

Disentangling Exploration from Exploitation^{*}

Alessandro Lizzeri[†] Eran Shmaya[‡] Leeat Yariv[§]

November 21, 2025

Abstract. Starting from Robbins (1952), the literature on experimentation via multi-armed bandits has wed exploration and exploitation. Nonetheless, in many applications, agents' exploration and exploitation need not be intertwined: a policymaker may assess new policies different than the status quo; an investor may evaluate projects outside her portfolio. We characterize the optimal experimentation policy when exploration and exploitation are disentangled in the case of Poisson bandits, allowing for general news structures. The optimal policy features complete learning asymptotically, exhibits lots of persistence, but cannot be identified by an index à la Gittins.

Keywords: Exploration and Exploitation, Poisson Bandits

JEL codes: C73, D81, D83, O35

^{*}We thank Arjada Bardhi, Matthew Ellman, Francesco Fabbri, Nicolas Klein, Santiago Oliveros, Xiaosheng Mu, Bruno Strulovici, as well as the Editor and four anonymous reviewers for helpful comments. We gratefully acknowledge financial support from the National Science Foundation, grant SES 1949381.

[†]Princeton University and NBER; lizzeri@princeton.edu

[‡]State University of New York at Stony Brook; eran.shmaya@stonybrook.edu

[§]Princeton University, CEPR, and NBER; lyariv@princeton.edu

1 Introduction

In various applications, decision-makers navigate a dynamic landscape by simultaneously taking actions and gathering insights about their environment. Donors collect information about the value of charitable organizations while potentially supporting some. Hospitals monitor how different specialists perform, even as they refer patients only to a chosen few. Individuals navigate their career paths by exploring opportunities within their organization or beyond.

The seminal work of Robbins (1952), Gittins (1979), and Gittins and Jones (1979), proposed a dynamic model that integrates learning with decision-making. In their classical multi-armed bandit problem, each action taken by an agent provides information solely about that specific action’s effectiveness. In practical applications, this framework assumes that, for instance, policymakers learn only from the policies they implement, investors gather insights exclusively from the returns of the stocks they currently hold, and employees evaluate their career prospects based solely on their experiences within the firms where they work. Like a bet on a slot machine, optimal choices balance the benefits of learning about the action (exploration) and its consequent payoff benefits (exploitation).

We propose a framework for studying settings in which exploration and exploitation are, and can be, disentangled. We depart from the slot machine model, and instead consider decision-makers who can learn about choices they might not immediately pursue. We characterize the resulting optimal policy and illustrate when the ability to disentangle exploration from exploitation is especially advantageous.

In our model, an agent encounters a recurring decision between two uncertain projects. These projects might be policies, stocks, job prospects, etc. To simplify, we assume that each project offers either a positive flow payoff if successful (good project) or no payoff if unsuccessful (bad project). The quality of each project is determined independently at the outset, with prior probabilities known to the agent.

In each moment, the agent decides which project to exploit; that is, which policy to implement, which investment to make, which job to choose, and so on. Her choices determine the overall payoff, calculated as the discounted sum of rewards obtained from exploitation. The agent learns incrementally throughout the process: at each moment, she possesses a unit of attention, or exploration, which she allocates between the two projects. The agent gains information about a project only if she pays some attention to it. In contrast to the traditional multi-armed bandit framework, our model’s operating assumption is that the agent can gather information through exploration, which is not necessarily tied to exploitation. It captures a wide range of environments: cases where acquiring information requires effort distinct from taking action (for example, researching securities or

evaluating policies), scenarios where payoffs materialize only in the long run (for instance, charitable giving or basic research funding), and situations where information comes from external or aggregate sources (say, hospital referrals or safety of commuting routes). It also fits settings where exploration and exploitation are carried out by different but aligned actors, such as intelligence agencies informing executive decisions or committees advising institutional leaders.

Specifically, when exploring a project, the agent can get conclusive information about its quality, which arrives at a Poisson rate. The arrival rate may differ depending on the explored project and whether it is good or bad. We consider general good news settings, where good news arrives faster than bad news, and receiving no news makes the agent increasingly pessimistic. We also consider general bad news settings, where bad news arrives more rapidly, and receiving no news makes the agent increasingly optimistic.

In many real-world scenarios, it may be reasonable to assume that the exploited project yields some valuable data automatically. To accommodate this possibility, we consider a spectrum of entanglement levels. An entanglement level α specifies the predetermined portion of exploration that must be allocated to the exploited project. When $\alpha = 0$, exploration and exploitation are entirely disentangled. Conversely, when $\alpha = 1$, exploration and exploitation are fully entangled and our environment admits several settings studied in the literature as special cases. Specifically, when news arrives only about good projects at positive rates, this aligns with the Keller, Rady, and Cripps (2005) (KRC) setting. When news arrives exclusively about bad projects at positive rates, this mirrors the Keller and Rady (2015) (KR) setting.¹

We first highlight a general long-run property of disentangled experimentation that starkly distinguishes our setting from the classic bandit model. Whenever some portion of exploration can be dedicated to an unexploited project ($\alpha < 1$), we show that an optimizing agent exploits the *realized* best project asymptotically. Intuitively, if there is any room for the agent to be swayed by information toward exploiting a different project than the one she already exploits, under any level of disentanglement she would obtain that information in the long run. The asymptotic optimality in our setting underscores a fundamental difference from the conventional setting, with full entanglement, where it is well-known that the agent’s exploitation need not converge to the ex-post optimal project.

We then turn to the characterization of the optimal policy and its consequences. We begin with a benchmark model where the exploration choice is straightforward and study how disentanglement affects optimal exploitation. Specifically, we assume that one project is known to be good, and therefore safe, as in KRC and KR, although we allow for Poisson

¹While special, these settings have been used to study a variety of applications, including delegation problems (Hörner and Samuelson, 2013; Guo, 2016), experimentation by committee (Strulovici, 2010), dynamics of discrimination (Bardhi, Guo, and Strulovici, 2020), and many others; see our literature review.

arrival rates of both good and bad news. In this case, the agent explores the uncertain, or risky, project as much as possible. With any level of positive entanglement ($\alpha > 0$), the agent’s exploitation choices constrain her exploration. The less entangled the agent, the looser the link between the payoffs she achieves in the short run and the information she obtains. She thus faces a variant of the standard exploration/exploitation dilemma. The optimal policy is characterized by a threshold on the posterior probability of the risky project’s favorability. When this threshold is surpassed, the agent chooses to exploit and further explore the risky project; otherwise, she exploits the safe project, minimally exploring it using the predetermined portion of her attention budget.

In Proposition 1, we demonstrate that the optimal threshold depends only on the maximum between the arrival rates of good and bad news. In either good or bad news settings, the analytical description of the optimal threshold is identical. The optimal policy exhibits different features, naturally. In particular, as in KRC and KR, with a high enough initial prior that the risky project is good, absent news arrival, the agent ultimately switches her exploitation in good news settings, but never does so in bad news settings. We then show that the value of disentanglement is higher under good news than under bad news, and that it responds non-monotonically both to the discount rate and to the prior probability that the risky project is good.

We then turn to the analysis of two projects with uncertain returns. To illuminate the forces within our model, we focus on scenarios where exploration and exploitation are entirely disentangled. This assumption implies that the agent optimizes exploitation by favoring the myopically optimal project at any given moment. However, determining the optimal exploration strategy is less straightforward and does not adhere to an index policy akin to Gittins’. The nature of the optimal policy depends on the structure of the news arrival process.

We start by examining balanced news settings, where both good and bad news arrive at equal rates for each project. The characterization of the optimal policy in this benchmark case provides the foundation for deriving the solution in the more intricate settings with good and bad news. With balanced news, the passage of time without any news does not provide any insight into the quality of a project. Consequently, the optimal policy is independent of the time elapsed without news, and the primary consideration is which project to explore at the outset.

In Proposition 2, we show that the optimal exploration strategy is determined via the comparison of a particular formulation of the information value associated with each project. This value is influenced not only by the rates at which news arrives but also by the relative rewards and prior probabilities assigned to each project’s success. Our characterization highlights two key deviations from the optimal policy observed in the traditional,

fully entangled environment. First, the optimal choice of which project to explore and to exploit is independent of the discount rate. Second, the optimal exploration strategy depends on the interplay between the parameters of both projects and, as noted, cannot be described via a separable index.

Moving to general good news settings, Proposition 3 shows that the agent still optimally exhibits a lot of persistence in her exploration. Absent news, the agent switches which project she explores at most once. This switch occurs only if the initially explored project aligns with the myopically optimal one.

This outcome is rooted in a fundamental principle of information economics: valuable information is actionable and influences which project is exploited. In general, actionable information manifests in two forms, either adverse news regarding the exploited project, or favorable news concerning the alternative project. To glean intuition for the persistence of optimal exploration, consider pure good news settings, where only good news arrives at a positive rate about either project, as in KRC. In such settings, in the short run, actionable information materializes only through positive news about the unexploited project. If the agent explores the unexploited project, absent news, she becomes increasingly pessimistic about it. Consequently, she has no incentive to switch either her exploited project or her explored project.

The optimality of persistent exploration starkly contrasts with predictions derived from the classical, fully entangled environment. In the classical good news setting, as the agent explores and exploits a project, her confidence in its potential diminishes gradually, leading to a reduction in its corresponding Gittins index. Eventually, the indices for both projects coincide, prompting the agent to alternate between the projects absent news, hence switching infinitely often. Subsequently, upon receiving positive news about either project, the agent indefinitely explores and exploits that project, effectively terminating further information gathering. In particular, with some probability, the agent ultimately exploits the project deemed inferior ex-post.

In general bad news settings, Proposition 4 illustrates that the structure of optimal exploration depends on the ranking of the maximal rewards the projects generate. Once the agent embarks on exploring the high-reward project, she continues indefinitely unless information arrives. Furthermore, in the absence of news, the agent inevitably explores the high-reward project at some point. Thus, similar to the dynamics observed in good news settings, the agent may switch her exploration at most once without news arrival.

To gain intuition, consider pure bad news settings, where only bad news arrives at positive rates, as in KR. In such settings, when the agent explores the high-reward project and no news is received, her confidence in the project progressively grows. Only negative news regarding that project would induce her to switch her exploited project. Therefore, it

remains optimal to continue exploring the high-reward project. One might question why the same logic wouldn't apply to the low-reward project. For the low-reward project, even if the agent maintains a sufficiently optimistic outlook, positive information about the high-reward project could still sway her exploitation choice. The only means of acquiring such information is by exploring the high-reward project for an extended period.

The distinction from the classical environment hinges on the nature of news arrival. In good news settings, the absence of news over time leads the agent to explore a project different from the one she exploits: disentanglement can be important even in the long run. By contrast, in bad news settings, exploration and exploitation ultimately converge on the same project, so disentanglement is relevant only in the short run.

In the settings we consider, the payoff benefits of disentanglement over entanglement are most pronounced when parameters fall within intermediate ranges: the discount rate, arrival rates of news, and initial beliefs regarding the viability of the projects under consideration. Collectively, our results show that when information and actions occur in sync, the ability to disentangle the two not only impacts behavioral predictions, but carries important implications for potential payoffs.

Our framework also provides a useful foundation for analyzing applications traditionally studied through the classical model. To illustrate, we examine the consequences of disentanglement in the context of a team problem, following [Bolton and Harris \(1999\)](#) and KRC. In our formulation, n homogeneous agents independently allocate a unit endowment of both exploration and exploitation across two common projects. As in KRC, one of the projects is safe and agents observe the outcomes of all exploration efforts, both their own and others', but benefit only from their own exploitation. This creates incentives to free ride on others' exploration, as agents can appropriate its informational benefits without bearing the associated costs. We show that such free-riding incentives persist under any positive degree of entanglement, while full efficiency obtains only under complete disentanglement. Under partial entanglement, the extent of free riding depends on the parameters and can vary non-monotonically in the degree of entanglement and the prior that the risky project is good.

Related Literature The multi-armed bandit problem was likely initially posed by [Thompson \(1933\)](#) in the context of clinical trials. Starting from [Robbins \(1952\)](#), the statistics literature has offered insights on the features of optimal policies. [Gittins \(1979\)](#) and [Gittins and Jones \(1979\)](#) present the first general index-based optimal policies. [Gittins, Glazebrook, and Weber \(2011\)](#) offer a survey of ensuing results. As already noted, the special case of Poisson bandits was introduced by [Keller et al. \(2005\)](#) (KRC) and [Keller and Rady \(2015\)](#) (KR), assuming two arms, only one of which yields uncertain rewards.

The basic multi-armed bandit setting has been utilized for a wide array of applications in economics, ranging from monopoly pricing decisions (Rothschild, 1974), to labor market choices and matching (Jovanovic, 1979; Miller, 1984), to venture capital (Bergemann and Hege, 1998), to the design of recommender systems (Che and Hörner, 2018), to team experimentation (Bolton and Harris, 1999; Strulovici, 2010, in addition to KRC and KR); for a survey, see Bergemann and Valimaki (2006).²

Our paper also relates to the literature on dynamic information acquisition, initiated by Wald (1947). In the most basic model, an agent can acquire costly signals in sequence, and determine when to stop information collection and take a decision. In our setting, the cost of exploring one project is the option value of exploring the other. Unlike the classical model, the cost is therefore changing and endogenous. Furthermore, while our setting is dynamic, it does not correspond to a stopping problem per se. The idea that payoffs are generated by actions that are unobservable, so that the decision needs to rely on information obtained elsewhere appears in Carnehl and Schneider (2023) and Georgiadis-Harris (2024), although they focus on different payoff structures and consider choices that occur before an exogenously determined time. Damiano, Li, and Suen (2020) extend the KRC setting with one safe project and pure good news by allowing the agent to *pause* exploration and exploitation at any time and collect auxiliary news at a cost. They characterize how and when auxiliary information is used. The closest connection to our paper lies in the discussion of the payoff consequences of disentanglement in the special case of one safe project and good news, which we present in Section 3.2. Our conclusions differ due to the distinct information costs our model implies. Our analysis of two risky projects in Sections 4 and 5, as well as our characterization result for one safe project in Proposition 1, where we allow for both general good and bad news, has no counterpart in their paper.

