the *assumption* that switching to the safe arm is an irreversible choice, while also making clear that this assumption is restrictive: players would like to switch back to the risky arm under circumstances that occur with positive probability. The second problem –unobservable actions, observed outcomes– is better understood, at least in the case of exponential bandits. The answer further underscores the differences between good and bad news bandits.

Rosenberg, Solan and Vieille (2007) prove that, when the switch to the safe arm is irreversible, the optimal policy remains an index policy (a generalization of a threshold policy defined shortly) under weak distributional assumptions.

Consider the case of two players (the focus of their paper). By round n, assuming player 1 has not switched to the safe arm yet, two possibilities arise. Either player 2 has not either, or he has in round $k < n$. Not observing player 2 stop is a good sign about 2's observed payoffs. As a result, player 1 uses a simple cut-off rule: he compares his private belief, based on his own observations only, to a threshold that is non-increasing with n. If and only if his belief exceeds this threshold, he continues experimenting. If, instead, player 2 switched to the safe arm in round k, player 1 faces the standard decision problem of a one-armed bandit, for which the solution is an index policy (with a cut-off in terms of his belief). Of course, his belief accounts for the decision of player 2 to quit at time k. Whether an index policy remains optimal when the irreversibility assumption is dropped is an important open question.

Bonatti and Hörner (2011) is based on the good news model, but involves a payoff externality. As soon as a player experiences a breakthrough, the game ends, and every player derives the same utility from it. However, experimenting (effort) is costly and its cost is privately borne. Think of collaborators working on a common project from different locations. Effort is not observable, but the output is automatically reported on Dropbox –if someone cracks the problem, it is immediately common knowledge among players. Importantly, effort is private information. Free-riding is present, but deviations do not lead to changes in the opponents' beliefs, as such deviations are not observed. As a result, the unique symmetric equilibrium, which as in KRC involves interior effort, leads to higher effort provision than with observable effort. This is because unexpected, but observed, shirking leads to a belief that is higher than it would be otherwise, and this optimism leads to

[5] As mentioned above, there exist efficient SSE in the Brownian case.

more effort. Hence, shirking is more attractive when it is observed, as it leads other players to work harder in the future (relative to how much they would work if shirking was not observed).

What is the ultimate fate of such collaborations? As Bonatti and Hörner show, effort is scaled down over time, so that the asymptotic belief makes agents just indifferent between working and not. The project dwindles over time, but is never abandoned.[6] Gordon, Marlats and Ménager (2015) consider a variation in which there is a fixed delay $\Delta > 0$ between a player's breakthrough and his collaborator observing it. Remarkably, the unique symmetric MPE exhibits periodic behavior, with players alternating over time phases in which they use the risky and the safe arm exclusively. (These phases are further divided up according to whether a player is likely to hear news from his collaborator or not, depending on whether his collaborator was using the risky arm Δ instants ago.)

With bad news, the logic is reversed, as Bonatti and Hörner (2015) show. Observability leads to more effort because unexpected shirking depresses collaborators, and so leads to lower effort. Hence, such a deviation leads to an automatic punishment, and so mitigates free-riding relative to unobserved effort. Surprisingly, the unique equilibrium is in mixed strategies. That is, players randomize over extremal strategies, with a switching time to the risky arm (if no breakdown has occurred until then) that is chosen at random (this is *not* equivalent to a pure strategy equilibrium with interior levels of experimentation).

The fundamental difference is related to the kind of deviations that are most attractive. Under good news, a player that briefly deviates to the safe arm unbeknown to the other players becomes more optimistic than them, as the lack of a breakthrough is not as statistically significant for him. Hence, his relative optimism strengthens his incentives to revert to experimentation. The deviation "self-corrects" and local first-order conditions suffice, leading to interior but pure strategies. Under bad news, a player that deviates the same way becomes more pessimistic than the other players, as he views the absence of a breakdown as a more likely event. Increased pessimism pushes him further towards the safe arm. Hence, local first-order conditions are no longer sufficient, and so randomization must occur over pure extremal strategies.

2.3 Extensions

Irreversibility. All but one of the papers mentioned in Section 2.1 involve reversible actions. A player can resume experimentation after playing safe, and vice-versa. This is not innocuous. Instead, Frick and Ishii (2015) assume that using the risky arm (the adoption of a new technology) is an irreversible choice.[7] Their model is not strategic, to the extent that there is a continuum of agents. However, these agents are forward-looking, and learn about the state of the world (whether adoption is profitable or not) from the adoption choices of others. The more others adopt, the more likely it is that an exponentially distributed signal publicly reveals whether the state is good or bad. Opportunities to adopt follow independent Poisson processes, one for each of the agents. Frick and Ishii show that the adoption patterns differ markedly across scenarios. With good news, there cannot be a region of beliefs for which agents are indifferent between adopting or not. This is because, if an agent is willing to invest immediately after t in the absence of news (as would be the case in the interior of such a region, if it existed), then he must strictly prefer to invest at time t. Indeed, the only other event that can occur is good news, in which case adoption is optimal as well. Hence, there is no option value to waiting. As a result, the equilibrium is extremal: all agents

[6] This inefficiency calls for regulation. Campbell, Ederer and Spinnewijn (2011) consider the problem of the optimal deadline as a way of alleviating the free-riding. The intervention could be purely informational: Bimpikis, Ehsani and Mostagir (2016) suggest introducing a "black-out" phase in which no information is shared, a policy that is shown to be optimal within some class of policies. The optimal informational mechanism remains unknown.

[7] Irreversibility presupposes that the unit resource cannot be split across arms.

that get the chance to adopt do so up to some date t_1. If no breakthrough was observed by that time, agents stop adopting. As a result, the adoption pattern (the fraction of adopters) turns out to be a concave function of time. In contrast, with bad news, there is a region of beliefs (and so, an interval of time) over which agents are indifferent, and so only a fraction of agents that get an opportunity to adopt do so, resulting in an S-shaped adoption pattern.

It is worth mentioning that irreversibility is not necessarily socially costly. But this depends on which action is irreversible. In the models of Section 2.1, the first-best is an equilibrium outcome when switching to the safe arm is irreversible. If switching to the safe arm is irreversible, it is no longer possible to be a by-stander while others experiment. And because one still benefits from the accelerated rate of learning that results from others experimenting, it is optimal to follow the first-best strategy when others do so. More generally, we speculate that making the risky action reversible and the safe action irreversible is socially valuable, as it encourages experimentation, which is always too low.[8]

Murto and Välimäki (2011, 2013; hereafter MV11, MV13) consider stopping games (that is, games where switching to the risky arm is irreversible) with observational learning and pure informational externalities.[9] In both papers, stopping corresponds to taking the action whose payoff depends on the state (in this sense, it is the risky one). The models have notable differences. In MV11, the state is binary, values are correlated but not identical, and it is a good news model: if the state is good, agents that have not stopped yet receive a private, independent signal revealing the state at a rate that is exponentially distributed. In MV13, the state space is much richer and is the (common) optimal time to invest. Signals are received at the beginning once and for all, but the signal space is rich (the monotone likelihood ratio property is assumed to impose some discipline). In both models, agents observe the earlier times at which other agents chose to stop, but nothing else. A robust finding across both papers is that, even as the number of players grows large, efficiency is not achieved. Stopping occurs too late, relative to first-best. This is particularly striking in MV13, as it highlights the role of time passing by, and how it differs from static models, as in common value auctions in which asymptotic efficiency obtains. Further, MV11 show how, with many players and continuous time, exit behavior displays waves of exit interwoven with calm periods of waiting, during which the public belief hardly budges. Exit waves occur because the exit of an agent implies a discrete jump in the public belief, leading with positive probability to the immediate exit of further agents whose belief was nearly the same, which in turn triggers a further cascade of agents leaving.

Payoff externalities. Most of the papers considered so far focus on pure informational externalities. From a theoretical point of view, this allows us to isolate informational incentives. From an economic point of view, it is very restrictive. Other externalities arise in many contexts. For instance, competition in the product market leads to payoff externalities (with more firms "experimenting" in the risky market, profits have to be shared among more firms). Decision-making procedures also lead to externalities, when agents' action cannot be selected of their own free will. Congresspeople do not decide whether they want to experiment in isolation, but rather make decisions by majority voting.

Some papers explore such issues. Cripps and Thomas (2015) consider externalities in terms of congestion costs.[10] Imagine stepping out of an unfamiliar airport, hoping to catch a cab at the curbside. Naturally, there are signs above your head providing potentially critical information for

[8] We are not aware of a single model in which strategic interactions lead to over-experimentation. It would be interesting to examine under which conditions this might happen (heterogeneity?).

[9] See also Rosenberg, Salomon and Vieille (2013) for a related analysis in continuous time for the two-player case.

[10] See also Thomas (2015).

those deciphering the local alphabet. But you are not one of them. As it turns out, nobody is ahead of you, so this is your lucky day: either the first cab will be for you, or this is not a spot where cabs are authorized to pick up fares. What do you do?

Suppose instead that there are three people in front of you, so that you are fourth in line (your upbringing compelling you to incur the requisite congestion cost if you choose to stay). It is clear to you that their knowledge of the local idiom is as deficient as your own. Yet, having come before you, they might have observed useful data, like a cab picking up somebody ahead of them. Communication would certainly help, but interpersonal skills are not your forte (it may be embarrassing to admit you?re standing in line with no assurance this is even a cab stand. So any cheap talk by people in line is likely to be uninformative anyway). What do you do?

This problem looks daunting: incomplete information (a binary state), coarse information sets (the number of people ahead of you in the line), and both informational and payoff externalities. Yet, at least for some range of parameters, Cripps and Thomas solve for the optimal policy. One of the key insights is that information is nested: the first person in line knows as much as everyone else, since he must have come before. Hence, provided that the queue is not so long that you balk immediately, it is optimal to mimic this first person's behavior. If he ever balks, then so should you, as it reveals that he hasn't seen a cab, and you cannot hope for a higher payoff than his. Yet, it isn't always optimal to stay as long as he does. Cripps and Thomas (2015) do not provide a general solution, unfortunately, but their analysis provides insights into what can go wrong and suggests alternative candidate strategies.