The idea that decision makers may be able to attend to or acquire information only up to a limit also appears in the rational inattention literature, see Sims (2003) and Maćkowiak, Matějka, and Wiederholt (2023)’s survey. Recent work considers dynamic attention allocation. For example, Che and Mierendorff (2019) consider an environment à la Wald (1947)—a stopping problem—in which a decision maker acquires information from different news sources, each providing conclusive news about the underlying state at a Poisson rate, prior to making an irreversible binary decision. Since the rates at which news arrives from either source may depend on the underlying state, the optimal policy balances the speed at which either news source delivers news and its “bias,” a trade-off different than

²The analysis in Che and Hörner (2018) relates to the special case of one safe project in our environment, which we discuss in Section 3. Eliaz, Fershtman, and Frug (2024) consider an extension of the basic model, where bandits—or tasks, in their framework—evolve when attended to, and payoffs also depend on unattended tasks. There is also recent empirical work that uses the basic multi-armed bandit setting in the context of pharmaceutical demand and physician prescribing behavior (see Crawford and Shum, 2005; Currie and MacLeod, 2020; Dickstein et al., 2021) and in the context of research and development (Zhuo, 2023).

the one underlying our agent’s problem. [Liang, Mu, and Syrgkanis \(2022\)](#) also study a variation of the Wald problem, where a decision maker allocates a fixed attention budget across multiple sources of information to learn about a decision-relevant state. Information sources are diffusion processes whose unknown drift is an attribute that contributes linearly to determine the state. In the optimal policy, the decision maker initially allocates all attention to the most informative source, then gradually incorporates additional sources until, eventually, attending to all sources.

There is also a literature in computer science that takes an algorithmic approach to identifying which arm is most desirable in a multi-armed bandit problem. [Bubeck, Munos, and Stoltz \(2011\)](#) is perhaps the most conceptually related paper. They focus on regret-minimizing exploration algorithms. There is no simultaneous exploitation, and the objective is the difference between the average payoff of the best arm and the average payoff obtained by the algorithm’s recommendation. See also [Audibert, Bubeck, and Munos \(2010\)](#) and the literature that followed.

2 The Model

Our Framework An agent allocates exploration and exploitation resources between two projects, L and H , in continuous time. Project $z = L, H$ is good with probability p_z and bad with the complementary probability $1 - p_z$. The quality of the projects is determined independently at the outset and persistent. The resulting rewards from project z , denoted by \tilde{R}_z , are random variables taking values in $\{0, R_z\}$. If project z is good, it pays a flow reward of $R_z > 0$; If project z is bad, it pays 0 forever. We assume $R_H > R_L > 0$. We also assume that $p_L, p_H > 0$ and $p_H < 1$ so that there is meaningful uncertainty about which project is superior.³

As in KRC, we assume that the agent has a unit of investment to allocate, capturing the exploitation aspect of the agent’s choice. At any moment, the agent’s instantaneous reward from investing $k_z \geq 0$ in exploiting project $z = L, H$ is given by:

$$k_L \tilde{R}_L + k_H \tilde{R}_H,$$

where $k_L + k_H = 1$ and \tilde{R}_z denotes the realized rewards from project $x = L, H$.⁴ We assume the agent’s exploitation policy is measurable with respect to the information available at any time. As is standard, payoffs are discounted at a fixed rate $r > 0$. Thus, if the agent

³The analysis is unchanged if, instead, we assume that a good project delivers lump-sum payoffs of R_z/μ at some Poisson rate μ .

⁴We later show that, for most of our analysis, the agent optimally chooses $k_z \in \{0, 1\}$ for $z = L, H$. We maintain this greater generality in order to contrast some of our results with the classical, fully entangled setting, where interior investments are sometimes utilized in the optimal policy.

invests k_z^t in exploiting project $z = L, H$ at time t , her resulting overall payoffs are given by:

$$\mathbb{E} \int_0^\infty r e^{-rt} (k_L^t \tilde{R}_L + k_H^t \tilde{R}_H) dt,$$

where \tilde{R}_L and \tilde{R}_H correspond to the realized rewards from project L and H , respectively. Importantly, the agent does not observe the rewards generated by her own exploitation, an assumption we soon discuss.⁵

Analogously, at any moment, the agent allocates a unit budget of attention, or information collection resources, across the projects. This is the exploration aspect of the agent's choice. If the agent spends a fraction $\alpha_z > 0$ of her attention budget exploring project $z = L, H$, she may receive conclusive news about project z . Specifically, if project z is good, the agent receives good news—a conclusive signal indicating that the project is good—at a Poisson rate $\alpha_z \lambda_z^g$ (and no news otherwise). Similarly, if project z is bad, the agent receives bad news—a conclusive signal asserting the project is bad—at a Poisson rate $\alpha_z \lambda_z^b$. We assume $\max\{\lambda_z^g, \lambda_z^b\} > 0$ and that $\text{sign}(\lambda_H^g - \lambda_H^b) = \text{sign}(\lambda_L^g - \lambda_L^b)$, with the convention that $\text{sign}(0) = 0$. That is, the agent has opportunities to learn and the information structure is similar across the two projects. The agent's exploration policy is also assumed to be measurable with respect to the information available at any time.

Whenever $\lambda_z^g - \lambda_z^b > 0$ for $z = L, H$, good news arrives at a higher rate than bad news. We refer to such environments as *good news settings*. Absent any news, the agent becomes increasingly pessimistic: no news is bad news. A special case corresponds to the frequently studied good news setting of KRC, which we term *pure good news*, where $\lambda_z^g > 0$ and $\lambda_z^b = 0$ for $z = L, H$. Conversely, whenever $\lambda_z^b - \lambda_z^g > 0$ for $z = L, H$, bad news arrives at a higher rate than good news. We refer to such settings as *bad news settings*. Absent any news, the agent becomes increasingly optimistic: no news is good news. A special case corresponds to the frequently studied bad news setting of KR, which we term *pure bad news*, where $\lambda_z^b > 0$ and $\lambda_z^g = 0$ for $z = L, H$. We refer to settings in which good and bad news arrive at precisely identical rates, $\lambda_z^g = \lambda_z^b$ for $z = L, H$, as *balanced news settings*. In balanced news settings, without the arrival of news, the agent's posterior belief that the explored project is good does not change. These settings are useful as central benchmark cases around which we construct some of our proofs.

Distinction between Exploration and Exploitation We assume that information about project quality is attainable independently of the agent's exploitation choices.

It is instructive to contrast this with the classical model. In the classical framework, the only signals about project value are the realized payoffs, which are observed exclusively

⁵Our model is equivalent to one in which the agent makes a single allocation decision at a random time, exponentially distributed with rate r , in which case no relevant feedback from payoffs is available.

when the project is exploited. The canonical example is the slot-machine (bandit) problem, where beliefs about success probabilities are updated solely from the frequency of past rewards. Of course, in general, information need not be immediately payoff-relevant. The critical assumption in the classical model is that any information about the quality of the project becomes available only when that project is exploited. While our assumption of non-observability of payoffs is certainly an extreme assumption, it is arguably just as extreme as the assumption in the classical model that *all* the information available is generated exclusively via exploitation.

Our model is designed to capture environments in which information about a project's value can be acquired independently of exploiting it. The model serves as a useful approximation across various settings. First, in many environments, acquiring information requires deliberate effort distinct from taking action. For example, investing in a security does not directly generate informational benefits beyond those obtained through research. When facing multiple securities, an agent allocates attention, rather than capital, to learn about their underlying values. It is this attention that produces information, not the act of investing itself.⁶

Second, in some domains, payoffs materialize only in the long run. Consider charitable giving: donors must investigate the operations of different nonprofits to assess impact. The act of donating, on its own, yields little short-run feedback.⁷ Similarly, funding agencies that support basic research may gain little short-term insight into a project's eventual value. Independent evaluations, conducted regardless of whether a project is funded, can generate informative signals for future funding decisions.

Finally, in many applications, information is only available through external or aggregate sources. A hospital referring patients to specialists may receive noisy or limited individual feedback, and thus rely on aggregate performance data, whether about past or potential providers, to guide future referrals. Likewise, a commuter experimenting with different walking routes gains little insight into neighborhood safety from a single uneventful trip. Instead, crime statistics across routes offer a more reliable basis for behavioral adjustment.

Our assumption is also relevant in settings where different agents are responsible for exploration and exploitation, a separation that arises naturally in many real-world contexts. For instance, intelligence agencies collect information to guide decisions made by members of the executive branch, such as the president. Likewise, hiring and promotion committees in academic departments evaluate candidates and present recommendations that inform decisions ultimately made by chairs or deans. Our model captures environ-

⁶Much of the relevant information concerns future events, such as earnings growth, that shape expectations about returns and are unrelated to past payoffs.

⁷Indeed, tools like <https://www.charitynavigator.org/> help donors assess organizational effectiveness.

ments in which the agent responsible for exploration and the one responsible for exploitation have aligned preferences.⁸

Partial Disentanglement Certainly, in many applications, rewards from exploitation choices do provide some information about the quality of the undertaken projects. For example, in policy settings, policymakers can actively engage—by studying reports, consulting expert committees, or tracking media coverage—to learn about the performance of existing or proposed policies. However, implemented policies often generate more immediate outcome statistics.

In order to capture such environments, as well as relate the commonly utilized exploration/exploitation model to ours, we consider the α -constrained decision process. In the α -constrained decision process, whenever the agent exploits a fraction $k_z \in [0, 1]$ in project $z = L, H$, she must allocate at least αk_z to exploring it: $\alpha_z \geq \alpha k_z$. When $\alpha = 1$, the agent must explore the project she exploits, leading to the standard exploration/exploitation trade-off. When $\alpha = 0$, exploration and exploitation are fully disentangled.

When $\alpha > 0$, the agent receives some information automatically about the project she exploits. She still has a budget of attention, which we set at $1 - \alpha$ for presentation simplicity, that she can allocate between the projects as before.⁹

General Approach Throughout, we characterize optimal policies up to measure-zero sets of time. Our analysis draws on dynamic programming principles but departs from the standard use of the Hamilton-Jacobi-Bellman equation in experimentation problems. We provide a new methodology that accommodates general environments with both projects being risky, where the state space is two dimensional.

Asymptotic Optimality We begin with a straightforward result that highlights the fact that the option to disentangle exploration from exploitation, corresponding to any $\alpha < 1$, has important implications on outcomes.

Proposition 0 (Asymptotic Optimality). *For all $\alpha < 1$, the agent exploits the best project asymptotically.*

⁸In Section 6 of our Online Appendix, we discuss some implications of heterogeneous preferences in such agency problems.

⁹Assuming the agent has a unit budget of attention as before would amount to a re-normalization and would not affect most of our analysis or insights.

Proposition 0 offers a fundamental contrast between our environment and the standard setup, where it is well known that the agent’s exploitation need not converge to the ex-post optimal project.

The proof of Proposition 0 holds for any number of projects and any payoff process. To prove this result, we need to show that the agent will eventually explore projects for a sufficiently long time so as to learn to exploit the best one. However, an impatient agent might prefer an exploration strategy that is more efficient in the short run. Assume, for instance, that p_L is close to 1, while p_H , λ_H^g , and λ_H^b are low. In the long run, the agent benefits from exploring project H . In the short run, exploring project H is not useful because, in expectation, it would take a long time to conclude that project H is good with sufficient likelihood to exploit it. In fact, in the classical environment, if project L is known to be good, a sufficiently impatient agent would never learn that project H is good as well. With $\alpha < 1$, the impatient agent may still explore project L initially: if λ_L^b is sufficiently high, the agent might initially explore project L since bad news will lead her to switch her exploited project. However, as we show in the proof, at some point, the short-run benefit from continuing to explore project L diminishes enough so that even an impatient agent will prefer to explore project H .

Proposition 0 also underscores the importance of our assumption that the agent is long-lived. If we replace our agent with a sequence of short-lived agents, each of whom lives for a fixed duration, then it may be that they all prefer to explore project L since neither is around long enough to benefit from exploring project H . Liang and Mu (2020) call this phenomenon a *learning trap*.

3 The Special Case of One Safe Project

As already noted, a heavily studied exploration/exploitation setting is that introduced by KRC, where project L is “safe:” $p_L = 1$. This environment is used in many applications and is a special case of our environment. It is an ideal starting point for our analysis, partly because of its familiarity, and partly because its simplicity allows us to introduce key ideas in our approach. These ideas will prove particularly helpful when we analyze the case of both projects being risky, where the optimal exploration choice becomes more nuanced.

3.1 Optimal Policy with a Safe Project

With one safe project, the choice of exploitation fully determines the optimal exploration policy. Since uncertainty is present only for project H , the agent always allocates at least

$1 - \alpha$ units of attention to exploring project H .¹⁰ The choice of exploitation determines payoffs as well as how the remaining α units of attention are allocated. This choice is affected by α and, as we soon show, there is some cutoff posterior $\bar{p}(\alpha)$ above which the agent exploits project H and below which she exploits project L . A myopic agent would exploit project H when it has a higher expected value, hence her cutoff posterior would be $p_M = \frac{R_L}{R_H}$. With $\alpha = 0$, even a patient agent exploits project H whenever it is myopically optimal, namely when $p_H \geq p_M$. When $\alpha > 0$, however, exploiting project H garners an informational advantage as it allows the agent to explore project H and learn at higher rates: she can dedicate her full attention to project H instead of only a fraction $1 - \alpha$ of it. The agent may choose to exploit project H at even lower posteriors than p_M , an instance of the exploration/exploitation trade-off. The following proposition characterizes the optimal exploitation strategy.¹¹

Proposition 1 (One Safe Project: Optimal Exploitation). *Let $\lambda = \max\{\lambda_H^g, \lambda_H^b\}$. For any $\alpha \in [0, 1]$, the agent optimally exploits project H whenever her posterior that project H is good exceeds $\bar{p}(\alpha)$, where*

$$\bar{p}(\alpha) = \frac{(r + \lambda(1 - \alpha))R_L}{(r + \lambda)R_H - \lambda\alpha R_L}.$$

The cutoff $\bar{p}(\alpha) \leq \frac{R_L}{R_H}$ is decreasing in α and R_H/R_L , and increasing in r . When $\alpha > 0$, it is decreasing in λ .