In Strulovici (2010), the externality comes about from the decision-making process. Each player faces an independent one-armed bandit. However, actions are not independent. All players must take the same action at every instant, which gets decided by majority voting. It is a good news model, and outcomes are observed.[11] As a result, players are divided into two groups: those who have found out that their arm is good, and stubbornly cast their vote for the risky arm (which is no longer risky for them), and those who have not yet received a good signal and grow anxious that the safe arm might be the better alternative. If the first group takes over, the uncertain voters will be doomed to use the risky arm, which is most likely bad for them. If instead the second group pre-empts this from happening by voting for the safe alternative while they still can, players from the first group will be frustrated winners. How does this trade-off play out? As it turns out, it leads to under-experimentation. This is not too surprising, as the existence of the first group imposes a cost which the social planner does not account for, on uncertain voters that might be stuck with a bad arm. Strikingly, as the number of voters grows large, experimentation is abandoned as soon as the myopic cut-off is reached.

This analysis suggests interesting questions for future research. What if the bandits are correlated, for instance, when outcomes or information is unobserved? In a jury context, the accused is either guilty or innocent. In such environments, understanding under which conditions information aggregation obtains seems an important problem.

3 Motivation

3.1 Many Options and Coarse Learning

Experimentation is often carried out by a third party whose interests are not necessarily aligned with those of the experimenter. Nobody cares about being a guinea pig, yet it is in society's best interest that some of us are. Consider the example of Waze, as discussed by Kremer, Mansour,

[11] See also Khromenkova (2015) for the analysis of the mixed news case.

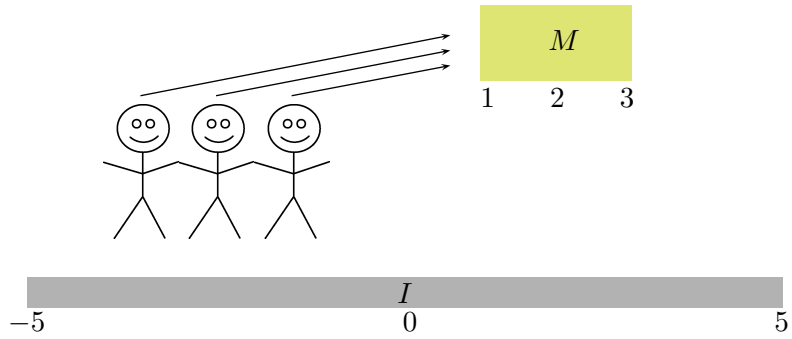


Figure 5: Experimentation cannot start if $\mathbf{E}[\pi_I] < a_M$.

and Perry (2014, KMP). Every day, motorists must arbitrage between two routes. Travel time is uncertain, and so is each motorist's payoff π_k, which is shared by the motorists traveling this route that day. Let us assume that it is uniformly drawn from $[a_k, b_k]$, with $k = M, I$ standing for the type of the route: $a_I < a_M < b_M \leq b_I$, so that the Merritt parkway is "safer" than I-95, but also $a_M + b_M > a_I + b_I$, so it is on average faster. Motorists cannot postpone or forward their daily commute, and arrive in a sequence. They do not care about experimenting, as travel times are independently distributed across days, and they need to travel only once a day. However, upon taking a route, their cell phone diligently relays travel time to Waze (and hence perfectly resolves uncertainty for that day) unbeknown to the motorist.[12]

Consider first the case in which $a_M = 1, b_M = 3$, whereas $a_I = -5, b_I = 5$ (with the usual stretch of imagination that negative payoffs demand). Our first motorist, an early bird aware of his position in the line, selects the M-route, as its expected payoff of 2 exceeds zero, I-95's expected payoff. The second motorist's choice is equally simple, *independent* of his knowledge of the first motorist's realized travel time. At worst, it is 1, which still beats the mean payoff from I-95. Hence there is nothing he can be told that would change his mind: route M it is. More generally, if $\mathbf{E}[\pi_I] < a_M$, all drivers take the Merritt parkway, and Waze loses its purpose (not to mention, to the dismay of Waze's innovators, Google's attention). This is captured by Figure 5.

Instead, suppose that $a_M = -1 < b_M = 5$. The early bird selects route M, as before, because its expected payoff exceeds I's expected payoff. Let us first assume that later motorists observe earlier ones' choices and payoffs (if not, they all herd on the M-route). This is what is called transparency here. If the first motorist's realized payoff falls below 0 (say -0.5), then the second motorist selects the I-route, resolving all residual uncertainty as far as the East Coast is concerned.

However, if the early bird's realized payoff is 0.5, the second (as well as all later) motorist mimicks the first, despite the potential benefits from the I-route. This is where Waze comes into play. Suppose that Waze commits to the following policy: it discloses to the second motorist whether the first motorist's travel time is below or above 1, and nothing more. If the realized payoff is 0.5, the second motorist is willing to experiment with the I-route nonetheless, as, on average, it still does as well as the M-route (0).

What if the first motorist's payoff is 1.5? The second motorist must be told that the early bird did well (more than 1 as a payoff), and so he will follow suit. The third motorist, however, need not be told what the second one was told. In particular, Waze can commit to tell him to take the M-route in one of two events. Either the second motorist took the M-route, as did the first, and their common realized travel time fell short of 2.33; or the early bird's travel time fell short of 1,

[12] As all stories, this one has practical limitations. In particular, choices do not affect travel times. This ignores congestion. But as every user of I-95 or US-101 will attest, there are some things logic cannot explain.

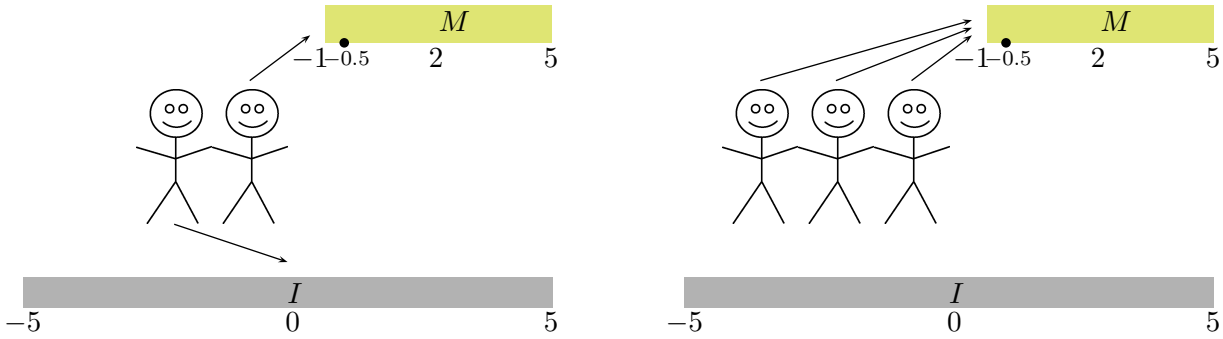


Figure 6: Choices of second and third agent under transparency.

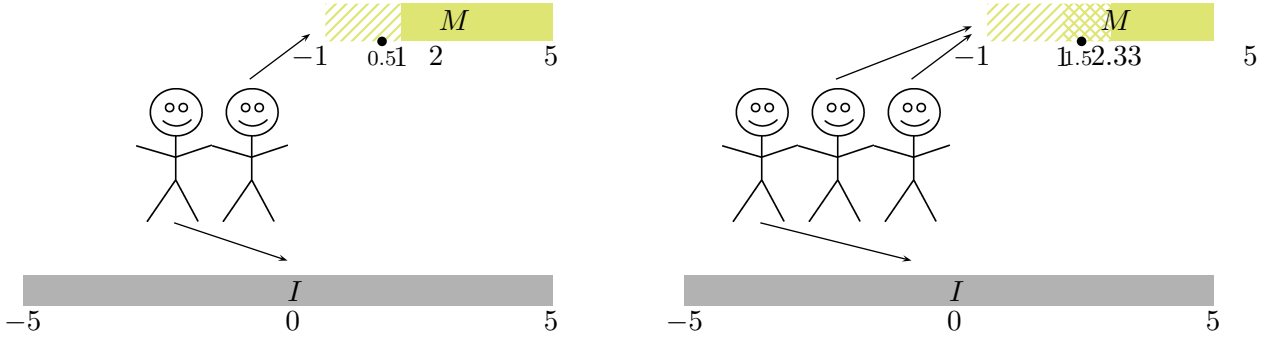


Figure 7: Choices of the second and third agent under the optimal recommendation policy.

so that the second experimented with the I-route, and the M-route turns out to be the faster one. Comparing these two possibilities, the choice of 2.33 as a threshold ensures that the third motorist is willing to experiment with the I-route if told so. And if he is not told to experiment (and neither were any of the first three), the fourth is willing to do so, so that, with four motorists or more, all uncertainty can be resolved, while making sure that each motorist remains sufficiently uncertain about the right choice that it is optimal for him to follow Waze's recommendation. See Figure 7 for an illustration.

KMP develop this logic further. With two actions and perfect learning, as in this example, the optimal recommendation mechanism has a simple partitional structure.[13] The partition pertains to the realized value of the interval $[a_M, b_M]$ (the safer action) into a collection of disjoint sets (not necessarily a partition, and not necessarily non-empty) $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_I$ with the property that, if the realized payoff of the first motorist is in $\mathcal{I}_i$, the i-th agent is the first to experiment with the action with the lower mean (and all later agents make the right choice). Partition strategies need not be optimal under stochastic learning, but KMP show that as the number of agents grows large, it does very well nonetheless (it is asymptotically optimal).