Although the cutoff $\bar{p}(\alpha)$ does not depend on whether good news or bad news arrive at higher rate, provided the maximal news arrival rate λ remains constant, the optimal policy differs between the two settings. In good news settings, if no news arrives, any amount of exploration of project H leads the agent to grow increasingly pessimistic about project H . If the agent starts by exploiting project L , she switches to exploiting project H only upon receiving good news. If the agent starts by exploiting project H , after a sufficiently long time without news, the agent becomes so pessimistic about that project that she switches to exploiting project L . In contrast, in bad news settings, if no news arrives, any amount of exploration of project H leads the agent to grow increasingly optimistic about project H . Therefore, for any $\alpha < 1$, if the agent starts by exploiting project L , absent bad news, she

¹⁰The results are the same if there is an exogenous baseline arrival rate of news on the risky project that is independent of the exploited project, where the exploitation decision generates additional information. Moreover, since only project H is worth exploring, the disentanglement level α matters only during exploitation of project L . Thus, the analysis remains unchanged with project-specific disentanglement levels.

¹¹The result is essentially implied by a combination of [Che and Hörner \(2018\)](#)'s Proposition 1 and the proof of Proposition 7 in Section D.5 of their Online Appendix, although they study a different environment and a different set of questions. Our method of proof is different and, we believe, instructive.

switches to exploiting project H at some point. If she starts by exploiting project H , she never switches unless bad news arrives.

The KRC and KR cutoffs correspond to $\bar{p}(1)$. As α decreases, the link between exploration and exploitation is relaxed and $\bar{p}(\alpha)$ increases toward the myopic cutoff p_M . When $\frac{R_H}{R_L}$ increases, gaining information on whether project H is good becomes more valuable and the cutoff $\bar{p}(\alpha)$ moves away from p_M . Last, as λ increases, exploration of project H becomes more appealing as it is expected to yield a conclusive signal more quickly. Again, the optimal cutoff $\bar{p}(\alpha)$ moves away from p_M .

In order to glean intuition for the derivation of the optimal cutoff, consider a good news setting. For any posterior p such that $pR_H \geq R_L$, it is certainly optimal for the agent to exploit project H : it generates higher expected payoffs and delivers more information. Assume then that $pR_H < R_L$. Call σ_L the strategy that specifies exploiting project L until news, and σ_Δ an alternative strategy that prescribes exploiting project H for a short time interval $\Delta > 0$ before returning to exploiting project L in the event that there is no news. The difference in payoffs between these two strategies is given by:

$$-\Delta r(R_L - pR_H) + (1 - \Delta r)p\lambda\Delta\alpha \frac{r}{r + (1 - \alpha)\lambda} (R_H - R_L) + O(\Delta^2). \quad (1)$$

The first term in equation (1) is the expected flow payoff difference between exploiting projects L and H . The second term is the expected discounted present value of information that reflects the possibility that, in the time interval Δ , the agent receives good news and optimally switches to exploiting project H . The arrival rate of bad news appears only in a term corresponding to the flow payoff during the interval of length Δ if bad news is received from project H : the agent already intends to switch back to project L absent news. Since the probability of such news, when project H is bad, is $O(\Delta)$, the corresponding term is $O(\Delta^2)$. At the cutoff $\bar{p}(\alpha)$, taking limits as $\Delta \rightarrow 0$, our proof illustrates that the expression in equation (1) approaches 0. This yields the formula appearing in Proposition 1.

An analogous construction holds for bad news settings, and the resulting cutoff depends on the maximal arrival rate for both good news and bad news settings. In particular, the cutoff corresponding to $\lambda_H^i > \lambda_H^j$, where $i, j \in \{g, b\}$ is the same as the cutoff corresponding to a setting with λ_H^i and $\lambda_H^j = \lambda_H^i - \epsilon$, with $\epsilon > 0$ as small as desired. It follows that the cutoff corresponding to a good news setting with good news arriving at a rate of λ is the same as the cutoff for a balanced news setting with arrival rate of λ . Similarly, the cutoff corresponding to a bad news setting with bad news arriving at a rate of λ is also the same as the cutoff for a balanced news setting with arrival rate of λ . Thus, the cutoff formulas for both good and bad news settings must coincide.

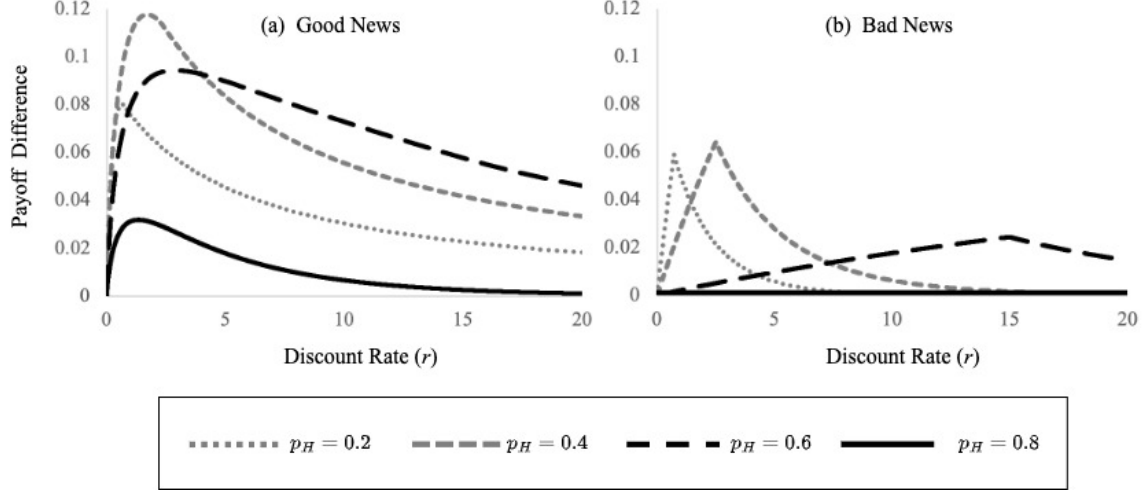


FIGURE 1: Payoff value of disentanglement for (a) pure good news settings, and (b) pure bad news settings when $R_L = 10$, $R_H = 15$, and $\lambda_H = 5$

3.2 Payoff Consequences of Disentanglement

Relaxing the entanglement constraint by reducing α can only improve the agent's expected payoff. We now identify features of the environment that make disentanglement particularly valuable.

Certainly, when R_H/R_L increases, the benefits of learning without forgoing payoffs are larger. Therefore, the value of disentanglement increases in R_H/R_L . In what follows, we inspect the dependence of payoffs on other parameters.

For any project rewards R_L and R_H , denote by $\Pi(p_H, r/\lambda; \alpha)$ the agent's expected payoff for the environment's parameters, an analytical formulation of which appears in the Appendix. To quantify the impacts of disentanglement, we first focus on the two extreme cases, $\alpha = 0$ and $\alpha = 1$, and consider the normalized payoff difference:

$$\Delta\Pi(p_H, r/\lambda) = \frac{\Pi(p_H, r/\lambda; 0) - \Pi(p_H, r/\lambda; 1)}{p_H R_H + (1 - p_H) R_L},$$

where the denominator represents the ex-ante value of the full information payoff and serves as a normalization factor. In Figure 1, we depict $\Delta\Pi(p_H, r/\lambda)$ for various parameters, focusing on the pure good and bad news settings, where $\lambda_H = \max\{\lambda_H^g, \lambda_H^b\}$ and $0 = \min\{\lambda_H^g, \lambda_H^b\}$.

As can be seen, the benefit of disentanglement is non-monotonic with respect to the discount rate r , and equivalently, with respect to the arrival rate λ of good news. Intuitively, when the agent is very patient ($r \rightarrow 0$) or when news arrives rapidly ($\lambda \rightarrow \infty$),

regardless of α , the agent can accumulate information with no substantial payoff consequences. Even in the classical environment, the agent may suffer payoff losses because she exploits the risky project for a long time, but the payoff consequences are minimal when the agent is very patient. The benefit of disentanglement is therefore small. When the agent is very impatient ($r \rightarrow \infty$) or when news arrives slowly ($\lambda \rightarrow 0$), short-run, or myopic payoffs approximate the agent's payoffs regardless of the level of disentanglement, which hence has little impact. It follows that the payoff consequences of disentanglement can be meaningful only for intermediate values of r/λ .

As Figure 1 illustrates, the effects of p_H are also non-monotonic. Consider first good news settings (depicted in the left panel). Suppose $p_H \leq \bar{p}(1)$, so that the probability that project H is good is lower than the cutoff in the classic environment. Regardless of the disentanglement level α , project L is exploited. The value of disentanglement is then only due to the ability to continue collecting information; it is increasing in the prior p_H that project H is good. When $p_H > \bar{p}(0) = R_L/R_H$, regardless of the disentanglement level α , project H is explored and exploited. Disentanglement is then beneficial only due to the continuation value in the eventuality that no news arrives and the posterior falls below R_L/R_H when a sufficiently long period transpires without news. The probability of no news is decreasing in p_H . Thus, the value of disentanglement is decreasing in the region $(\bar{p}(0), 1)$. The benefits of disentanglement are, therefore, highest in the $(\bar{p}(1), \bar{p}(0))$ region. In this region, when $\alpha = 1$, the agent exploits a sub-optimal project for its exploration value. Disentanglement limits the payoff loss associated with such exploration.

Consider now bad news settings (depicted in the right panel). As in good news settings, when $p_H < \bar{p}(1)$, regardless of α , project L is exploited. The value of disentanglement is due to the information it affords. This value is increasing in the prior likelihood that project H is good. When $p_H > \bar{p}(1)$, in the classical environment, with $\alpha = 1$, the agent explores and exploits project H . Absent news, the agent becomes increasingly optimistic and continues exploring project H . This persistence in the explored and exploited project generates a kink in payoffs, noted by KR, which yields the kink seen in Figure 1. Disentanglement leads the agent to exploit project L for posteriors higher than $\bar{p}(1)$, when its expected payoffs are higher than those from project H . The benefit from doing so decreases with the probability that project H is, in fact, the better project. When $p_H > \bar{p}(0)$, regardless of the level of disentanglement, the agent explores and exploits project H and switches the project she exploits only upon seeing bad news. Thus, expected payoffs are independent of α in the region $(\bar{p}(0), 1)$.¹²

The following corollary summarizes our discussion.

¹²In fact, Propositions A and B in the Appendix imply that the value of disentanglement is single-peaked in p_H in both news settings. See Section 1 in the Online Appendix for details.

Corollary 1 (One Safe Project: Comparative Statics). *The disentanglement value $\Delta\Pi(p_H, r/\lambda)$ is non-monotonic in each of its arguments. It is maximized at p_H^* such that $\bar{p}(1) < p_H^* < \bar{p}(0)$ in good news settings and at $p_H = \bar{p}(1)$ in bad news settings.*

As Figure 1 suggests, the value of disentanglement is higher under pure good news than under pure bad news with the same arrival rate, holding other parameters constant. This result follows directly from Propositions A and B in the Appendix. Intuitively, disentanglement is valuable only when the agent seeks to exploit project L , the safe project. Disentanglement allows the agent to explore project H , where uncertainty remains, even while exploiting project L . Good news about project H is then more valuable than bad news: only good news would prompt the agent to shift away from exploiting project L and start exploiting project H . Consequently, the advantage of disentanglement—allowing the agent to gather more information about project H while continuing to exploit project L —is particularly pronounced in the pure good news setting.¹³

In terms of the degree of disentanglement α , increasing it tightens the agent’s constraint, and this reduces her expected payoffs. However, the relationship between expected payoffs and α is neither concave nor convex. To see this, consider, for instance, the balanced news setting. For any $p_H \in (\bar{p}(1), \frac{R_L}{R_H})$, there exists α^* such that $\bar{p}(\alpha^*) = p_H$. Using the monotonicity of $\bar{p}(\cdot)$ in Proposition 1, at the outset, the agent exploits the risky project H for any $\alpha > \alpha^*$. Furthermore, in a balanced news setting, the only way the agent updates her posterior, and changes her exploited project, is by receiving news. Therefore, the agent’s expected payoffs are constant in α for $\alpha > \alpha^*$. However, for $\alpha < \alpha^*$, expected payoffs are strictly decreasing and concave in α ; see the Appendix for details.¹⁴ In particular, expected payoffs are neither concave nor convex in α over the interval $[0, 1]$.

4 Two Risky Projects: Balanced News and Indexability

We now analyze the general case of two risky projects, where $p_L, p_H \in (0, 1)$. For tractability, we assume full disentanglement, $\alpha = 0$. In this case, the agent’s optimal exploitation choices are simple: she always chooses the myopically optimal project, which we term the *favorable* project. That is, project x is favorable, while project y is unfavorable, if $p_x R_x > p_y R_y$. Both projects are favorable when their expected values coincide. The focus of our analysis is, therefore, on the characterization of optimal exploration.

¹³Section 1 of the Online Appendix contains a formal proof of this result.

¹⁴When $\lambda = \lambda_H^g = \lambda_H^b$ and $\alpha < \alpha^*$, the resulting expected payoff is given by:

$$R_L + \frac{\lambda(1-\alpha)}{r + \lambda(1-\alpha)} p_H (R_H - R_L),$$

which is concave in α .

In this section, we analyze balanced news settings in which $\lambda_z^b = \lambda_z^g = \lambda_z$ for $z = L, H$. The analysis of these settings proves instrumental for the characterization of optimal policies in good and bad news settings, which follow. While rarely studied in the literature, these settings reflect environments in which the arrival rate of news does not depend on its valence. For example, when assessing the efficacy of a menu of medical treatments using clinical trials, the arrival rate of news depends on the number of patients and the rate at which they are treated, but not necessarily on the quality of the treatments per se. Similarly, when researching the promise of an investment opportunity, the arrival rate of news often depends on the scope and speed of investigation, not explicitly on the quality of the investment option.

As we show, the analysis of the balanced news settings also starkly illustrates that the optimal policy cannot be characterized via an index à la [Gittins \(1979\)](#) in our environment.

4.1 Optimal Policies Under Balanced News

A key feature across all news settings is that the optimal policy becomes straightforward after news is revealed. Specifically, if either project is revealed to be bad or if project H is revealed to be good, exploration no longer has value. If project L is revealed to be good, the policy reverts to the case described in Section 3. Given that the post-news decision is resolved simply, our analysis focuses primarily on the features of the optimal policy before news arrives.

Suppose the agent optimally explores one of the projects at the outset. Absent news, the agent's posterior probabilities and, therefore, her decision problem do not change. In particular, in the optimal policy, the agent does not switch the project she explores unless news arrives. The agent's exploration choice is then effectively a static problem corresponding to her decision of which project to start exploring at the outset.

In order to characterize the optimal policy, the following notation is useful. When project x is favorable, we define $\tilde{p}_x \equiv p_x$. When project x is unfavorable, we define $\tilde{p}_x \equiv \min(p_y R_y / R_x, 1)$. The probability $\tilde{p}_x \geq p_x$ is, therefore, a modification of the probability that any project $x = L, H$ is good.

Proposition 2 (Optimal Exploration in Balanced News Settings). *Suppose $\lambda_z^b = \lambda_z^g = \lambda_z$ for $z = L, H$. Any optimal exploration strategy entails exploring project x until news arrives, where $\lambda_x(1 - \tilde{p}_x) \geq \lambda_y(1 - \tilde{p}_y)$, with $y \neq x$.*

Intuitively, the agent selects the project that is most “informative.” A higher arrival rate of news certainly increases the appeal of exploring a project. In addition, information is useful only when it affects exploitation decisions. When the agent explores the favorable

project, only bad news triggers a switch in exploitation. Bad news on project x can arrive only for a bad project x , which occurs with probability $1 - p_x$. In contrast, exploration of an unfavorable project y may or may not lead to a change in exploitation choices, even if good news arrives. Indeed, if the agent is sufficiently optimistic about project x , good news on project y would not sway her exploitation choices. In such cases, exploring project y before learning the quality of project x is of no value. Hence, the probability adjustment factor in the proposition, which raises the hurdle for exploring unfavorable projects.

The optimal exploration strategy is generally unique, with two exceptions. First, whenever the knife-edge condition that $\lambda_x(1 - \tilde{p}_x) = \lambda_y(1 - \tilde{p}_y)$ for $y \neq x$ holds, any exploration strategy is optimal. Second, if project H is explored and good news arrives, the agent exploits project H forever. Any ensuing exploration is immaterial and thus optimal.

In the classical environment, when exploration and exploitation are fully entangled, each project is associated with a (Gittins) index that depends only on that project's parameters. Specifically, the index for a project z is given by $p_z R_z \frac{(r + \lambda_z)}{(r + p_z \lambda_z)}$. In the optimal policy, the agent explores and exploits the project with the higher index. With balanced news, the agent switches away from exploring and exploiting project z only upon receiving news.

In our environment, with exploration disentangled from exploitation, the expected reward $p_x R_x$ of each project x serves as a separable index for exploitation: the agent optimally exploits whichever project generates the highest expected reward. The agent may switch her exploited project twice when exploration starts from an unfavorable project L : first, if she learns her initially unfavorable project L is good and, second, if she later learns her initially exploited project H is, in fact, good (as $R_H > R_L$). This already highlights the importance of disentanglement, as exploration and exploitation need not track one another. Furthermore, as we show below, there is no separable index that underlies optimal exploration. Intuitively, the value of exploring the unfavorable project depends on the returns of the favorable project.

The ability to disentangle exploration from exploitation alters the model's comparative statics. We highlight two key differences from the classical environment. First, and in sharp contrast to the classical setting, the optimal policy derived here is independent of the discount rate. Second, the effect of prior beliefs on the exploration strategy is more nuanced. In the classical environment, a project becomes monotonically more attractive to explore as its prior probability of being good increases. In our setting, however, the prior's influence depends on whether the project is currently favorable. For illustration, suppose project L is favorable, so that $p_L R_L \geq p_H R_H$. The choice of which project to explore reduces to a comparison between $\lambda_L(1 - p_L)$ and $\lambda_H(1 - p_L R_L / R_H)$. As the prior p_L of project L increases to 1, the first term converges to 0 and the second term converges to $\lambda_H(1 - R_L / R_H) > 0$, making project L unambiguously less attractive to explore. In con-

trast, increasing the prior p_H of the unfavorable project has no effect on this trade-off until project H itself becomes favorable.

4.2 No Exploration Index

As mentioned above, in the classical environment, Gittins (1979)'s characterization of the optimal policy holds. That is, each project is associated with an index that only depends on the parameters of that project. At any point, the agent explores and exploits the project with the highest current index. While Proposition 2 offers a simple characterization of the optimal policy, we now show that, in our setting, optimal exploration is not governed by an index à la Gittins (1979).

Suppose that the optimal policy in a balanced news setting can be described via an index tailored to each project. We denote by $I(p, R, \lambda)$ the index corresponding to a project with a probability p of being good, an arbitrary reward $R > 0$ conditional on being good, and a rate of news arrival—good or bad—of λ .

Consider three hypothetical projects. Project $i = 1, 2, 3$ is governed by a probability p_i that it is good, associated with a flow reward of $R_i > 0$, and a news arrival rate of $\lambda_i > 0$. Suppose that

$$R_1 > p_2 R_2 > p_1 R_1 \quad \text{and} \quad \lambda_2(1 - p_2) < \lambda_1 \left(1 - \frac{p_2 R_2}{R_1}\right).$$

Then, using Proposition 2, when the agent has access to projects 1 and 2, she optimally exploits project 2, but explores project 1. That is, $I(p_1, R_1, \lambda_1) > I(p_2, R_2, \lambda_2)$.

Suppose now that

$$p_2 R_2 > R_3 > p_3 R_3 > p_1 R_1.$$

This implies that, when the agent has access to projects 2 and 3, she optimally explores and exploits project 2. That is, $I(p_2, R_2, \lambda_2) > I(p_3, R_3, \lambda_3)$.

Suppose, further, that λ_3 is high enough so that

$$\lambda_3(1 - p_3) > \lambda_1 \left(1 - \frac{p_3 R_3}{R_1}\right).$$

This implies that, when the agent has access to projects 1 and 3, she explores and exploits project 3. Therefore, $I(p_3, R_3, \lambda_3) > I(p_1, R_1, \lambda_1)$, establishing a cycle, in contradiction. Although this construction is done for the balanced news setting, it is robust to small perturbations of parameters. In particular, the optimal exploration policy is not generally governed by an index for either good or bad news settings either. Thus,

Corollary 2 (No Exploration Index). *The optimal exploration policy is not governed by an index.*

We emphasize that this conclusion is not driven by an excess number of degrees of freedom. Both our environment and the classical benchmark share the same fundamental degrees of freedom in the problem’s description—namely, the parameters p_i , R_i , and λ_i for $i = 1, 2$. In Section 2 of the Online Appendix, we show that non-indexability holds for arbitrary levels of disentanglement.

5 Optimal Policies Under Good and Bad News

In this section, we characterize the optimal policies in both good and bad news settings with two risky projects. We show that the optimal policy entails very few switches of either the exploited or the explored project. Furthermore, unlike the special case in which one project is safe, the information structure has a substantial impact on the shape of the optimal policy. In Section 3 of the Online Appendix, we demonstrate that our results remain robust to small but positive levels of entanglement (namely, $\alpha > 0$).¹⁵

5.1 Good News Settings

We now analyze good news settings. Before describing our general characterization, consider the following example, highlighting the impacts of disentanglement when both projects are risky.

Example 1 (Good News: Ex-ante Identical Projects) Suppose the two projects are ex-ante identical: $p_L = p_H$ and $R_L = R_H$.¹⁶ Furthermore, consider the pure good news setting in which $\lambda_z^b = 0$ and $\lambda_z^g = \lambda > 0$ for $z = L, H$.

In the classical environment with $\alpha = 1$, the optimal strategy requires splitting exploration and exploitation equally between the two projects until receiving news. Intuitively, consider a discrete time approximation of this problem. If the agent explores and exploits project x , the corresponding Gittins index declines absent news—the agent becomes more pessimistic about project x . She should then immediately

¹⁵The general analysis for $\alpha \in (0, 1)$ is more intricate since, in that case, when the agent explores one project while exploiting another, the absence of news affects the posterior probabilities for both projects. Moreover, depending on the value of α , the relative news arrival rates, and the prior likelihoods that the projects are good, the posterior may evolve more rapidly for either explored or exploited project.

¹⁶Strictly speaking, this violates our assumption that $R_H > R_L$, which generates the non-trivial scenarios for Section 3. We assume equal rewards here to simplify our illustration of the stark effects of disentanglement.

switch to project y , only to switch back an instant later. In the limit, splitting exploration and exploitation equally across the two projects leads the two indices to decline at the same rate and maintains the incentive to continue with such a split. We can interpret this strategy as requiring the agent to switch between projects infinitely often.¹⁷

In contrast, in our setting with $\alpha = 0$, an optimal policy requires indefinite disentanglement, i.e., exploiting one project and exploring the other indefinitely, until the arrival of good news. If the agent exploits project x and explores project y at the outset for any infinitesimal time interval, project x becomes favorable, so continuing to exploit project x is optimal. Furthermore, information is useful to the agent only if it leads her to change her exploited project. Good news on project x would not alter her exploitation choices; only good news on the unfavorable project would. This means that it is optimal for the agent to use her entire exploration budget on project y : any splitting of exploration resources between the two projects is sub-optimal since it reduces the effective rate at which news arrives on the unfavorable project. As a consequence, with full disentanglement, the agent switches her exploitation choices at most once and *never* switches her exploration choice prior to receiving news. Of course, the agent is indifferent as to which project she explores and which she exploits at the outset given the complete symmetry of the problem. In fact, the agent can also choose at random which project to start exploring. The contrast with the classical environment is that such randomization cannot proceed with a split of exploration or exploitation for a non-negligible duration.

In general, in the classic environment, when projects are heterogeneous, the agent initially explores and exploits the project with the higher Gittins index. Absent news, that project's Gittins index declines over time, until it reaches equality with the index of the other project. Upon such indifference, the agent splits exploration and exploitation to maintain her indifference. We can interpret this splitting of attention, or exploration resources, as the limit of sequential immediate switches in discrete time (see [Gittins et al., 2011](#)). As we now show, such rapid switches *never* occur when exploration and exploitation are disentangled.

Consider then a disentangled setting with good (or balanced) news, where $\lambda_x^g \geq \lambda_x^b$ for $x = L, H$. Whenever project x is favorable, so that $p_x R_x \geq p_y R_y$, exploiting project x is optimal. When the agent explores project x , receiving no news makes her increasingly pessimistic. We denote by $\bar{t}_x(p_L, p_H)$ the time it takes the agent to reach equality between the expected values of both projects. If $p_x R_x = p_y R_y$, then $\bar{t}_x(p_L, p_H) = 0$; otherwise,

¹⁷See case (v) in Section 3.3.2 of [Gittins et al. \(2011\)](#) for details.

$\bar{t}_x(p_L, p_H) > 0$.¹⁸ Specifically, after exploring project x for a duration $\bar{t}_x(p_L, p_H)$ without receiving news, the agent's posterior that project x is good is precisely $p_y R_y / R_x$. That is,

$$\frac{p_x e^{-\lambda_x^g \bar{t}_x(p_L, p_H)}}{p_x e^{-\lambda_x^g \bar{t}_x(p_L, p_H)} + (1 - p_x) e^{-\lambda_x^b \bar{t}_x(p_L, p_H)}} = p_y R_y / R_x.$$

Simplifying, whenever $\lambda_z^g > \lambda_z^b$, we obtain:

$$\bar{t}_x(p_L, p_H) = \frac{1}{\lambda_x^g - \lambda_x^b} \ln \left(\frac{p_x (R_x - p_y R_y)}{p_y R_y (1 - p_x)} \right).$$

We now state our result characterizing optimal exploration in this setting.¹⁹

Proposition 3 (Optimal Exploration in Good News Settings). *Suppose $\lambda_z^g > \lambda_z^b$ for $z = L, H$ and that project x is favorable, so that $p_x R_x \geq p_y R_y$. An optimal exploration strategy is described as follows.*

- *If, at any time the agent explores project y , she never switches absent news.*
- *If the agent initially explores project x , then if, absent news, she switches to exploring project y , she does so at a time $T \leq \bar{t}_x(p_L, p_H)$.*

Furthermore, if $\lambda_x^b = 0$, there is an optimal strategy in which the agent never switches her explored project absent news.

Proposition 3 illustrates that disentanglement dramatically reduces the expected number of switches prescribed by the optimal policy. The exploited project can be switched at most twice: starting from project H , good news about project L could lead to one switch if the agent becomes sufficiently pessimistic about project H , and later good news about project H would lead to a switch back to project H . The explored project can be switched at most once before any news arrives. In fact, if the agent explores the unfavorable project initially, she never switches the explored project absent news, no matter how pessimistic she becomes about this project. Of course, when she becomes pessimistic about the explored project, she also becomes increasingly confident that she is exploiting the superior

¹⁸If $p_x R_x > p_y R_y$ in a balanced news setting, exploring project x does not change the agent's posterior and we set $\bar{t}_x(p_L, p_H) = \infty$. When $p_x R_x \leq p_y R_y$, we denote $\bar{t}_x(p_L, p_H) = 0$ even when $\lambda_x^g = \lambda_x^b$ and the agent does not alter her prior as time passes without information.

¹⁹As stated at the outset, we ignore 0-measure sets. When we say the agent explores project y at some time, we mean the agent explores project y for a positive-measure set of times. Switching to a project x implies that there is an ensuing positive-measure set of times at which the agent explores project x .