3.2 Few Options and Rich Learning

Sometimes once is not enough. Learning takes time, the early bird's reassurances notwithstanding. Let us follow Che and Hörner (2016) and assume that payoffs are coarse –either high (1) or low (0)– but learning is not. Specifically, if the state of the world is good, selecting the risky option is

[13] The main focus of KMP is on the case in which the designer maximizes the expectation of the average (undiscounted) payoff of the agents, $I^{-1}\sum_{n=1}^{I}\pi^i$, where I is the number of agents and π^i is agent i's realized utility.

best for everyone, but few of those that have tried bother to let others know. Specifically, assume that, if and only if the state is good, (positive) feedback arrives at a random time, exponentially distributed with an intensity proportional to the (infinitesimal) mass of agents that consume at that point in time. That is, the rate at which feedback arrives is

$$\mathbf{P}[\text{positive feedback}] = \lambda \mu_t \cdot \omega \mathrm{d}t,$$

where $\lambda > 0$ is some velocity parameter, μ_t is the fraction that consumes at time t, and $\omega = 0, 1$ is the state. Each consumer consumes at most once. A fraction $\rho > 0$ of consumers have a positive opportunity cost $c \in (p_0, 1)$ (*regular* agents), where p_0 is the probability that $\omega = 1$. The remaining agents (fans) have sufficiently unattractive outside options that they experiment no matter what, although they are just as critical in providing feedback. Because $c > p_0$, these are the only ones experimenting at the start. They are the seeds that jump-start the process. However, if Waze (or its equivalent) simply relays the feedback that these agents provide, if any, the remaining ρ fraction of agents remain on the sideline, unless positive feedback is reported. As a result, experimentation is too slow. To accelerate it, the designer may *spam* these agents, by recommending a fraction α_t of these agents that they try out the product. (Of course, this requires that these agents rely on the designer as their exclusive source of information.) The designer commits to this policy, and selects at random this fraction of guinea pigs. If one of the early consumers has reported that the payoff is high, the designer and consumers have aligned interests. If not, the designer's belief evolves according to

$$\dot{p}_t = -\lambda(1 - \rho + \rho\alpha_t)p_t(1 - p_t). \tag{3}$$

The designer can capitalize on the possibility that he has learnt the state to make spamming credible, provided that spamming is meted out proportionally to the likelihood of this event. Conditional on being told to consume, an agent's expected utility from consuming is

$$\frac{\frac{1-p_0}{1-p_t}\alpha_t p_t + \frac{p_0-p_t}{1-p_t} \cdot 1}{\frac{1-p_0}{1-p_t}\alpha_t + \frac{p_0-p_t}{1-p_t}} = \frac{(1-p_0)p_t\alpha_t + (p_0 - p_t)}{(1-p_0)\alpha_t + (p_0 - p_t)},$$

given that $(p_0 - p_t)/(1 - p_t)$ is the probability assigned by this consumer to the designer having learnt the state by time t, and $(1 - p_0)/(1 - p_t)$ is the complementary probability that he has not. Setting this equal to c, and solving for α_t, we obtain

$$\alpha_t = \frac{(1-c)(p_0 - p_t)}{(1-p_0)(c - p_t)}. \tag{4}$$

Spamming is not credible at the very beginning ($\alpha_0 = 0$) because it is virtually impossible that the designer has learnt something useful by then. But as time passes by, spamming becomes more credible, and the rate of spamming accelerates. Equations (3) and (4) provide a pair of equations that can be solved for the maximum spamming α, and the corresponding belief p.

However, if the designer is benevolent, spamming has to cease eventually. Persistent lack of feedback from consumers reinforces his belief that the payoff is low, so that experimentation has to stop. When? With only two types of agents, the threshold at which it stops coincides with the cut-off $\underline{p}$ that prevails absent the incentive constraint that constrains the amount of spamming. This is because, at the point at which experimentation with regular agents stops, the continuation game is identical to the one without incentive constraints: only fans experiment. As a result, the first order condition that dictates when spamming ceases is the same in both problems. Incentives

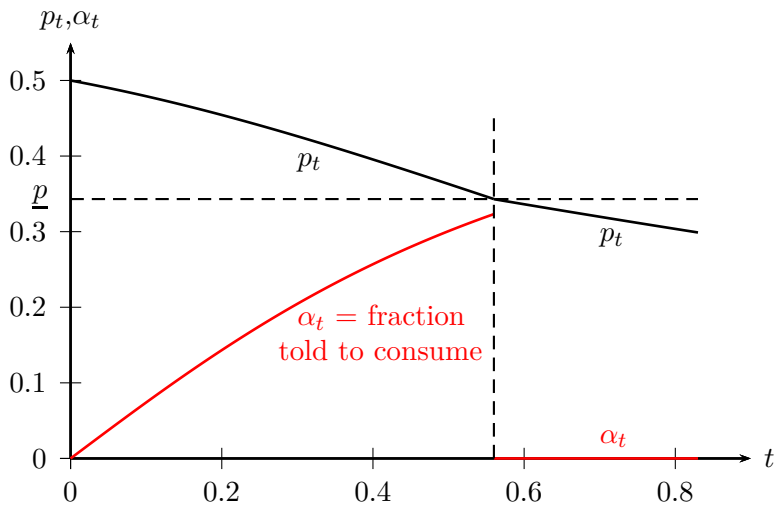


Figure 8: Experimentation rate and belief evolution.

constrain the rate of experimentation from regular agents, but when considering when to stop, this rate is irrelevant for the marginal value of continuing for another instant.[14]

Overall, the pattern of experimentation over time is hump-shaped. Experimentation is most valuable at the start, but this is when the agents' suspicions prevent it from being in full swing. It then rises gradually as the designer's recommendations gain credibility, and finally tapers off as growing pessimism means that high-cost agents should not be carrying out such experimentation. See Figure 8.

An important assumption made so far in our discussion of these two papers is that recommendations are confidential. If spamming is public (reviews, endorsements, etc.), the designer is further constrained. In the last example, it is then optimal for the designer to behave as follows. If he gets feedback at any time, he tells everyone to experiment from that point on. If he fails to, he picks at random a time at which he claims that he did, and gets everyone to experiment from that point on, but only until some later time at which, if he still hasn't heard positive feedback, he concedes that his claim was unwarranted, and experimentation (from regular agents) stops. This spamming occurs at most once. The distribution over the random time is chosen to make experimentation start as early as possible, subject to the claim that he received feedback being sufficiently credible that agents are willing to follow his recommendation over the next phase.

The models we have discussed so far assume that the principal's objective is to maximize social welfare. This may be the right objective for many recommender systems (like Waze, news websites recommending stories to read, or review websites recommending places to visit or apps to download) since the equilibrium with full information disclosure is likely to have inefficiently low experimentation because of the positive externality learning by early consumers has on later ones. That said, if the recommender's objective is to maximize the number of users (for example, because they can monetize traffic to their website) then the objective function can become more complicated, especially in the presence of competition between recommender platforms. That problem has been studied by He (2016). She considers differentiated platforms and agents making single-homing

[14]With more than two types, this reasoning carries over to the lowest type among those agents with a positive opportunity cost. But it does not apply to those with a higher cost. Because the rate of experimentation is too slow, the designer optimally chooses to spam such agents a little longer than he would absent incentive constraints. His reduced experimentation "capacity" in the continuation makes their willingness to partially experiment valuable, and he spams at beliefs below those at which they should in the unconstrained problem.

decisions which one to join based on the experimentation and recommendation policies the platforms follow.

When agents choose platforms without knowing whether they will be early or late in the queue for recommendations, the optimal policy is to maximize efficiency of dynamic learning, consistent with what we have described so far. However, if agents have some information regarding whether they are likely to be early or late, then a new economic force appears. A platform attracts more early users by recommending the myopically optimal product/route instead of the dynamically-efficient one (that takes into account the option value of learning and making better recommendations to future agents) in order to attract more early customers. The platform with a larger early market share can obtain a competitive advantage in the future because learning from more early consumers can lead to superior recommendations to future customers. He (2016) shows that a merger of the platforms can make the recommendation system more efficient, because a monopolist can internalize the business stealing effect.

4 Delegation

Information is only one of the many tools available to the principal. Authority is another. In two very different environments, Guo (2016) and Grenadier, Malenko and Malenko (2015, GMM) examine the scope for and the structure of delegation in dynamic problems. In both papers, the direction of the conflict of interest (expressed in terms of timing of the optimal decision) changes the answer drastically. To understand why, we start with a simple example.

4.1 A Simple Example

Consider a simple cheap-talk model based on Crawford and Sobel (1982, CS).[15] We assume that the state is uniformly distributed on $[0,1]$, and that preferences are quadratic. The receiver is unbiased, with preferences $-(t-y)^2$, where t is the state of the world (the sender's type) and $y \in \mathbf{R}_+$ is the action by the receiver. The sender has preferences $-(t-(y+b))^2$, where $b \in \mathbf{R}$. We distinguish (i) *positive* bias, $b > 0$, and (ii) *negative* bias, $b < 0$. Throughout we assume $|b| < 1/2$, for the problem to be interesting (babbling is the commitment solution otherwise). In CS, this distinction is irrelevant, as it is a matter of normalization. Here instead, we assume that t also indexes time, and that, as time passes by, opportunities vanish. That is, if the receiver acts at time t, his available actions are elements of $Y_t = \{y \in \mathbf{R} : y \geq t\}$. Delaying is always an option, but acting can only take place once (it ends the game). We ignore any discounting or cost of delay, the possibility of never taking an action (add a large penalty for doing so) and leave aside the technical modeling details regarding continuous-time strategies. Let us think of the sender being able to send a message in $[0,1]$ at any time t and impose standard measurability restrictions.

Because time flows from "left to right," we will argue that the positive and negative bias cases are very different, that is, admit very different sets of equilibria.

Positive bias (bias for delay). In that case, we claim that the receiver can do as well as under commitment in the standard model of CS in which types have no temporal meaning. The commitment case has been solved by Holmström (1984) and in greater generality by Melumad and Shibano (1991). It involves setting:

$$y(t) = \begin{cases} t+b & \text{if } t \leq 1-2b, \\ 1-b & \text{if } t > 1-2b. \end{cases}$$

[15] A related example appears in a preliminary draft by Gordon and de Villemeur (2010).