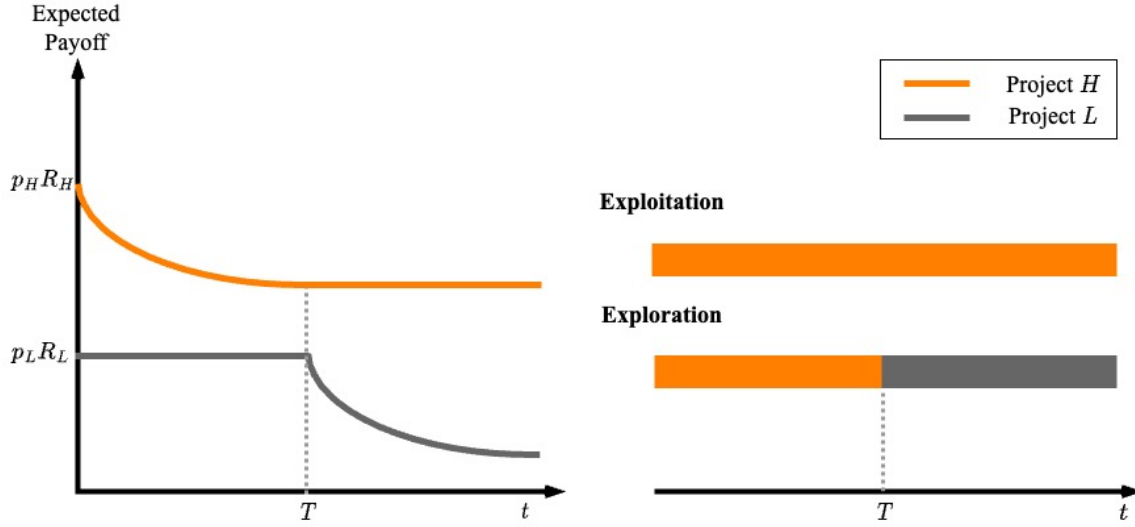


FIGURE 2: Optimal policy with two risky projects in good news settings

project.²⁰

The role of disentanglement is evident in the optimal policy described in Proposition 3. Under this policy, eventually, absent news, the exploited project must differ from the explored project. Indeed, since the optimal policy prescribes indefinite exploration of a project, eventually the posterior probability that the explored project is good must be low enough to make the other project favorable and, therefore, exploited.

Figure 2 depicts the exploration and exploitation patterns in good news settings. In the figure, project H is initially favorable and both explored and exploited. As time progresses without news, the agent's confidence in project H diminishes. However, at time T , the agent switches to exploring project L even though project H remains the more favorable option. In the absence of news, the agent continues to exploit project H indefinitely. Starting from time T , the projects she explores and exploits diverge.

In order to understand the logic of Proposition 3, consider first the case in which the optimal exploration strategy prescribes exploring the unfavorable project y initially, so that the agent explores and exploits different projects. The value of exploring project y depends only on the rate at which good news arrives: receiving bad news on project y retains project x as favorable. Thus, similar to the setting with one safe project, the value of exploring project y depends on λ_y^g , but not on λ_y^b . In particular, this value is the same if we increase the arrival rate of bad news so that news is balanced, $\lambda_y^b = \lambda_y^g$. In this case, as

²⁰While the first switch may occur sooner in our model than in the standard entangled environment, it is also the *final* switch. In contrast, full entanglement yields infinitely many switches at infinitesimal intervals, leading to a lower average time between switches.

discussed in Section 4.1, the agent's posteriors do not change absent news. If it is optimal to explore project y at some point, it is also optimal to explore project y after any amount of time that has passed without news. We can conclude that it must also be optimal to continue exploring project y in the absence of news when $\lambda_y^b < \lambda_y^g$.

Consider now the case in which the optimal strategy prescribes exploring the favorable project x initially, so that the agent explores and exploits the same project at first. Why can it be optimal for the agent to switch the project she explores when $\lambda_x^b > 0$? Suppose, for instance, that $p_H R_H > R_L \geq p_L R_L$. In this scenario, exploring project L initially is not useful: even good news on project L would not lead the agent to switch her exploited project. Instead, if $\lambda_H^b > 0$ and the agent explores project H , she would switch to exploiting project L upon receiving bad news on project H : exploring project H is valuable. When, instead, $p_L R_L < p_H R_H < R_L$, good news on project L would lead the agent to switch to exploiting project L , implying that exploring project L can be useful. The determination of the switching time T depends on the relative magnitudes of $p_H R_H$, $p_L R_L$, and the arrival rates of news on the two projects.²¹

Why can the agent not switch exploration of the favorable project x after a duration $T > \bar{t}_x(p_L, p_H)$ without receiving news? By the definition of $\bar{t}_x(p_L, p_H)$, after such a duration T without news, project x becomes unfavorable. Our previous arguments then imply that, absent news, indefinite exploration of project x beyond time T is optimal.

As is common in individual decision problems, multiplicity of optimal policies is rare, as the following claim illustrates.

Claim 1 (Uniqueness of Optimal Policy). *Suppose that $\lambda_z^g > \lambda_z^b$ for $z = L, H$ and that the favorable project x is such that $R_y > p_x R_x > p_y R_y$. Then, generically, the optimal policy is unique.*²²

The claim captures settings in which good news on the unfavorable project is immediately valuable: it alters the exploited project. Multiplicity can arise when information has no instantaneous value, regardless of which project is explored. This is the case when $p_H R_H > R_L$ in a pure good news setting, with $\lambda_z^b = 0$ for $z = L, H$. In this parameter region, the agent exploits project H initially even after good news on project L . Therefore, exploration has no instantaneous value. Of course, good news on project L leads the agent to explore project H , and a sufficiently long time without news would lead the agent to switch to exploiting project L . In this case, there are multiple optimal strategies that differ in which project is explored initially.

²¹The bound $T \leq \bar{t}_x(p_L, p_H)$ is tight for some parameters. See Section 4 in our Online Appendix for details.

²²The optimal policy is unique as long as $\lambda_x^b R_y \neq \lambda_y^g R_x$, which excludes a zero-measure set of parameters.

We now turn to a discussion of the initial exploration choice. For expositional simplicity, we focus on the special case of pure good news settings, where $\lambda_x^b = 0$, $x = L, H$. In this case, Proposition 3 indicates that an optimal policy has the agent explore the same project until receiving news, implying that the initial choice is permanent absent news. We also restrict attention to the case $p_L R_L < p_H R_H < R_L$, where information on both projects is valuable at the outset. Indeed, exploring project H for a sufficiently long time would make the agent pessimistic about the quality of that project and, absent news, the agent would switch her exploited project after a duration $\bar{t}_H(p_L, p_H)$. Exploring project L is also valuable: receiving good news on that project would lead the agent to immediately switch the project she exploits. In particular, for this set of parameters, exploring either project can be optimal depending on the difference between news' arrival rates.

In line with our previous notation, we denote $\tilde{p}_L \equiv p_H R_H / R_L$. Thus, \tilde{p}_L corresponds to the prior that project L is good at which the agent is indifferent between the two projects.

Claim 2 (Initial Choice with Pure Good News). *Suppose $\lambda_z^b = 0$ for $z = L, H$ and that $p_L R_L < p_H R_H < R_L$. It is optimal to explore project H if and only if $\lambda_H^g \frac{w - \rho_L}{1 - \rho_L} (1 - p_H) \geq \lambda_L^g (1 - \tilde{p}_L)$, where $w = e^{-r \bar{t}_H(p_L, p_H)}$ and $\rho_L = \lambda_L^g / (r + \lambda_L^g)$.*

The specification in the claim is reminiscent of the one appearing in Proposition 2, with the added multiplier $\frac{w - \rho_L}{1 - \rho_L}$ for project H . As already noted, Proposition 3 indicates that in the pure good news setting, we only need to compare two cases, differing in which project is explored until news. Suppose that exploring project H is optimal. At time $\bar{t}_H(p_L, p_H)$, project L becomes favorable and the agent explores and exploits different projects. As described in the intuition for Proposition 3, the value of exploring project H depends only on the arrival rate of good news: receiving bad news on project H sustains project L as favorable. Thus, the value of exploring project H depends on λ_H^g , but not on λ_H^b . Consequently, starting at $\bar{t}_H(p_L, p_H)$, the expected payoffs from this problem are the same as those in an auxiliary balanced news problem with arrival rate of λ_H^g for both good and bad news. The determination of which project to explore must then conform with the characterization in Proposition 2.

In contrast with the balanced news setting, when $p_H R_H > p_L R_L$, the initial comparison includes the factor $\frac{w - \rho_L}{1 - \rho_L}$, penalizing the exploration of project H . To understand this penalty, note that, absent news, if the agent explores project H , she switches the exploited project only after a duration $\bar{t}_H(p_L, p_H)$. The larger this duration, the longer the period in which exploration without news does not affect the agent's exploitation, and the less appealing it is to explore project H . Indeed, w and $\frac{w - \rho_L}{1 - \rho_L}$ decrease with $\bar{t}_H(p_L, p_H)$. If both projects are favorable, so that $\bar{t}_H(p_L, p_H) = 0$, or if the agent is infinitely patient ($r = 0$), then

$w = 1$ and the claim's inequality boils down to the comparison in Proposition 2. Similar characterizations hold for other cases of prior probabilities that either project is good.

This claim offers another way to show the way by which disentanglement of exploration from exploitation has bite. Although project H is optimally exploited at the outset, it is optimal to explore project L whenever $\rho_L > w$, i.e., when news arrival on project L is fairly rapid. Similarly, as the agent becomes more and more impatient, with r increasing indefinitely, both w and ρ_L approach 0, and the agent explores project L . Since $p_L R_L < p_H R_H < R_L$, in these circumstances, the agent would exploit project H initially regardless of which project she explores. She switches the project she exploits only if she learns that project L is good. Furthermore, unlike the comparative statics in the classical entangled environment, exploration of project L becomes more appealing as p_H increases.

In general, comparing the payoffs generated by the optimal policy in our setting to those generated in the classical environment yields similar insights to those observed when one of the projects is safe, as presented in Corollary 1. When arrival rates λ_L^g and λ_H^g are very high or when the discount rates are very low, the agent can achieve payoffs close to those corresponding to a complete information setting in both environments. Similarly, when arrival rates λ_L^g and λ_H^g are very low, or discount rates are very high, the agent receives an expected payoff approximating the myopic expected payoff in both environments. In particular, the benefits of disentanglement are most pronounced for intermediate levels of arrival and discount rates. Similarly, the benefits of disentanglement are non-monotonic in the prior p_H .

5.2 Bad News Settings

We now turn to bad news settings. Before characterizing the optimal policy, consider the following example, which complements Example 1 and illustrates some of the qualitative differences between the information structures we consider.

Example 2 (Bad News: Project L is Favorable) Suppose that $\lambda_z^g = 0$ and that $\lambda_z^b = \lambda_z > 0$ for $z = L, H$. Furthermore, suppose project L is favorable, so that $p_L R_L > p_H R_H$.

In the classical bandit environment, if the wedge between the projects' expected values, or between their arrival rates $(\lambda_L - \lambda_H)$, is sufficiently high, the agent explores and exploits project L . Absent news, the agent becomes increasingly optimistic about project L and thus continues exploring and exploiting project L indefinitely. If project L is indeed good, then the agent never receives bad news on project L and therefore never learns whether project H is good.

In contrast, with full disentanglement, even if the agent explores project L at the outset, which is optimal if λ_L is high enough, she does not do so indefinitely. Switching

the exploited project can occur both upon learning that project L is bad, and when becoming increasingly optimistic about project H . As the duration of exploration of project L increases, so does the posterior p_L , implying that the likelihood of learning that project L is bad vanishes, as does the value of continuing to explore it. Consequently, switching to exploring project H is eventually optimal. Thus, in bad news settings, disentanglement may lead to more switching of the explored projects than in the classical environment.

The observation in Example 2 that, in the classical environment, the agent explores and exploits the same project indefinitely unless news arrives is clearly quite general. At the outset, if the Gittins index is higher for project x , that project is explored and exploited. Absent bad news, the agent becomes more optimistic about the quality of project x and its associated Gittins index increases. In contrast, in our environment, when exploration and exploitation are disentangled, the agent may optimally switch the project she explores.

Proposition 4 (Optimal Exploration in Bad News Settings). *Suppose $\lambda_z^b > \lambda_z^g$ for $z = L, H$. The optimal exploration strategy is described as follows.*

- *If the agent initially explores project H , she never switches absent news.*
- *If the agent initially explores project L , she switches after a period $T < \infty$ without news.*

In contrast with the optimal policy in good news settings, characterized in Proposition 3, in bad news settings, the optimal policy never entails exploring project L forever. The intuition is similar to that appearing in Example 2. When the agent explores project L , absent news, she becomes increasingly optimistic about its prospect. Consequently, regardless of news' arrival rates, after a sufficiently long period of exploring project L without news, the agent exploits project L and the likelihood she learns project L is bad becomes vanishingly small. The value of exploring project H , however, remains strictly positive.

Figure 3 depicts the exploration and exploitation patterns in bad news settings. In the figure, project H is initially favorable and therefore exploited, but project L is explored, say, because it features a high news arrival rate. As time progresses without news, the agent's confidence in project L increases. At time t_1 , project L becomes favorable, and the agent switches to exploiting it. By time T , the rate of learning on project L has flattened, and the agent switches to exploring project H , while continuing to exploit project L . Without news, the agent becomes increasingly optimistic about project H . At time t_2 , project H becomes favorable again and the agent switches to exploiting it. Absent news, the agent continues to explore and exploit project H indefinitely.

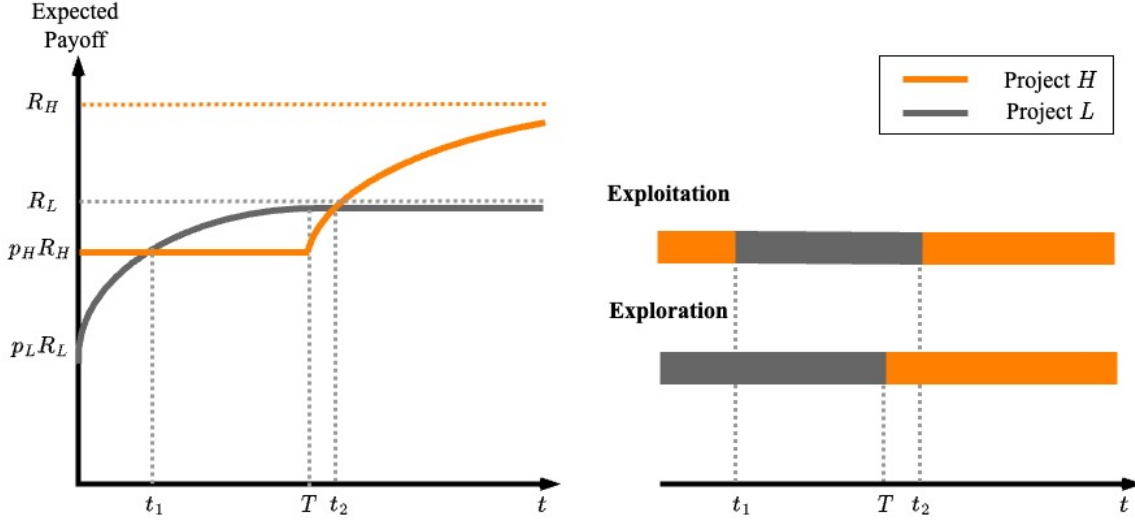


FIGURE 3: Optimal policy with two risky projects in bad news settings

Proving that the optimal policy involves no switching when project H is initially explored is involved. However, when $p_H R_H \geq R_L$, the claim is straightforward: the agent explores project H from the start since even good news about project L would not prompt her to switch her exploited project; exploring project L is not decision relevant.

When $R_L > p_H R_H \geq p_L R_L$, it is useful to consider an auxiliary problem in which the agent receives balanced news about project H at the original rate λ_H^b for both good and bad news; the original arrival rates are used when project L is explored. In the auxiliary problem, the agent has more information than in the original problem. If, in the original problem, exploring project H is optimal, then it is also optimal to explore project H in the auxiliary problem, where it is more informative. Absent news, the agent would optimally explore project H until news in the auxiliary problem: her posteriors do not change. The agent can emulate that same strategy even in the original problem. Furthermore, exploring and exploiting project H until news generates the same payoffs in both problems. Since it is optimal in the auxiliary problem, which affords the agent more information, it must be optimal in the original problem as well. The remaining case in which $p_L R_L > p_H R_H$ is discussed in the Appendix. The proof of Proposition 4 also illustrates that the optimal policy is unique until news arrives.²³

As for the initial choice of projects, our discussion above suggests that whenever $p_H R_H \geq R_L$, the agent begins by exploring project H . When $R_L > p_H R_H$, project L is explored initially when bad news' arrival rate for project L is sufficiently high. The proof of Proposition

²³If any news arrives about project H , or if bad news arrives about project L , later exploration has no effect on payoffs. In particular, after such news arrives, any exploration strategy is optimal.

4 provides the relevant parameter comparisons governing the choice of which project is optimally explored first.

In terms of comparative statics, in bad news settings—unlike good news settings—as r grows indefinitely, the agent optimally explores and exploits the same project: the only news that would change short-term exploitation is bad news on the exploited project. The comparison of payoffs generated by the optimal policy with and without disentanglement is similar to that observed for good news settings and that described with one safe project in Corollary 1. In particular, the benefits of disentanglement are most pronounced for intermediate values of parameters.

5.3 The Role of Disentanglement with Different News Processes

The agent’s optimal behavior differs sharply between good news and bad news settings. In good news settings, as time passes without news, the agent eventually explores a project different from the one she exploits. By contrast, in bad news settings, exploration and exploitation eventually converge on the same project, so disentanglement matters only in the short run. In fact, under bad news, the agent must ultimately explore and exploit project H , whereas under good news, she may persistently explore or exploit either project L or H .

This divergence arises from the distinct role of learning from the passage of time. With good news, the agent becomes increasingly pessimistic about the project being explored, prompting her to exploit the alternative project; changes in exploitation decisions are therefore triggered by good news. With bad news, the agent becomes increasingly optimistic about the project she explores, leading her to eventually align exploration and exploitation; here, changes in exploitation arise only after bad news. Moreover, as Proposition 4 shows, in bad news settings, the agent must eventually switch to exploring project H if she begins with exploring project L .

6 Strategic Disentangled Experimentation

We now incorporate disentangled exploration and exploitation into a strategic team setting, building on the framework of KRC. As in KRC, we consider a group of n identical agents, each of whom must decide, at each moment in time, how to allocate a unit endowment of both exploration and exploitation across two common projects, one of which is safe as in Section 3. Following KRC, we assume that all agents observe the outcomes of exploration—both their own and of others—but derive benefits only from their own exploitation. All other aspects of the exploration and exploitation technologies mirror those

in the single-agent model. In particular, while exploration outcomes are fully shared, payoff outcomes from exploitation remain unobserved by any agent. For tractability, we focus on the balanced news case with $0 \leq \alpha \leq 1$.

Consider a Markov strategy profile in which, absent news, agent i exploits the risky project with intensity k_i , $i = 1, \dots, n$.²⁴ Any agent i then allocates $1 - \alpha + \alpha k_i$ of her exploration endowment to the risky project. Let $k_{-i} = \sum_{j \neq i} k_j$. Agent i 's payoff under this profile is given by

$$\begin{aligned} V^{Team}(k_i, k_{-i}, p_H) &= \\ &= \frac{(k_i p_H R_H + (1 - k_i) R_L) \cdot r + (p_H R_H + (1 - p_H) R_L) \cdot (\alpha k_i + \alpha k_{-i} + n(1 - \alpha)) \lambda}{r + (\alpha k_i + \alpha k_{-i} + n(1 - \alpha)) \lambda}. \end{aligned}$$

Therefore, agent i 's best-response to opponents' strategies depends only on $k_{-i} = \sum_{j \neq i} k_j$ and is given by:

$$\beta_i(k_{-i}) = \begin{cases} 1 & \text{if } p_H > \bar{p}_\alpha(n) \text{ or } [\underline{p}_\alpha(n) \leq p_H \leq \bar{p}_\alpha(n) \text{ and } k_{-i} < \hat{k}] \\ [0, 1] & \text{if } \underline{p}_\alpha(n) \leq p_H \leq \bar{p}_\alpha(n) \text{ and } k_{-i} = \hat{k} \\ 0 & \text{if } p_H < \underline{p}_\alpha(n) \text{ or } [\underline{p}_\alpha(n) \leq p_H \leq \bar{p}_\alpha(n) \text{ and } k_{-i} > \hat{k}], \end{cases}$$

where the boundary conditions are determined as follows. The threshold $\underline{p}_\alpha(n)$ is obtained by solving $V^{Team}(0, 0, p) = V^{Team}(1, 0, p)$ for p . At this threshold, agent i is indifferent when none of the other agents exploits the risky project. The threshold $\bar{p}_\alpha(n)$ is obtained by solving $V^{Team}(0, n-1, p) = V^{Team}(1, n-1, p)$ for p so that agent i is indifferent when all other agents fully exploit the risky project. The two thresholds satisfy $\underline{p}_\alpha(n) < \bar{p}_\alpha(n)$. Intuitively, an individual agent's exploitation impacts the team's learning less when all other agents exploit and explore the risky project relative to when all other agents exploit the safe project and only partially explore the risky project. Finally, for any $p_H \in [\underline{p}_\alpha(n), \bar{p}_\alpha(n)]$, the threshold \hat{k} corresponds to the overall exploitation by others that makes the agent indifferent. Formally,

$$\begin{aligned} \underline{p}_\alpha(n) &= \frac{(r + n\lambda(1 - \alpha)) R_L}{(r + \lambda(\alpha + n(1 - \alpha))) R_H - \lambda\alpha R_L}, \\ \bar{p}_\alpha(n) &= \frac{(r + (n - \alpha)\lambda) R_L}{(r + n\lambda) R_H - \lambda\alpha R_L}, \text{ and} \\ \hat{k} &= \frac{p_H(R_H - R_L)}{(R_L - p_H R_H)} - \frac{(r + n(1 - \alpha)\lambda)}{\lambda\alpha}. \end{aligned} \tag{2}$$

²⁴Since we consider a balanced news setting, posteriors are flat in the absence of news, and Markov strategies are therefore constant.

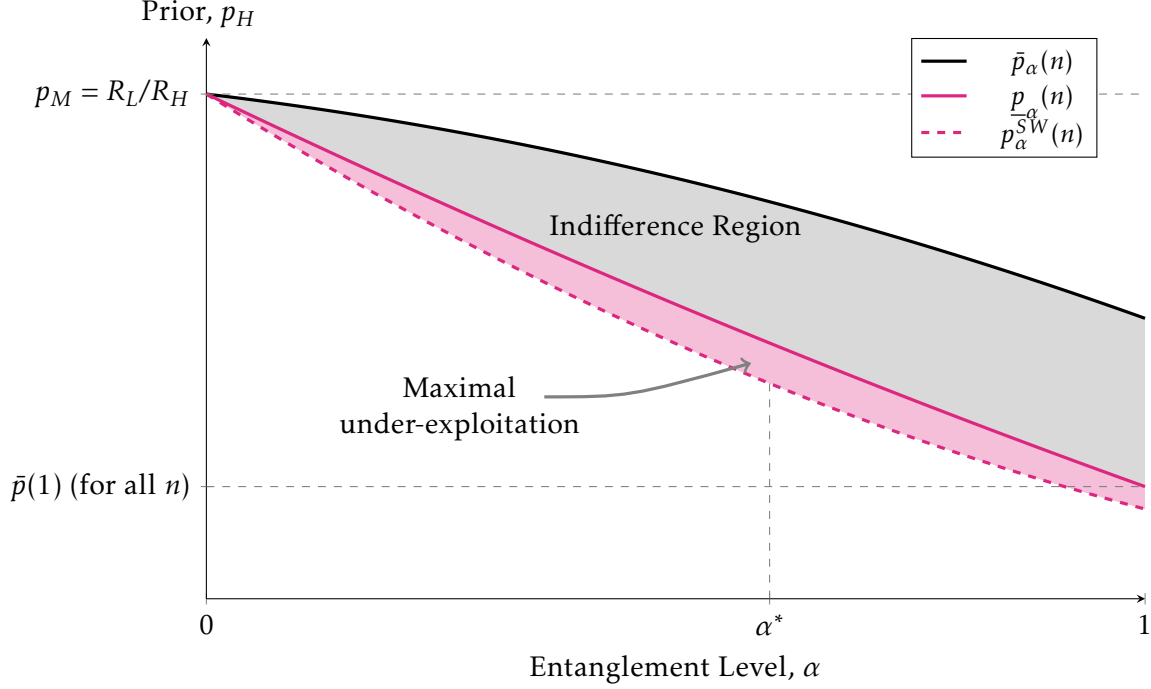


FIGURE 4: Dependence of regions of risky project exploitation on disentanglement

Thus, for every α , team exploration and exploitation induces a game with strategic substitutes. For $n > 1$, the exploitation intensity in the unique symmetric equilibrium, k^* , is given by:²⁵

$$k^* = \begin{cases} 1 & \text{if } p_H \geq \bar{p}_\alpha(n) \\ \hat{k}/(n-1) & \text{if } \underline{p}_\alpha(n) \leq p_H \leq \bar{p}_\alpha(n) \\ 0 & \text{if } p_H \leq \underline{p}_\alpha(n). \end{cases}$$

Now, consider the interval $(\underline{p}_\alpha(n), \bar{p}_\alpha(n))$ over which both projects are actively exploited. In the classical environment, when $\alpha = 1$, the lower bound $\underline{p}_\alpha(n)$ is constant in n and matches the single-agent cutoff, as in KRC. In contrast, in the fully disentangled case, when $\alpha = 0$, the indifference region collapses to the single point p_M . In fact, both $\underline{p}_\alpha(n)$ and $\bar{p}_\alpha(n)$ decrease in α , while the distance between them increases in α , see Figure 4. For intermediate cases, where $\alpha < 1$, the number of agents n plays a crucial role. Both bounds rise with n , reflecting the increase in “free” exploration provided by additional agents. As $n \rightarrow \infty$, this effect drives both bounds toward the myopic threshold, with the interval converging to $p_M = R_L/R_H$.

²⁵As in KRC, there may also be asymmetric equilibria.

The entanglement parameter α affects equilibrium exploitation intensity. As in the single-agent case, a higher α leads to more exploitation of the risky project. This is reflected by the exploitation rate k^* increasing in α and, consequently, by the entire indifference region shifting downwards. Of course, in the team problem, in the region $(\underline{p}_\alpha(n), \bar{p}_\alpha(n))$, each agent is indifferent between exploiting either project. As α increases, the cost of exploring project H rises since it becomes more tightly linked to risky exploitation. To restore the agent's indifference, the equilibrium must adjust to also reduce the benefit of exploring project H . This adjustment is achieved through an increase in the overall exploration level, k^* , undertaken by all team members.

To assess the team equilibrium's inefficiency, we use as benchmark the first-best allocation chosen by a utilitarian social planner. This allocation is the solution to the "cooperative problem" of KRC. The planner's problem is equivalent to that of a single agent facing a news arrival rate of $n\lambda$: that is, news arrives n times faster. Applying our results from Section 3 of the paper, the planner's optimal exploitation threshold is thus given by:

$$p_\alpha^{SW}(n) = \frac{(r + n\lambda(1 - \alpha))R_L}{(r + n\lambda)R_H - n\lambda\alpha R_L}. \quad (3)$$

This threshold, also depicted in Figure 4 here, is consistent with KRC's cooperative solution when $\alpha = 1$.²⁶ By comparing Equations (2) and (3), we can establish that, for all $n > 1$ and $\alpha > 0$, the social planner's threshold is strictly lower than the lower equilibrium threshold for exploitation of the risky project: $p_\alpha^{SW}(n) < \underline{p}_\alpha(n)$. These thresholds coincide for $\alpha = 0$ and $n = 1$. Therefore, for $\alpha > 0$, $n > 1$, and for any $p_H \in (p_\alpha^{SW}(n), \underline{p}_\alpha(n))$, the socially optimal solution for the team is to maximally exploit the risky project, whereas no exploitation of the risky project occurs in equilibrium. For values of p_H in the interval $[\underline{p}_\alpha(n), \bar{p}_\alpha(n)]$, the social planner would prescribe full exploitation of the risky project, whereas in equilibrium, only partial exploitation occurs: $k^* < 1$. Figure 5 depicts the inefficiencies in the intensity of risky project exploitation for different levels of α . The gap between the planner's solution and the equilibrium allocation reflects the presence of positive externalities: in equilibrium, agents do not internalize the informational benefits their own risky exploitation provides to others.