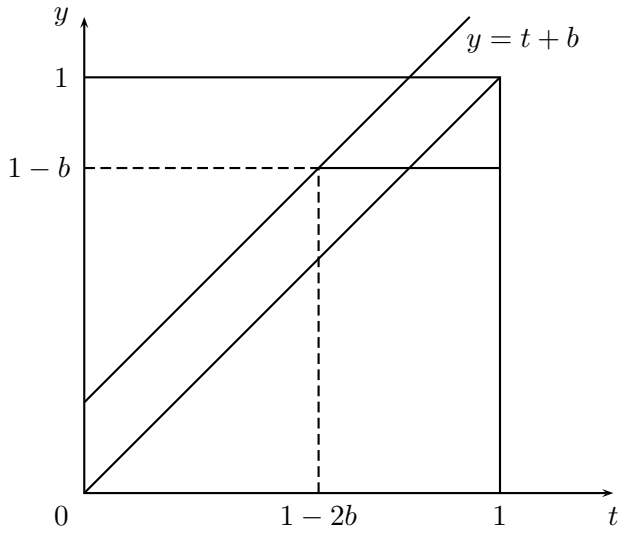


Figure 9: Optimal Delegation.

See Figure 9. We focus on the equilibrium that is best for the receiver.[16]

Why does this require no commitment? Time endows the receiver with the necessary power. When the sender tells him to act at time $t \leq 1-b$, the receiver learns that the state is $t-b$, but this action is no longer available. So the best he can then do is to take action t. Similarly, once time $1-b$ arrives, he knows that the state is in $[1-2b, 1]$, and his expected action $1-b$ matches what he would do in the next split second had perfect revelation continued –given that it won't, it remains optimal in the absence of information.

Note that (unlike in the commitment case) there are lots of other equilibria. For instance, babbling is always an equilibrium, and so are many other messaging strategies.

Negative bias (bias for action). With a negative bias, it is clear that the commitment optimum (which now involves a "floor" concerning types in $[0, 2|b|]$, pooled at b, and perfect separation over high types, who obtain $t-|b|$) is no longer incentive compatible. If a type $t \geq 2|b|$ reveals his type, the receiver can take action t if he can take action $t-|b|$, so that sequential rationality prevents him from taking the sender's optimal action.

Does this mean that we are back in the world of CS? In addition to the equilibria of CS, additional ones arise because of the role of time. For instance, consider a two-partition equilibrium, in which types in $[0, t_1]$ get mapped into action $y_1 > t_1/2$ and types in $(t_1, 1]$ get mapped into $y_2 = (1+t_1)/2$. This can be achieved by waiting till time y_1 (messages are interpreted as "babble" until then), and then senders separating at time y_1 between those whose type is above t_1 and those whose type is below.[17] Type t_1 must be indifferent, which requires

$$t_1 - b - y_1 = y_2 - (t_1 - b), \text{ or } y_1 = 2(t_1 - b) - y_2 = 2(t_1 - b) - (1+t_1)/2.$$

[16] In CS, *ex ante* payoffs are aligned and so this also gives the highest payoff to the sender. But this is only true for any allocation in which the receiver takes his preferred action given his information (which is not necessarily the case in our environment, in which there might be a lower bound on the action he can take).

[17] If senders do so by randomizing over $[0, t_1]$ or $(t_1, 1]$ depending on whether their type is above or below t_1 at time y_1, all messages are on path, so that what would happen if the sender deviated at time t_1 is irrelevant –the receiver will act at that time. And the receiver has nothing to gain from waiting any longer, if he expects no further meaningful information to be transmitted.

The receiver must prefer waiting until time t_1 to acting at time $1/2$ (his favorite in the absence of any information), namely,

$$\int_0^1 (t-1/2)^2 \mathrm{d}t \geq \int_0^{t_1} (t-y_1)^2 \mathrm{d}t + \int_{t_1}^1 (t-y_2)^2 \mathrm{d}t$$

or

$$4t_1^2 - (3+16b)t_1 + 8b(1+2b) \leq 0,$$

an expression that reduces to $1/2+2b$ when $t_1 = (1+4b)/2$, the value of t_1 given the partition $\{[0,t_1],(t_1,1]\}$ in the original CS-model. Hence, for $b < 1/4$ (the usual condition for non-babbling equilibria to exist in CS), we have a range of values of t_1, namely,

$$t_1 \in \left[\frac{1}{2}+2b, \frac{1}{2}+2b+\frac{1}{8}\left(\sqrt{9-32b}-1\right)\right],$$

for which such equilibria exist (this interval is non-empty by our assumption that $b < 1/4$).[18] But note that they are all worse than the original two-partition equilibrium of CS, as the higher value pushes t_1 above what it would have been if $y_1 = t_1/2$. The problem is that the lower interval is already too large, compared to the higher one.

It turns out that in our uniform-quadratic set-up, this is a general feature: no pure-strategy equilibrium involving one-way communication improves on the best CS equilibrium.[19] This is because, first, and as in CS, an equilibrium must be partitional. Because the sender's preferences satisfy the single-crossing property, if types $t < t'$ are both mapped into the same outcome y, then so must all types in the interval $[t,t']$. Letting $[t,t']$ be the maximum interval associated with y, either $y = (t+t')/2$ because the receiver finds out about the sender being in that range before this time arrives, or $y > (t+t')/2$ because he learns about it at time y. Surely, the type $y+|b|$ is in the interval $[t,t']$ and so each interval is of length at least $2|b|$, and so there are finitely many intervals, $\mathcal{I}_1,\ldots,\mathcal{I}_K$. The action associated to the highest interval, $\mathcal{I}_K$, must be its midpoint, as it is best for the receiver to take if the time comes. Consider the problem of maximizing the receiver's expected utility over sequences $\{y_1,\ldots,y_K\}$, $t_0 = 0,\ldots,t_K = 1$, such that each y_k is at least as high as the midpoint of $\mathcal{I}_k$, and each $t_k - |b|$ is equidistant from y_k and y_{k-1}. It is easy to see that it is maximized at the midpoints.

4.2 Economic Applications

Guo and GMM consider two dynamic principal-agent models in which there is uncertainty (an unknown state of the world) and risk (imperfect learning via the observation process, though in GMM the principal only learns via the agent's reports). In both models, the agent has private information that does not evolve over time. The principal and the agent have a conflict of interest. In Guo, it is an experimentation model in which the agent has a vested interest in playing risky or safe (depending on his bias), but also private information at the start regarding the viability of the project. In GMM, the principal faces a problem of real options (an irreversible investment decision) based on a publicly observable signal process (an exogenous geometric Brownian motion) that relates to the project's profitability via a factor that is the agent's private information.[20]

[18] This class of equilibria involves the standard CS equilibrium of partition size 2. The same argument applies to any CS equilibrium, independent of its partition size.

[19] Of course, one can do better with either a randomization device or the receiver talking as well, along the lines of long cheap-talk (Aumann and Hart, 2003), which becomes especially natural in this dynamic environment.

[20] The main difference isn't Poisson vs. Brownian uncertainty. Rather, it lies in the updating. In Guo, the public signal is informative about the agent's type, whereas as in GMM, it is not, which simplifies somewhat the updating.

In both models, the principal designs an optimal mechanism, with or without commitment. The principal has authority, and can dictate the agent's choices based on his reports. It matters whether the agent's bias leads him to favor delaying the (optimally, irreversible) decision or not. If he enjoys a longer phase of learning, the principal has to counteract the agent's incentives to act too late; if he enjoys a shorter phase of learning, inefficiently early stopping/investment have to be prevented. This is the dichotomy already present in the simple example of the previous section.

In the case in which for any given prior belief the agent prefers to switch to the safe arm later than the principal, Guo shows that the optimal policy (from the principal's viewpoint) can be implemented by a *sliding rule*: based on the observed history, in which experimentation was pursued throughout by the agent, the principal updates his belief as if the agent was equally uninformed regarding the project's potential. There is a cut-off (in belief space) that he picks at which he dictates the agent to stop. Until then, the agent is free to stop on his own, which might happen if the agent's private information is so bad that earlier termination might be preferable from his point of view despite his bias. The simplicity of this rule is remarkable. Furthermore, as in our example, and for the same reasons, the principal need no commitment to implement it. If he were free to grant authority to stop or continue experimentation at every moment, he would find it optimal to leave it up to the agent to decide up to the moment at which this threshold is being crossed.[21]

In GMM, the bias is exactly as in CS: upon exercise of the option to invest, the agent receives an additional b to the investment payoff. Positive bias means the agent is biased towards early exercise. As in our example, this implies that delegation dominates communication (at least in the simple class of equilibria with one-way communication, as described above). Indeed, GMM show that such (stationary) equilibria involve a partition of the possible values of the signal process into intervals. The agent recommends that the principal invest as soon as the observable process enters a high enough interval, where "high enough" depends on his information. At that point, the principal invests. In contrast, when bias is negative, so that late exercise is favored by the agent, the outcome of delegation, which often involves full revelation, can also be achieved in the game with cheap talk and no commitment.

5 Information Design

In this section, we abstract from exogenous learning. In the absence of a message by the informed player, the other player does not learn anything. At the same time, we enrich the model by allowing the state to change over time, and by endowing the informed player with a richer set of verifiable messages.

5.1 An Investment Example

A decision-maker (DM) must decide whether to invest ($a = I$) or not ($a = N$) in a project with uncertain returns. Her decision is a function of her belief in the state of the world $\omega = 0, 1$, with a return $r(\omega)$. Doing so is optimal if and only if $p = \mathbf{P}[\omega = 1] \geq p^*$. Accordingly, the invest region is $I = \{p : p \geq p^*\}$, see Figure 10. Unfortunately, the DM does not observe the state, and must rely on an advisor's goodwill to get information. The advisor receives a fixed payoff of 1 from investment. Our players are trained statisticians, not economists: they have no idea about the potential benefits of contractual payments, but know all about statistical experiments *à la* Blackwell. In particular, the advisor can commit to a splitting of his prior belief into an arbitrary distribution over posterior

[21] Guo also considers the case of bias toward early action, and also finds that the optimal contract with commitment is time-inconsistent and so can?t be implemented without commitment.