The interval $[p_\alpha^{SW}(n), \underline{p}_\alpha(n)]$ can be viewed as a region of maximal under-exploitation. The size of this interval is determined by the gap $\underline{p}_\alpha(n) - p_\alpha^{SW}(n)$, which is strictly increasing in α for $r > \hat{r} = \frac{\lambda(R_H - R_L)\sqrt{n}}{R_H}$. However, for $r < \hat{r}$ there is an interior maximum $\alpha^*(r)$ such that the size of the gap is decreasing for $\alpha > \alpha^*(r)$, as depicted in Figure 4 above.²⁷

²⁶As established in Section 3, the thresholds for good, balanced, and bad news are identical in the single-agent problem.

²⁷See Section 5 in our Online Appendix for precise derivations.

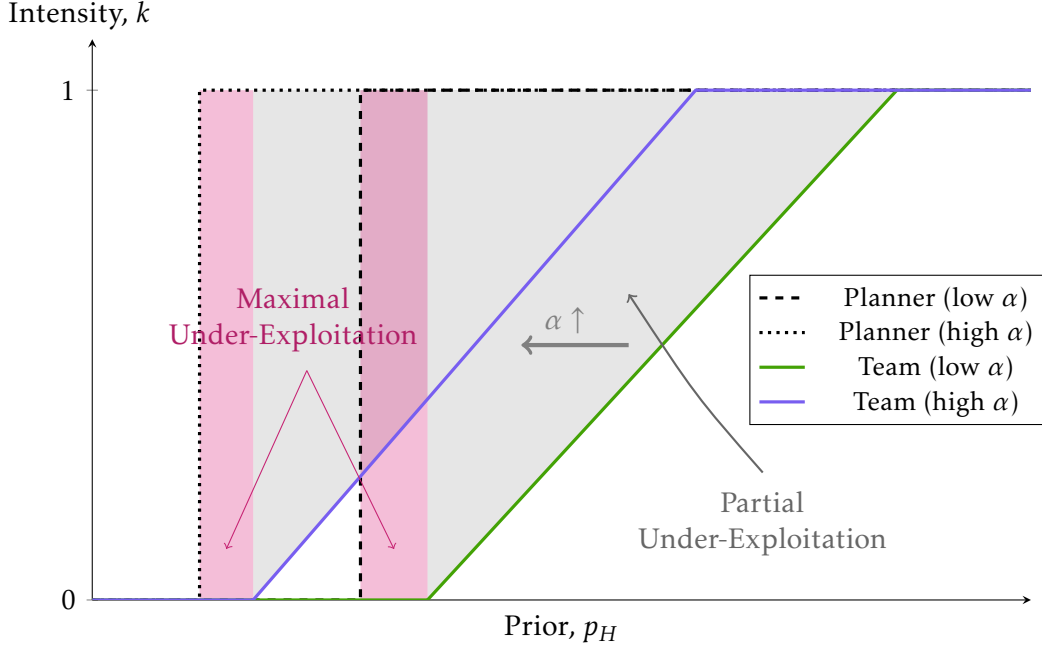


FIGURE 5: Inefficiencies in the intensity of risky project exploitation. The figure depicts the optimal rate of risky project exploitation for two disentanglement levels and describes the team equilibrium choices (“Team”) and the utilitarian social planner’s choice (“Planner”)

Intuitively, the relationship between the region of maximal under-exploitation and α arises from a subtle trade-off between the cost of exploration and the incentive to free-ride. As α increases from 0 to 1, the effective cost for an individual to explore the risky project rises. Initially, for low and intermediate values of α , this increasing cost amplifies the incentive to free-ride on the exploration efforts of other team members, causing the inefficiency gap between the team’s equilibrium and the social optimum to grow. However, as α becomes very high, the cost of exploration can become prohibitively high. The resulting scarcity of information increases the marginal value for any single agent to conduct her own exploration. Patient agents (low discount rate r) place higher weight on the future. Therefore, information considerations, which affect future payoffs, lead this effect to counteract the free-riding motive. In fact, at the extreme, when agents are infinitely patient ($r = 0$), the maximal under-exploitation region vanishes for $\alpha = 0$ and $\alpha = 1$; that is, $p_0^{SW}(n) = \underline{p}_0(n) = R_L/R_H$ and $p_1^{SW}(n) = \underline{p}_1(n) = 0$. However, for intermediate values of α , there is still a non-trivial region of maximal under-exploitation.

Recall also that the equilibrium intensity of exploitation k^* is strictly increasing in α as long as p_H remains in $[\underline{p}_\alpha(n), \bar{p}_\alpha(n)]$.²⁸ Therefore, while greater entanglement creates the

²⁸The equilibrium investment is continuous: $k^* = 0$ at $p_H = \underline{p}_\alpha(n)$ and $k^* = 1$ at $p_H = \bar{p}_\alpha(n)$.

free-riding problem, it also mitigates the *degree* of under-exploitation when exploitation does occur. We call the extent of free riding the distance between the action chosen by team members and the action that would be chosen by the social planner. We can conclude with the following.

Proposition 5 (Free Riding and Disentanglement).

1. *The team equilibrium is socially efficient if and only if $\alpha = 0$, $p_H \leq p_\alpha^{SW}(n)$, or $p_H \geq \bar{p}_\alpha(n)$. For all $\alpha > 0$ and $n > 1$, the team equilibrium entails under-exploration whenever $p_H \in [p_\alpha^{SW}(n), \bar{p}_\alpha(n)]$.*
2. *The size of the region of maximal under-exploitation $[p_\alpha^{SW}(n), \bar{p}_\alpha(n)]$ is strictly increasing in α for high discount rates ($r > \hat{r}$). For low discount rates ($r < \hat{r}$), the size of this region exhibits an interior maximum with respect to α .*
3. *For $p_H \in [\underline{p}_\alpha(n), \bar{p}_\alpha(n)]$, the intensity of exploitation strictly increases in α .*

Our analysis shows that free riding incentives identified by Bolton and Harris (1999) and Keller et al. (2005) persist for any positive level of entanglement α . The relation between the extent of free riding and the level of entanglement is subtle and depends on the region of the prior. For values of p_H such that agents choose interior levels of exploitation intensities k^* , these intensities are increasing in α . Therefore, an increase in α that still leaves p_H in $[\underline{p}_\alpha(n), \bar{p}_\alpha(n)]$ leads agents to come strictly closer to the socially optimal value of k . However, the size of the region of maximal under exploitation is either monotonically increasing in α or initially increasing, up to an interior maximum.

7 Concluding Remarks

This paper presents a new framework for studying experimentation that, unlike the conventional multi-arm bandit paradigm, permits agents to disentangle exploration from exploitation. Our findings are applicable to the extensively studied case of Poisson bandits, accommodating more than one risky project and general good and bad news settings. We demonstrate that the optimal policy entails full learning asymptotically, displays significant persistence, yet cannot be discerned through an index like Gittins'. The value of disentanglement depends non-monotonically on parameter values.

Absent news, disentanglement is utilized at different phases of the experimentation process, depending on the format of news. In good news settings, the agent optimally explores and exploits different projects after enough time has passed. In bad news settings, the agent may explore and exploit different projects only in the short run. After a long duration, she necessarily explores and exploits the same project.

We hope our framework can be used for a variety of applications previously studied only through the lens of the classical bandit environment, including dynamic social choice (as in [Strulovici, 2010](#)), expert delegation (as in [Guo, 2016](#)), job search (as in [Jovanovic, 1979](#); [Miller, 1984](#)), and more. Given the sharp contrast between our model’s predictions and those of the classical bandit framework, future research could develop econometric methods to distinguish between the two using population data. Future work could also enrich our model by allowing risk aversion and diversification motives, heterogeneous disentanglement levels across projects, or different information processes.

8 Appendix

8.1 Preliminaries

Proof of Proposition 0. Let U_t be the continuation payoff according to the optimal policy at time t , and let V denote the full-information payoff, when the realized quality of each project is known.

Denote by M_t the myopic payoff—the value of the favorable project—given the information the agent has at time t under the optimal policy.²⁹ Let $m_t = \mathbb{E}M_t$. The value m_t is increasing in t since the agent’s information improves over time. Also, $m_t \leq \mathbb{E}V$. Therefore, the limit $m_\infty = \lim_{t \rightarrow \infty} m_t$ exists.

For any $\varepsilon > 0$, let T be sufficiently large so that, by exploring both projects at the same rate for a period of time T , the agent can achieve a continuation payoff of $\mathbb{E}V - \varepsilon$.

The following inequalities must hold for all t :

$$(1 - e^{-rT/(1-\alpha)})m_t + e^{-rT/(1-\alpha)}(\mathbb{E}V - \varepsilon) \leq \mathbb{E}U_t \leq r \int_{\tau=0}^{\infty} e^{-r\tau} m_{t+\tau} \leq m_\infty.$$

The left inequality follows from the fact that the agent can exploit the favorable project for a period of time $T/(1 - \alpha)$ and use the exploration schedule described above to achieve a continuation payoff of at least $\mathbb{E}V - \varepsilon$ from time $t + T/(1 - \alpha)$ onwards. The right-most inequality follows from the fact that, with any strategy, the conditional expectation of the flow payoff at time $t + \tau$ is smaller than the conditional expectation of the myopic payoff at that time: the most the agent can get at any time $t + \tau$ is $m_{t+\tau}$.

By taking the limit $t \rightarrow \infty$ we obtain that $m_\infty \geq \mathbb{E}V - \varepsilon$. Since this is true for every ε and since $m_t \leq \mathbb{E}V$ for all t , we get that $m_\infty = \mathbb{E}V$. It follows that $\lim_{t \rightarrow \infty} \mathbb{E}U_t = \mathbb{E}V$. Finally, since U_t is also a sub-martingale, U_t must converge to V almost surely, as desired. ■

²⁹Formally, $M_t = \max\{p_L^t R_L, p_H^t R_H\}$, where p_L^t and p_H^t are the time- t posterior probabilities that project L and project H are good, respectively.

For much of our analysis, it will be useful to note that when an agent discounts at a rate r and receives news arriving at a rate λ , the expected discount at the time \tilde{t}_λ at which news first arrives is given by:

$$\mathbb{E}(e^{-r\tilde{t}_\lambda}) = \int_0^\infty \lambda e^{-\lambda t} e^{-rt} dt = \frac{\lambda}{r + \lambda}. \quad (4)$$

8.2 One Safe Project: Proofs and Additional Analysis

Proof of Proposition 1. Consider first the decision problem with balanced news arriving at a rate of $\lambda = \lambda_H^g = \lambda_H^b$. Denote this problem by Γ_{BL} . Absent news, the posterior that the risky project is good remains constant. Thus, the time elapsed exploring a project does not affect which project should be exploited, implying that the optimal strategy is constant as long as no news arrives. If the agent explores and exploits project H until news, the resulting expected payoff is:

$$pR_H + (1-p)\frac{\lambda}{r+\lambda}R_L.$$

The first term corresponds to a realized good project H , where, regardless of news, the agent gets rewards. The second term corresponds to a realized bad project H . The agent switches to project L only when bad news about project H arrives, with an expected discount of $\frac{\lambda}{r+\lambda}$. Analogous logic implies that the payoff of exploiting project L while exploring project H at a rate of $1 - \alpha$ is:

$$R_L + p\frac{\lambda(1-\alpha)}{r+\lambda(1-\alpha)}(R_H - R_L).$$

In particular, since the difference in payoffs between exploiting project H and project L is monotonic in p , there is a cutoff $\bar{p}(\alpha)$ such that if $p > \bar{p}(\alpha)$, the agent explores and exploits project H until news, while if $p < \bar{p}(\alpha)$, the agent explores and exploits project L indefinitely. At the cutoff $\bar{p}(\alpha)$, the agent is indifferent. Equating the two expected payoff expressions yields the value of $\bar{p}(\alpha)$ in the statement of the proposition.

We now move to a good-news setting Γ_G , where $\lambda_H^g > \lambda_H^b$, so that $\lambda = \max\{\lambda_H^g, \lambda_H^b\} = \lambda_H^g$. We claim that a strategy of the following form must be optimal: there is a T (or a \hat{p}) such that, absent news, the agent exploits project H for $t < T$ (or $p > \hat{p}$) and exploits project L for $t \geq T$ ($p \leq \hat{p}$), where T is such that exploration of project H for a duration of T leads p to decline to \hat{p} . Consider an auxiliary problem Γ_A with the following modified news process: if the agent exploits project L and explores project H at a rate of $1 - \alpha$, she receives both good and bad news on project H at a rate of λ_H^g ; If the agent exploits project H , she receives news as prescribed in problem Γ_G . The candidate strategy delivers the same payoffs in Γ_A and in Γ_G : (i) when exploiting project H , news arrives at the same rate in both problems;

(ii) when exploiting project L , the additional arrival rate of bad news is not advantageous since only good news on project H would lead the agent to switch projects. Furthermore, payoffs in Γ_A must be weakly higher than in Γ_G . Thus, if the candidate strategy is optimal in Γ_A , it must be optimal in Γ_{Good} . We now show that such a strategy is optimal in Γ_A . In this problem, absent news, beliefs do not change when exploiting project L . Therefore, if at any point it is optimal to exploit project L at time t in Γ_A , then absent news, it is also optimal to do so at later times.