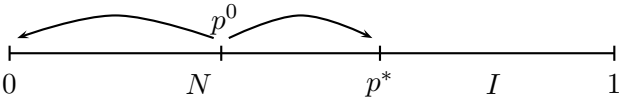


Figure 10: Investment as a function of the belief.

beliefs (a mean-preserving spread), with the DM only observing the realized posterior.[22] If the game is one-shot, the best experiment from the advisor's point of view is rather immediate. To maximize the probability that the DM invests, it is best to make the posterior belief that leads to an investment to be such that the DM is just indifferent, that is, equal to p^*, and the other posterior equal to 0.[23] See Figure 10. The resulting payoff is p^0/p^* (more generally, $\min\{1, p^0/p^*\}$, as $p^0 \gtrless p^*$). This is the celebrated concavification formula of Aumann and Maschler (1995), which states that the maximum payoff over possible splittings is equal to the smallest concave function no less than the payoff function given a belief (namely, $\min\{1, p^0/p^*\} = \text{cav}\mathbf{1}_{\{p \geq p^*\}}(p^0)$).

We call the corresponding strategy *greedy*. The greedy strategy is the strategy that minimizes information revelation subject to maximizing the one-shot payoff. Formally, it is defined for environments in which there is a binary action (such as investing or not) and the payoff to be maximized (the advisor's) is independent of the true state, although it may depend on the action. Let us define the region of beliefs I for which investing is optimal from the DM's point of view (here, $[p^*, 1]$). Then the greedy strategy specifies that no information be disclosed if $p \in I$. Otherwise, it specifies that the DM be told one of two posteriors p_I and p_N, with weights λ and $1-\lambda$ (so $p = \lambda p_I + (1-\lambda)p_N$), with $p_I \in I$, in a way that maximizes the weight $\lambda \in [0, 1]$. Such a greedy strategy is generically unique and can be solved using linear programming.

Now suppose that the investment decision is repeated twice, yet the state is persistent: it remains the same with probability $\rho \geq 1/2$, and it switches with probability $1-\rho$. Hence, the posterior belief, $\phi(p)$ lies somewhere between p and $1/2$, depending on ρ. The advisor's payoff is $\mathbf{1}_{\{a_1=I\}} + \beta \mathbf{1}_{\{a_2=I\}}$, where a_n is the action in round n, and $\beta \geq 0$ is the weight assigned to the second round.

To apply Aumann and Maschler's formula, we must compute the payoff $\pi(p)$, where p is the posterior belief of the DM *after* the splitting, but *before* the investment decision in the first round. Because at that stage, the DM invests if and only if $p \geq p^*$, $\pi(\cdot)$ jumps up by 1 at p^*, but this fixed benefit is sunk and simply added to what accrues from the second period. If $p \geq \phi^{-1}(p^*)$, then no information should be disclosed in the second period, because $\phi(p) \geq p^*$. In contrast, disclosing additional information is optimal if $p < \phi^{-1}(p^*)$. As is clear from the left panel in Figure 11, the graph of the resulting payoff differs markedly according to whether $p^* \gtrless 1/2$ (equivalently, $p^* \gtrless \phi^{-1}(p^*)$). In case $p^* > 1/2$ the optimal strategy is greedy for all β. The intuition is as follows. The best case scenario for period 2 payoffs is when the agent reveals nothing in period 1. In that case, since $\phi(p) < p^*$, the optimal disclosure in period 2 is to split the belief between p^* and 0. Now consider the agent following the greedy strategy in the first period, which also induces beliefs 0 and p^*. After that, whatever the interim beliefs, in the beginning of the second stage, the beliefs are less than p^* and in both cases it is optimal to induce posteriors of 0 and p^*. Since in that

[22] Note that the advisor need not observe the realized state, but somehow must have access to a technology that generates any posterior distribution.

[23] Such static problems are considered in greater generality by Rayo and Segal (2010), and Kamenica and Gentzkow (2011).

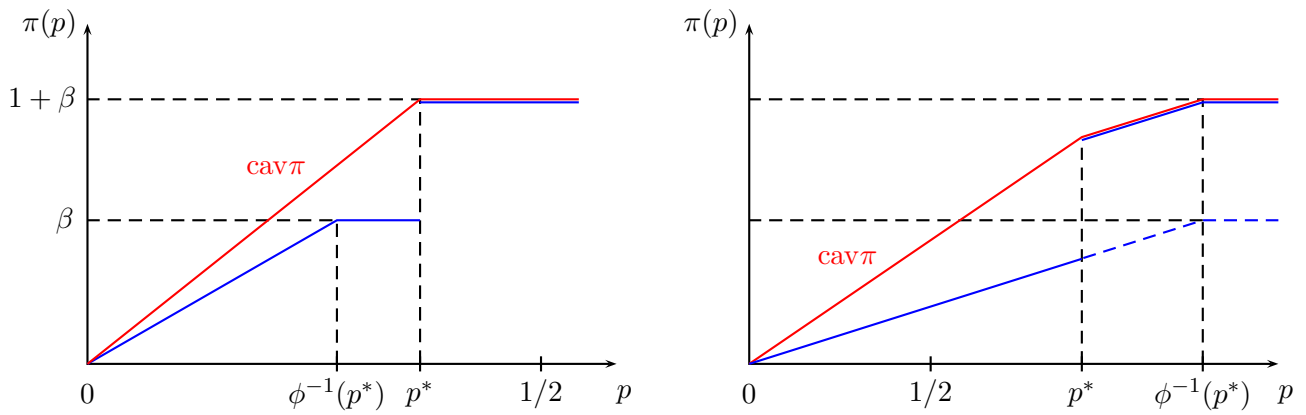


Figure 11: Optimal strategy in the two-period example.

range the payoff function is linear in beliefs, there is no cost for second-period payoffs if the agent is greedy in the first period. In contrast, if $1/2 > p^*$, greedy is optimal only if β is sufficiently small, because if $p^* < 1/2$ then if the belief at the end of period 1 is p^* then the belief at the beginning of period 2 is above p^* and hence disclosure in period one creates a probability distribution over period-2 beliefs on a concave part of the value function. So in this case there is a tradeoff between maximizing period 1 and period 2 payoffs.

Figure 12 shows why the greedy strategy is no longer optimal once β is large enough. For $p < \phi^{-1}(p^*)$, the optimal strategy consists in releasing just enough information for the posterior belief to be equal to $\phi^{-1}(p^*)$. Hence, in this example, the optimality of the greedy strategy breaks down if the future is important enough, and beliefs drift towards the interior of the Investment region. This is easy to understand: if beliefs drift outwards, as on the right panel of Figure 11, then both splittings according to the greedy strategy, and movements due to mean-reversion keep all beliefs inside the region over which π is linear. Given the martingale property of beliefs, this ensures that there is no loss in using the greedy strategy. Once beliefs drift towards the investment region, splitting the belief all the way to p^* leads to a cost for the future, as mean-reversion pushes the posterior belief into a range over which π no longer shares the same slope than in the range where the prior belief lies.

How robust is the optimality of the greedy strategy? Remarkably, Renault, Solan and Vieille (2015, RSV) show that, in the infinitely (discounted) repeated investment game, the greedy strategy is optimal for an arbitrary two-state irreducible Markov chain. Despite the fact that the future may outweigh the present, stationarity ensures that the scenario depicted in Figure 11 does not occur, or rather that, whenever the intertemporal cost cannot be avoided, the greedy strategy achieves the best trade-off.

Despite this remarkable result, the conditions under which the greedy strategy is optimal are very special. Suppose, for instance, that some extraneous public signal, obtained at the beginning of each of the two rounds, determines whether $p^* = 0.2$ or 0.8, with both being equally likely, but independent across rounds. To simplify, we assume that $\rho = 1$, so that the state is the same in both rounds. Let $p^0 = 0.2$. Suppose that $p^* = 0.8$ in the first round. By not disclosing any information in the first round, the advisor receives a payoff of $\beta/2$, because the prior belief suffices tomorrow, with probability 1/2. In contrast, by following the greedy strategy, the advisor splits the prior into 0 (with probability 3/4) and 0.8 (with probability 1/4), resulting in a payoff of $(1+\beta)/4 < \beta/2$ if $\beta > 1$. Hence, the greedy strategy is not optimal if the future is sufficiently important. If the DM

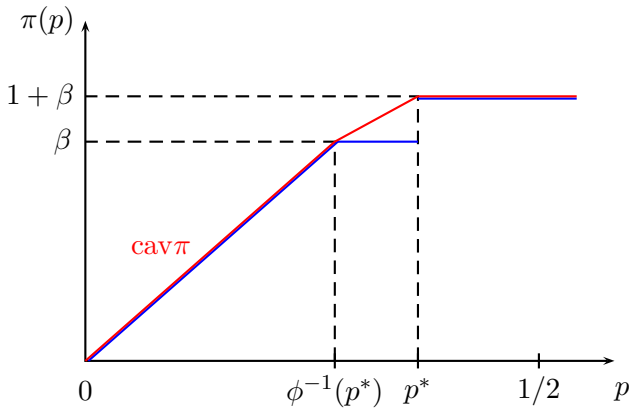


Figure 12: Greedy is not optimal when β is large.

invests once and only once, the threshold patience further decreases to $\beta \geq 1/2$.

This example relies on p^* being random, which implies that the relevant distance between the prior belief and the boundary of the investment region might become more favorable by waiting, rather than giving away information. The same phenomenon arises without randomness, by simply considering more states.

Indeed, RSV construct the following three-state counter-example, illustrated in Figure 13. The horizon is infinite, and the discount factor is $\delta < 1$. The DM invests in the light-gray area, the triangle with summits $\omega^*, \omega_1, \varepsilon\omega_1 + (1-\varepsilon)\omega_2$, where $\varepsilon > 0$ is a small number. Suppose that the prior is p_1 and the state's persistence is $1/2$, so that given the invariant distribution ω_2, the belief absent any disclosure after one round is

$$\frac{1}{2}p_1 + \frac{1}{2}\omega_2 = \varepsilon\omega_1 + \frac{1}{2}\omega_2 + (\frac{1}{2} - \varepsilon)\omega_3.$$

The greedy strategy calls for a splitting between ω_3, and ω^*, with probability 4ε on ω^*. The resulting payoff is $4\varepsilon\omega^* < 4\varepsilon$. By waiting one round and splitting the resulting belief between p_2 (with probability $1/(2(1-\varepsilon))$) and p_3, the resulting payoff is $1/(2(1-\varepsilon))$, so that, as long as $4\varepsilon < \delta(1-\delta)/(2(1-\varepsilon))$, the latter strategy dominates.

5.2 Beeps

Ely (2015) tells the following relatable story of "beeps" to motivate the optimal design of information disclosure. A researcher clears her email inbox and begins to work productively at her desk. She believes that emails arrive according to a known stochastic process, and she feels compelled to stop working and check her email when her belief ν that a new email has arrived is strictly greater than some threshold $p^* \in (0,1)$. Assume the researcher gets utility from working but not from checking email since she will become distracted on her computer (so in this stylized example, her compulsion to check email is purely pathological). Her discount rate is $r > 0$.

The researcher can set up her computer to produce an audible beep when an email is received. Perhaps surprisingly, this "distraction" may actually increase the expected time until the researcher checks her email. The beep will compel the researcher to immediately check her email (since her posterior belief μ will jump to 1), but the knowledge that the beep mechanism is enabled (and functional) allows her to work indefinitely with the certainty that no email has arrived, as long as she has not heard a beep. Many other beeping mechanisms are imaginable, so the question

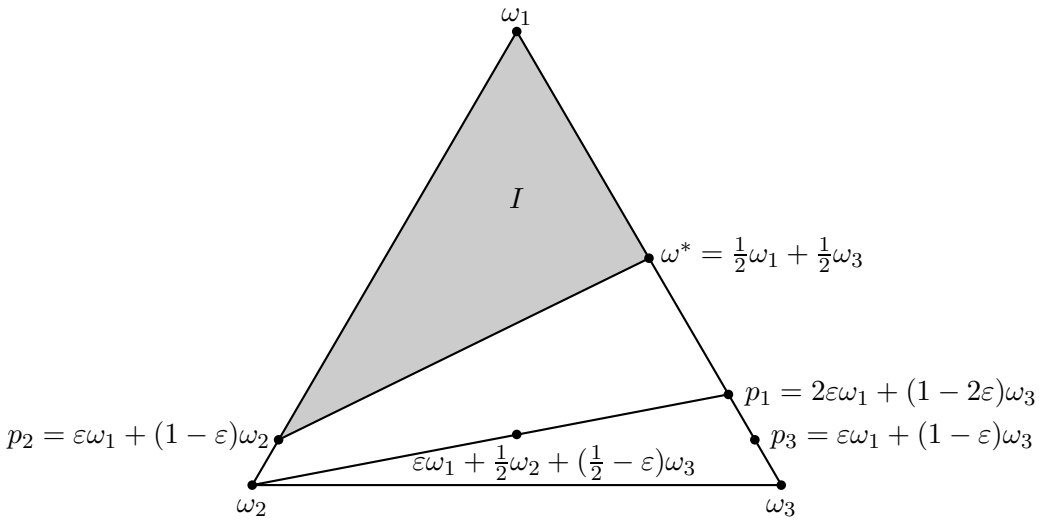


Figure 13: Greedy is not optimal with three states.

naturally arises: What informational filtering policy maximizes the expected discounted time until the agent checks her email?

The set of states is $\Omega = \{0, 1\}$ indicating whether or not an email has arrived, so state 1 is absorbing. Emails arrive at Poisson rate λ, but we work in discrete time with periods of length Δ. Hence, the transition probability from state 0 to state 1 is $m := 1 - e^{-\lambda\Delta}$. Let μ_t (ν_t) denote the agent's posterior belief, at time t before (after) receiving the principal's message, that an email has arrived. According to Bayesian updating, when the agent does not receive new information her beliefs obey the following law of motion:

$$\mu_{t+1} = f(\nu_t) = \nu_t + m(1 - \nu_t).$$

In the initial state, we assume no emails have arrived so it is common knowledge that $\mu_0 = \nu_0 = 0$. The agent checks her email iff $\nu_t > p^*$. The principal's utility can be expressed directly as a function of the agent's posterior belief:

$$u(\nu) = \begin{cases} 1 & \text{if } \nu \leq p^*, \\ 0 & \text{if } \nu > p^*. \end{cases}$$

The principal's discount factor is $\delta = e^{-r\Delta}$. The principal commits to an information disclosure policy, so the agent always interprets the principal's messages correctly.

The solution to this problem is an immediate corollary of the optimality of the greedy strategy, as established by RSV.[24] Here, it is straightforward to compute.

Case 1: $\mu \leq p^*$. Since the agent is taking the desired action (working), the principal releases no information.

[24]Since this model of beeps satisfies the conditions of Theorem 5 in RSV, it follows that the greedy strategy is optimal for any initial distribution, in particular, for the choice of $\mu_0 = \nu_0 = 0$. Indeed, this example of beeps is formally equivalent to the investment model in RSV in which the agent's state-dependent payoff from investing $r : \Omega \to \mathbf{R}$ is given by $r(0) = \frac{p^*}{1-p^*}$ and $r(1) = -1$. For then the agent invests and hence the principal receives payoff 1 if and only if $0 \leq r(\nu) = -\nu + (1-\nu)p^*/(1-p^*)$, *i.e.*, $\nu \leq p^*$.

Case 2: $\mu > p^*$. The principal chooses the greedy binary splitting at μ, *i.e.*, the agent solves $\max a_I$ under the constraints $\mu = a_I p_I + a_N p_N$, over probabilities p_I, p_N, and weights a_I, a_N, with $p_I \leq p^*$, $a_I, a_N \geq 0$, $a_I + a_N = 1$. It is easy to see that the maximum is achieved by $p_I = p^*$ and $p_N = 1$, so that

$$a_I = \frac{1-\mu}{1-p^*}, \quad a_N = \frac{\mu - p^*}{1-p^*}.$$

Therefore the principal should release two different messages, once that keeps the agent's beliefs at p^* and one that sends the agent's belief to 1.

Just because the greedy strategy is not robust does not imply that there are no interesting economic settings in which it is optimal. Ely provides a few such examples. More generally, however, little is known on the structure of the optimal strategy. Plainly, the principal's problem is a standard Markov decision problem, where the state variable is the agent's posterior belief, and the control is any splitting over beliefs. Hence, in the discounted case, it suffices to solve the optimality equation

$$V(\mu_t) = \max_{\substack{p \in \Delta(\Delta(\Omega)) \\ \mathbf{E}_p \nu_t = \mu_t}} \mathbf{E}_p[(1-\delta)u(\nu_t) + \delta V(f(\nu_t))],$$

or, in Aumann and Maschler's compact notation,

$$V = \operatorname{cav}[(1-\delta)u + \delta(V \circ f)],$$

with the obvious definitions. Well-known numerical methods exist to solve such a functional equation (*e.g.*, value iteration, the method used by Ely; but also policy iteration, linear programming, etc.). Perhaps structural properties of the optimal policy can be found in some interesting classes of models (supermodular preferences, first-order stochastic dominance on the state transitions).[25]

5.3 Design without commitment

Consider now the game between a firm and a regulator analyzed in Orlov, Skrzypacz and Zryumov (2016) (OSZ). The firm sells a product (say, a medical device) that is either safe or unsafe. While the product is on the market, the firm and the regulator observe noisy information about user experience, and update their beliefs about the safety of the product. The regulator has the power to recall the product from the market, at which point learning stops (or the reputation of the product is sufficiently tarnished that the firm cannot re-introduce it to the market without a major redesign). From the regulator's point of view, this is a real options problem. The firm and regulator incentives are partially aligned. They agree that only a safe product should be left on the market. They differ in their patience, or relative costs of type-one and type-two errors. Assume that the optimal stopping belief for the firm would be higher than for the regulator, so that the firm would like to experiment longer than the regulator. While the regulator has the decision power, the firm has control over the collection of additional information. Namely, there is a binary random variable that is correlated with the safety of the product and the firm can perform experiments to reveal information about that additional signal. The information collection technology is as before: the firm freely chooses a posterior distribution over the signal, subject to a martingale constraint. Importantly, whatever the firm learns, it has to disclose, even if it is *ex post* not in its best interest. As a result, even though information is valuable, the firm may decide not to learn some information (at least for a while) to persuade the regulator to wait a bit longer.

[25] By now, there are some papers deriving the optimal policy in some specific contexts. Smolin (2015), for instance, considers a related problem, with the twist that the agent's action affect transitions.

There are several differences between this model and the two discussed previously. First, the firm and the regulator have no commitment power. The paper analyzes Markov Perfect equilibria, with the joint belief about the safety of the drug and the realization of the noisy signal as state variables. Second, the sender's preferred action depends on the state. Third, the regulator's strategy is not a fixed belief cut-off, but a best-reply to the firm's information strategy. Finally, beliefs move stochastically in response to public news, even in the absence of disclosure by the receiver.

The resulting equilibrium dynamics depend on the size of the conflict between the firm and regulator. If the conflict is small (as measured by the difference in autarky threshold beliefs, that is, the optimal stopping beliefs when only public news are available), then the firm postpones learning about the additional signal until a belief is reached, at which bad news about the signal moves the posterior belief about the safety to the firm's optimal exercise point. At that belief it is an optimal strategy for the firm to fully reveal the additional information: if the news is bad, it gets its optimal stopping implemented; if it is good, the regulator postpones a recall as much as possible. In equilibrium, the regulator has *ex post* regret for waiting too long if the news is bad. But if the news is good, the posterior is strictly below his exercise threshold, so waiting is optimal *ex post*. In equilibrium, information disclosure is inefficiently late from the regulator's viewpoint. The firm manages to make the regulator wait (for some parameters, even pass his autarky threshold) because the discrete full revelation of the signal is credible and valuable to the regulator. In equilibrium, some recalls are forced by the regulator (after good news about the signal followed by bad public news) and some are voluntary (after bad signal realizations), so the firm is able to persuade the regulator into a form of a compromise.

When the difference in preferences between the two players is large, waiting for such full information disclosure is no longer sufficiently valuable to the regulator, and the equilibrium changes dramatically. When the regulator's belief reaches the autarky threshold, the firm "pipets" information about the signal, leading to either a reflection of beliefs at this threshold, or a discrete jump to reveal bad news. That is reminiscent of the greedy strategy above: the firm reveals the smallest amount of information to keep the regulator in the "no-recall" region, an equivalent of the "investment" region in the previous papers. Such persuasion is profitable for the firm (it manages to keep the product longer on the market), but somewhat paradoxically has no value for the regulator (his equilibrium payoff is the same as if no extra signal was available). The intuition comes from the real-options literature: the regulator's autarky value function satisfies smooth pasting condition at the exercise threshold. Since the firm's optimal strategy "pipets" information, it induces beliefs only on the linear part of the regulator's value function and that creates no value.

This stylized model can also reflect internal organization problems in which there is a conflict of interest about timing of exercise of a real option (as in Grenadier, Malenko, Malenko (2015) or Guo (2016)) and one agent has decision rights and another one can credibly disclose additional noisy information about the state. In some cases, the agent would like to exercise the option sooner (for example, if it is a decision when to launch a product). OSZ show that the agent then immediately reveals all information about the noisy signal and can be hurt by having access to the information. If the receiver expects full immediate information disclosure, it is rational for him to wait. Hence, an impatient sender is better off revealing the information immediately.

6 Richer Worlds

A significant shortcoming of all papers reviewed so far is the restriction to two independent arms. This is probably not the best way to think of the process of discovery and innovation. There are not simply two ways to go, but many, and trying out a new technique close to an existing one is

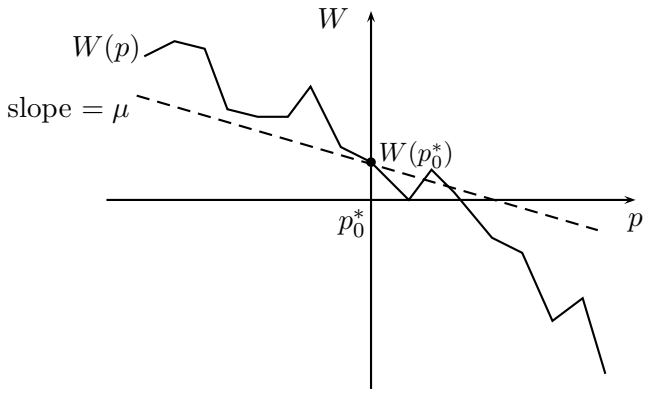


Figure 14: Products and possible outcomes.

likely to yield similar results. In contrast, trying out a very different technique is risky, but may pay big. Several models attempt to model such environments. These include Akcigit and Liu (2015), Bonatti and Rantakari (2016), Garfagnini and Strulovici (2016), Callander and Matouschek (2015). Here, we showcase Callander (2011).

Callander introduces the following model of learning by trial-and-error. There is a continuum of products that can be tried, $p \in \mathbf{R}$. Each product is mapped into some outcome, $W(p) \in \mathbf{R}$, and entails a cost $W(p)^2$.

At the start, the value of one and only one product is known, p_0^*, with value $W(p_0^*) > 0$. The player's belief about $W(\cdot)$ is modeled as a two-sided Wiener process. That is, we have two independent one-dimensional Wiener processes $W_- = (W_-(p) : p \geq 0)$ and $W_+ = (W_+(p) : p \geq 0)$, with parameters $(-\mu, \sigma)$ and (μ, σ) respectively, and initial value $W_-(0) = W_+(0) = W(p_0^*)$, and we define

$$W(p) = \begin{cases} W_+(p - p_0^*) & \text{if } p \geq p_0^*, \\ W_-(p_0^* - p) & \text{if } p < p_0^*. \end{cases}$$

It is assumed that $\mu < 0$. See Figure 14. Trying out a given p reveals the corresponding value of $W(p)$.

What values of p should be sampled? This is a difficult problem, already interesting when the agent is myopic (a sequence of entrepreneurs, each with one chance only), the focus of Callander's analysis. In a given round n, after history h^n, the entrepreneur chooses p so as to minimize $\mathbf{E}[W(p) \mid h^n]^2 + \mathbf{Var}[W(p) \mid h^n]$.[26] Let $\bar{p}, \underline{p}$ denote the highest and lowest value of p_m for $m < n$ along that history. Because $\mu \leq 0$,

$$p < \underline{p} \quad \Rightarrow \quad \mathbf{E}[W(p) \mid h^n]^2 + \mathbf{Var}[W(p) \mid h^n] > W(\underline{p})^2.$$

Hence, by induction, no value of $p < p_0^*$ is ever chosen. Note that the same argument implies that if $W(\bar{p}) \leq 0$, then choosing a product $p > \bar{p}$ is strictly worse than choosing $\bar{p}$. Hence, if a product p is ever chosen so that $p = \bar{p}$ and $W(p) \leq 0$, no product to the right of $\bar{p}$ will ever be sampled.

Next, notice that $p_m < p_{m'}$ are consecutive products (that is, no product in between their values has been chosen along that history), and $\operatorname{sgn} W(p_m) = \operatorname{sgn} W(p_{m'})$, then

$$p \in (p_m, p_{m'}) \quad \Rightarrow \quad \mathbf{E}[W(p) \mid h^n]^2 + \mathbf{Var}[W(p) \mid h^n] > \min\{W(p_m)^2, W(p_{m'})^2\}.$$

[26] This is because $\mathbf{E}[X^2] = \mathbf{E}[X]^2 + \mathbf{Var}[X]$.

Hence also, if $p_m < p_{m'}$ are tried products, with $\operatorname{sgn} W(p_m) = \operatorname{sgn} W(p_{m'})$, and all products $p \in (p_m, p_{m'})$ that were ever chosen are such that $\operatorname{sgn} W(p) = \operatorname{sgn} W(p_m)$. Then, if a product is ever chosen again in the interval $(p_m, p_{m'})$, it must be a known product, namely a product that minimizes the cost among those known products.

This leaves us with two cases: (i) trying a product $p > \bar{p}$, and (ii) trying a product in an interval $(p_m, p_{m'})$ of consecutive products such that $\operatorname{sgn} W(p_m) \neq \operatorname{sgn} W(p_{m'})$. Note that in that case it must be that $W(p_m) > 0 > W(p_{m'})$ (if one is 0 it is clearly optimal to choose it forever). This is because, by our previous observations, products to the right of a product with $\operatorname{sgn} W(p) = -1$ are never chosen. Hence, the only configuration that can arise in which there are tried products with both negative and positive values must be such that all products with negative values lie to the right of the products with positive value.

Choosing a product $p > \bar{p}$: Consider first trying a product at the "frontier," that is, larger than any product tried so far. As discussed, we may assume $W(\bar{p}) > 0$, otherwise such a choice cannot be optimal. We must solve

$$\min_{p \geq \bar{p}} \left\{ \mathbf{E}[W(p) \mid h^n]^2 + \mathbf{Var}[W(p) \mid h^n] \right\} = \min_{p \geq \bar{p}} \left\{ (W(\bar{p}) + \mu(p - \bar{p}))^2 + (p - \bar{p})\sigma^2 \right\},$$

clearly, a convex function in p, with first-order condition

$$\mu(p - \bar{p}) = -\frac{\sigma^2}{2\mu} - W(\bar{p}),$$

which is positive if and only if

$$W(\bar{p}) > -\frac{\sigma^2}{2\mu} =: \alpha.$$

(Recall that $\mu < 0$). Hence, such a choice is only a candidate to optimality if $W(\bar{p}) > \alpha$; if it is chosen, it is set so that

$$\mathbf{E}[W(p)] = W(\bar{p}) + \mu(p - \bar{p}) = \alpha.$$

In particular, this must be the optimal choice p_1^* in the first period $n = 1$ if $W(p_0^*) > \alpha$. (Interestingly, p_1^* is single-peaked in μ.) See left panel in Figure 15. Otherwise, $p_n = p_0^*$ for all $n \geq 1$, as the first outcome is "good enough."

Note that the cost of choosing this candidate is (plugging in the first-order condition into the objective)

$$2\alpha W(\bar{p}) - \alpha^2.$$

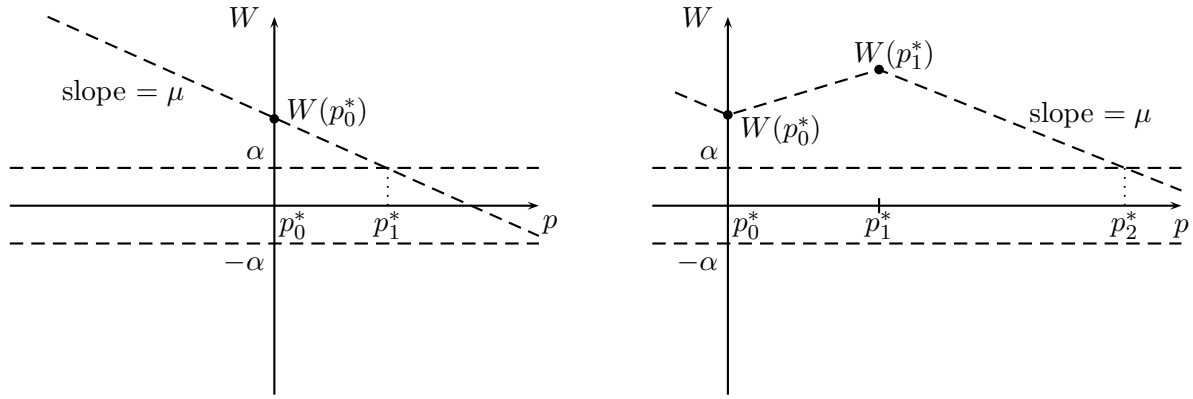
Hence the relevant comparison is whether this is larger than the cost of the best alternative so far. In particular, this choice is dominated by the product $p = \arg\min\{W(p_m)^2 : m = 0, \ldots, n-1\}$ if and only if

$$W(\bar{p}) > \frac{\bar{W}^2 + \alpha^2}{2\alpha},$$

where $\bar{W} = W(p)$.

Choosing a product $p \in (p_m, p_{m'})$, where $\operatorname{sgn} W(p_m) \neq \operatorname{sgn} W(p_{m'})$: Recall that we may assume that no product in the interval $(p_m, p_{m'})$ has ever been tried so far, and that $W(p_m) > 0 > W(p_{m'})$. This means that we can assume that

$$\frac{W(p_{m'}) - W(p_m)}{p_{m'} - p_m} < -\mu.$$



First period: p_1^* Second period: p_2^*

Figure 15: Exploration phase.

That is, the line connecting the points $(p_m, W(p_m))$ and $(p_{m'}, W(p_{m'}))$ is steeper than the drift line of the Wiener process. This is because of our earlier analysis: when $p_{m'}$ was chosen, it was picked either as the rightest product so far, in which case the resulting negative value (relative to a positive expected value of α) means that the slope is indeed steeper. Or it was not the rightest product, but then it was to the left of the first product that resulted in a negative value, meaning that the slope is even steeper.

We must minimize

$$\left(W(p_m) + \frac{p - p_m}{p_{m'} - p_m}(W(p_{m'}) - W(p_m))\right)^2 + \frac{(p - p_m)(p_{m'} - p)}{p_{m'} - p_m}\sigma^2,$$

which gives as candidate the solution to

$$\mathbf{E}[W(p) \mid h^n] = \mu(p_m, p_{m'})\left(1 - 2\frac{p - p_m}{p_{m'} - p_m}\right), \text{ where } \mu(p_m, p_{m'}) := -\frac{\sigma^2}{2\frac{W(p_{m'}) - W(p_m)}{p_{m'} - p_m}},$$

which is verified to improve (in expectations) on $p_{m'}$ if and only if

$$-W(p_{m'}) > \mu(p_m, p_{m'}).$$

Of course, one must additionally compare this expected payoff to the payoff $\bar{W}$ of the best product so far. The specific conditions are tedious, but with probability one, this process (of "triangulation") stops. See Figure 16.

To summarize, the game always starts with an exploratory phase, in which new products (to the right of the existing ones) are chosen. This process might stop, with the agent settling on the one with the best performance (which need not be the one furthest to the right). If at some point in the exploratory phase, a negative outcome results, the policy shifts to a triangulation phase, picking either a new product within the unique interval of products over which the realized values change sign, or the best product so far. Eventually, search stops.

It would be interesting to see how these dynamics change when entrepreneurs are forward-looking. Solving Callander's model for patient agents is an important open problem in this area.

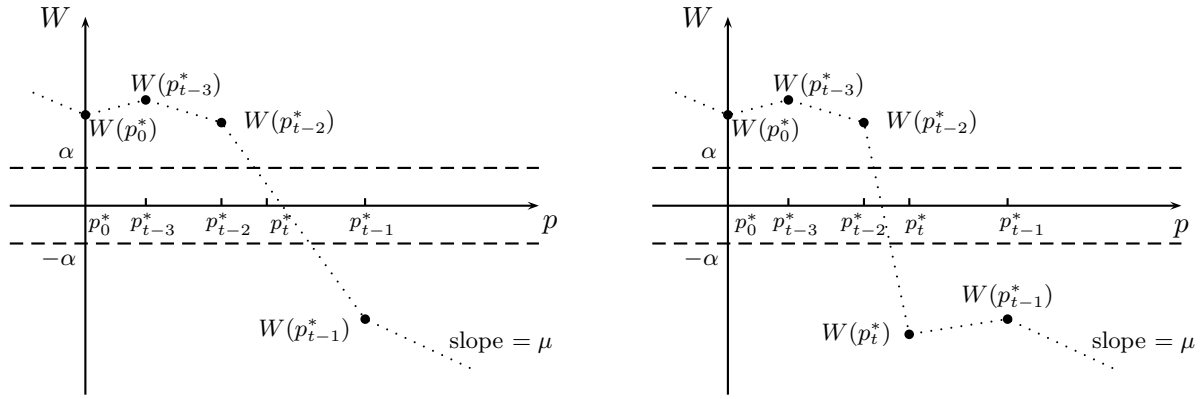


Figure 16: Triangulation phase, first and second step.

References

[1] Akcigit, U., and Q. Liu (2016). "The Role of Information in Innovation and Competition," *Journal of the European Economic Association*, **14**, 828–870.

[2] Aumann, R.J. and S. Hart (2003). "Long cheap talk," *Econometrica*, **71**, 1619–1660.

[3] Aumann, R.J., and M. Maschler (1995). *Repeated Games with Incomplete Information*, MIT Press.

[4] Bimpikis, K., S. Ehsani, and M. Mostagir (2016). "Designing Dynamic Contests?," working paper.

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

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

[7] Bonatti, A., and J. Hörner (2015). "Learning to Disagree in a Game of Experimentation," working paper, Cowles Foundation for Research in Economics.

[8] Bonatti, A., and H. Rantakari (2016). "The Politics of Compromise," *American Economic Review*, **106**, 229–259.

[9] Callander, S. (2011). "Searching and Learning by Trial and Error," *American Economic Review*, **101**, 2277–2308.

[10] Callander, S., and N. Matouschek (2015). "Managing on Rugged Landscapes," working paper, Stanford GSB.

[11] Campbell, A., F. Ederer, and J. Spinnewijn (2014). "Delay and Deadlines: Freeriding and Information Revelation in Partnerships," *American Economic Journal: Microeconomics*, **6**, 163–204.

[12] Che, Y-K. and J. Hörner (2015). "Optimal Design for Social Learning," working paper, Cowles Foundation for Research in Economics.

[13] Cohen, A., and E. Solan (2013). "Bandit Problems with Levy Processes," *Mathematics of Operations Research*, **38**, 92–107.

[14] Crawford, V., and J. Sobel (1982). "Strategic information transmission," *Econometrica*, **50**, 1431–51.

[15] Cripps, M., and C. Thomas (2015). "Strategic Experimentation in Queues," working paper, UCL.

[16] Das, K. (2015). "The Role of Heterogeneity in a Model of Strategic Experimentation," working paper, University of Exeter.

[17] Ely, J. (2015). "Beeps," working paper, Northwestern University.

[18] Frick, M., and Y. Ishii (2015). "Innovation Adoption by Forward-Looking Social Learners," working paper, Harvard.

[19] Garfagnini, U., and B. Strulovici (2016). "Social Experimentation with Interdependent and Expanding Technologies," *Review of Economic Studies*, forthcoming.

[20] Gordon, S., C. Marlats, and L. Ménager (2015). "Observation Delays in Teams," working paper, Paris.

[21] Gordon, S., E.B. de Villemeur (2010). "Strategic Advice on a Timing Decision," draft, Montréal.

[22] Grenadier, S., A. Malenko, and N. Malenko (2015). "Timing Decisions in Organizations: Communication and Authority in a Dynamic Environment," working paper, MIT.

[23] Guo, Y. (2016). "Dynamic Delegation of Experimentation," *American Economic Review*, **106**, 1969–2008.

[24] He, Johanna (2015). "Competition in Social Learning", working paper, Stanford University .

[25] Heidhues, P., S. Rady, and P. Strack (2015). "Strategic experimentation with private payoffs," *Journal of Economic Theory*, **159**, 531–551.

[26] Holmström, J. (1984). "On The Theory of Delegation," in *Bayesian Models in Economic Theory*, eds. Marcel Boyer and Richard Kihlstrom, North-Holland Publishing Co., Amsterdam.

[27] Hörner, J., N. Klein, and S. Rady (2015). "Strongly Symmetric Equilibria in Bandit Games," working paper, Yale University.

[28] Kamenica, E., and M. Gentzkow (2011). "Bayesian Persuasion," *American Economic Review*, **101**, 2590–2615.

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

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

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

[32] Khromenkova, D. (2015). "Collective Experimentation with Breakdowns and Breakthroughs," working paper, Mannheim University.

[33] Klein, N., and S. Rady (2011). "Negatively Correlated Bandits," *Review of Economic Studies*, **78**, 693–792.

[34] Kremer, I., Y. Mansour, and M. Perry (2014). "Implementing the "Wisdom of the Crowd"," *Journal of Political Economy*, **122**, 988–1012.

[35] Melumad, N., and T. Shibano (1991). "Communication in settings with no transfers," *RAND Journal of Economics*, **22**, 173–198.

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

[37] Murto, P., and J. Välimäki (2013). "Delay and information aggregation in stopping games with private information," *Journal of Economic Theory*, **148**, 2404–2435.

[38] Orlov, D., A. Skrzypacz and P. Zryumov (2016). "Persuading the Regulator To Wait," working paper, Stanford GSB.

[39] Rayo, L., and I. Segal (2010). "Optimal Information Disclosure," *Journal of Political Economy*, **118**, 949–987.

[40] Renault, J., E. Solan, and N. Vieille (2015). "Optimal Dynamic Information Provision," working paper, arXiv:1407.5649.

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

[42] Rosenberg, D., A. Salomon, and N. Vieille (2013). "On Games of Strategic Experimentation," *Games and Economic Behavior*, **82**, 31–51.

[43] Smolin, A. (2015). "Optimal Feedback Design," working paper, Yale University.

[44] Strulovici, B. (2010). "Learning while Voting: Determinants of Collective Experimentation," *Econometrica*, **78**, 933–971.

[45] Thomas, C. (2015). "Strategic Experimentation with Congestion," working paper, The University of Texas at Austin.