We want to show that $\hat{p} = \bar{p}(\alpha)$. The value of exploiting project L is the same in Γ_A and in Γ_{BL} (with news arrival rate of $\lambda = \lambda_H^g$), whereas the value of exploiting project H is lower in Γ_A . Therefore, $\hat{p} \geq \bar{p}(\alpha)$. To show that $\hat{p} \leq \bar{p}(\alpha)$, notice that exploiting project L until news is optimal if any alternative strategy delivers weakly lower payoffs. Consider the alternative strategy that prescribes exploiting project H for a time interval Δ before returning to exploiting project L in the event that there is no news. For this alternative strategy to deliver weakly lower payoffs, it must be that

$$-\Delta r(R_L - \hat{p}R_H) + (1 - \Delta r)\hat{p}\lambda_H^g\Delta\alpha \frac{r}{r + (1 - \alpha)\lambda_H^g}(R_H - R_L) \leq 0.$$

Taking limits as $\Delta \rightarrow 0$ and simplifying we obtain that this requires that $\hat{p} \leq \bar{p}(\alpha)$.

The argument for the bad news setting with $\lambda_H^g < \lambda_H^b$ is similar and therefore omitted. ■

We now obtain expressions for the agent's payoffs, which underlie some of the results described in Section 3.2. We focus on the case of full disentanglement, $\alpha = 0$, where the cutoff posterior is $\bar{p}(0) = \frac{R_L}{R_H}$, which is relevant for our analysis there. Denote by $\Omega(p) = \frac{1-p}{p}$ the odds ratio when the agent believes project H is good with probability p .

Proposition A (Expected Payoffs with Full Disentanglement) *Consider pure news settings with $\lambda = \max\{\lambda_H^g, \lambda_H^b\}$ and $0 = \min\{\lambda_H^g, \lambda_H^b\}$. For full disentanglement, $\alpha = 0$, and posterior p that project H is good,*

1. *Good news ($\lambda = \lambda_H^g$):*
 - (a) *If $p \leq \bar{p}(0)$, expected payoffs are $R_L + p \frac{\lambda}{r+\lambda}(R_H - R_L)$;*
 - (b) *If $p \geq \bar{p}(0)$, expected payoffs are $pR_H + (1 - p) \left[\frac{\Omega(p)}{\Omega(\bar{p}(0))} \right]^{r/\lambda} \frac{\lambda}{r+\lambda} R_L$.*
2. *Bad news ($\lambda = \lambda_H^b$):*
 - (a) *If $p \leq \bar{p}(0)$, expected payoffs are $R_L + p \left[\frac{\Omega(\bar{p}(0))}{\Omega(p)} \right]^{r/\lambda} \frac{\lambda}{r+\lambda}(R_H - R_L)$;*
 - (b) *If $p \geq \bar{p}(0)$, expected payoffs are $pR_H + (1 - p) \frac{\lambda}{r+\lambda} R_L$.*

Proof of Proposition A. The terms corresponding to parts 1.a and 2.b have already been calculated in the proof of Proposition 1.

We first prove part 1.b. Consider good news settings and suppose $p \geq \bar{p}(0) = \frac{R_L}{R_H}$. Set β to satisfy $p = \beta\bar{p}(0) + (1 - \beta)$, so that β is the probability such that, if the agent explores a good project H —generating either good news and a posterior of 1, or no news—she will reach the posterior $\bar{p}(0)$. Let z be such that $\beta = pz + (1 - p)$, so that z is the probability that the agent reaches the posterior $\bar{p}(0)$, conditional on project H being good. Simple algebra yields that $z = \frac{\Omega(p)}{\Omega(\bar{p}(0))}$. Let \bar{t} denote the exploration duration of project H after which, absent news, the agent reaches the posterior $\bar{p}(0)$. Since good news arrives at an exponential rate of λ , we can write $z = e^{-\lambda\bar{t}}$. Thus, the discount factor at time \bar{t} can be written as $z^{r/\lambda}$.

Consider an auxiliary problem Γ_A in which, after reaching the posterior $\bar{p}(0)$, the agent receives balanced news on project H , with arrival rate of $\lambda = \lambda_H^g$, no matter which project she exploits. The optimal strategy in our setting Γ_{Good} is optimal in Γ_A and generates the same expected payoffs in both problems. Furthermore, in Γ_A , absent news, the agent is indifferent between exploiting project L or project H when reaching $\bar{p}(0) = \frac{R_L}{R_H}$. Thus, the payoffs from utilizing the optimal strategy in our setting coincide with those derived from the exploitation of project H until news in Γ_A .

In both our problem Γ_{Good} and the auxiliary problem Γ_A , if the agent exploits project H indefinitely, regardless of whether news arrives, she obtains the expected payoff pR_H . Once the belief reaches $\bar{p}(0)$, news in Γ_A becomes balanced, so that the posterior remains constant until news arrives. At this point, in Γ_A , the agent is indifferent between (i) continuing to exploit project H until news arrives and (ii) switching immediately to project L , which coincides with the optimal policy in Γ_{Good} . Suppose the agent continues exploiting project H until receiving bad news. If project H is in fact bad, which occurs with probability $1 - p$, then upon the arrival of bad news she receives a payoff of R_L (rather than 0). The expected discount factor at the time this news arrives is $z^{r/\lambda} \cdot \frac{\lambda}{r + \lambda}$, where the first term captures discounting up to the moment when the belief reaches $\bar{p}(0)$, and the second term follows from 4. Thus, the agent's expected payoff is:

$$pR_H + (1 - p)z^{r/\lambda} \frac{\lambda}{r + \lambda} R_L,$$

corresponding to the statement in part 1.b of the proposition.

We now turn to part 2.a. Consider bad news settings and suppose $p \leq \bar{p}(0) = \frac{R_L}{R_H}$. Similar arguments to those used for good news settings imply that if we define $\tilde{z} = \frac{\Omega(\bar{p}(0))}{\Omega(p)}$, then $z^{r/\lambda}$ captures the discount factor at the time \bar{t} it takes to reach $\bar{p}(0)$ when exploring project H without news.

Consider an auxiliary problem Γ_A analogous to the one considered before, whereby

after reaching $\bar{p}(0)$, the agent receives balanced news, with arrival rate of $\lambda = \lambda_H^b$. The optimal strategy in our setting is optimal in Γ_A and, additionally, generates the same expected payoffs in both problems. In both problems, until the posterior $\bar{p}(0)$ is reached, the agent exploits project L and can only learn bad news about project H . Therefore, for such posteriors, she does not switch her exploited project. The benefit of responding to news starting from $\bar{p}(0)$ is that, when project H is good, which occurs with probability p , when news arrives, associated with an expected discount of $\frac{\lambda}{r+\lambda}$, the agent switches to project H and receives R_H . Thus, the agent's expected payoff is:

$$R_L + pz^{r/\lambda \frac{\lambda}{r+\lambda}}(R_H - R_L),$$

which corresponds to the expression stated in part 2.a of the proposition. ■

In Section 4.2, we evaluated the expected payoff benefit of disentangling exploration from exploitation. The description of payoffs when there is full entanglement, $\alpha = 1$, follows from KRC's and KR's analysis. Recalling that $\bar{p}(1) = \frac{rR_L}{R_H(r+\lambda_H) - R_L\lambda_H}$ and using the same notation as above, we have:

Proposition B (Expected Payoffs with Full Entanglement) *Consider pure news settings with $\lambda = \max\{\lambda_H^g, \lambda_H^b\}$ and $0 = \min\{\lambda_H^g, \lambda_H^b\}$. For full entanglement, $\alpha = 1$, and posterior p that project H is good,*

1. *Good news ($\lambda = \lambda_H^g$):*
 - (a) *If $p \leq \bar{p}(1)$, expected payoffs are R_L ;*
 - (b) *If $p \geq \bar{p}(1)$, expected payoffs are $pR_H + \frac{1-p}{1-\bar{p}(1)} \left[\frac{\Omega(p)}{\Omega(\bar{p}(1))} \right]^{r/\lambda} (R_L - \bar{p}(1)R_H)$.*
2. *Bad news ($\lambda = \lambda_H^b$):*
 - (a) *If $p \leq \bar{p}(1)$, expected payoffs are R_L ;*
 - (b) *If $p \geq \bar{p}(1)$, expected payoffs are $pR_H + (1-p)\frac{\lambda}{r+\lambda}R_L$.*

8.3 Two Risky Projects: Proofs

Proof of Proposition 2. Denote by $\rho_z = \lambda_z/(r + \lambda_z)$ for $z = L, H$ the expected discount at the time at which news arrives on project z . Let $e_0 = \max\{p_L R_L, p_H R_H\}$ be the expected payoff absent any information. Let e_z be the expected payoff generated when the agent knows whether project z is good, but has no access to information on the other project. Finally, let e^* denote the expected payoff the agent receives when she has complete information on the quality of both projects.

If the agent explores project x until news, and then switches to exploring project $y \neq x$, her expected payoff is

$$(1 - \rho_x)e_0 + \rho_x(1 - \rho_y)e_x + \rho_x\rho_y e^*.$$

In particular, exploring project x first is optimal whenever

$$(1 - \rho_x)e_0 + \rho_x(1 - \rho_y)e_x \geq (1 - \rho_y)e_0 + \rho_y(1 - \rho_x)e_y.$$

Equivalently,

$$\rho_x(1 - \rho_y)(e_x - e_0) \geq \rho_y(1 - \rho_x)(e_y - e_0),$$

or

$$\frac{\rho_x}{1 - \rho_x}(e_x - e_0) \geq \frac{\rho_y}{1 - \rho_y}(e_y - e_0),$$

which translates to

$$\lambda_x(e_x - e_0) \geq \lambda_y(e_y - e_0). \quad (5)$$

If project x is favorable, then $e_0 = p_x R_x$, and $e_x = p_x R_x + (1 - p_x)p_y R_y$. Therefore, $e_x - e_0 = (1 - p_x)p_y R_y$. If project x is unfavorable, then $e_0 = p_y R_y$ and $e_x = p_x \max(R_x, p_y R_y) + (1 - p_x)p_y R_y$, so $e_x - e_0 = p_x \max(R_x, p_y R_y) - p_x p_y R_y = p_x(R_x - p_y R_y)^+ = p_x R_x(1 - \tilde{p}_x)$, where $\tilde{p}_x = \min(p_y R_y / R_x, 1)$. By substituting into equation (5), we conclude that projects are compared via $\lambda_x(1 - \tilde{p}_x)$, where $\tilde{p}_x = p_x$ when project x is favorable and $\tilde{p}_x = \min(p_y R_y / R_x, 1)$ when project x is unfavorable, as stated in the proposition. \blacksquare

Proof of Proposition 3. Suppose project x is favorable, so that $p_x R_x \geq p_y R_y$. We need to show that it is optimal for the agent to either explore project x for a period T absent news, with $0 \leq T \leq \bar{t}_x(p_L, p_H)$, after which project y is explored until news is received; or to explore project x until news arrives, denoted as exploring x for a duration $T = \infty$. Whenever the agent receives news on one project, but not the other, she reverts to exploring the uncertain project. Once the agent learns the realization of both projects, the exploration strategy has no payoff impacts. For simplicity, we assume the agent reverts to exploring project x in that case. We denote by σ_T the strategy induced by each such $T \in [0, \bar{t}_H(p_L, p_H)] \cup \{\infty\}$.

Given the original decision problem Γ_{Good} , consider an auxiliary problem Γ_A with the following modified news process:

1. If the agent explores project y , she receives both good and bad news at a rate of λ_y^g .
2. If the agent explores project x and, by that time, has already explored it for at least $\bar{t}_x(p_L, p_H)$, then she receives both good and bad news at rate λ_x^g .

3. If the agent explores project x and by that moment she has explored project x for a period smaller than $\bar{t}_x(p_L, p_H)$, she receives good news at a rate of λ_x^g and bad news at a rate of λ_x^b .

Under any exploration strategy, and at any point in time, the agent is at least as well informed in Γ_A as in Γ_{Good} . In particular, the optimal expected payoff that can be achieved in Γ_A is weakly higher than the optimal expected payoff that can be achieved in Γ_{Good} .

Claim A1 For any $T \in [0, \bar{t}_x(p_L, p_H)] \cup \{\infty\}$, the strategy σ_T generates the same expected payoff in Γ_A as it does in Γ_{Good} .

Proof of Claim A.1 For $T \leq \bar{t}_x(p_L, p_H)$, the agent receives information at the same arrival rate in both Γ_{Good} and Γ_A during the initial duration of T . If news arrives during that period, the resulting optimal exploitation is identical in both problems: if good news arrives, exploit project x indefinitely if $x = H$ or until good news arrives from project y if $x = L$, and if bad news arrives from project x , then exploit project y indefinitely. Absent news, project x remains favorable when the agent switches to exploring project y . Thus, from then on, only good news on project y alters her exploitation. Since the arrival rate of good news on project y is the same in Γ_{Good} and Γ_A , the resulting expected payoffs coincide as well.

Suppose now that $T = \infty$, so that the agent explores project x until receiving news. Until time $\bar{t}_x(p_L, p_H)$, news arrives at the same rate in both Γ_{Good} and Γ_A . Absent news, at time $\bar{t}_x(p_L, p_H)$, the agent is indifferent between the two projects: they are both favorable. At any $t > \bar{t}_x(p_L, p_H)$, absent news, it is optimal to exploit project y in both Γ_{Good} and Γ_A . Only good news on project x then alters exploitation, and good news arrives at the same rate in Γ_{Good} and Γ_A . Therefore, the resulting expected payoffs coincide.

Claim A.2 There exists $T \in [0, \bar{t}_x(p_L, p_H)] \cup \{\infty\}$ such that σ_T is optimal in Γ_A .

Proof of Claim A.2 In Γ_A , if the agent explores project y and sees no news, her belief about the quality of project y does not change. Therefore, by dynamic-programming principles, if it is optimal for the agent to explore project y at any point then, absent news, it is also optimal to explore project y at any later point. Similarly, if the agent has explored project x for a period of at least $\bar{t}_x(p_L, p_H)$, continuing to explore project x until news is optimal. The conclusion follows.

Claims A.1 and A.2 illustrate the optimality of the class of strategies specified in the proposition. We now turn to showing that in settings with pure good news on at least one project, exploration switches only upon receiving news.

Claim A.3 If $\lambda_x^b = 0$, there exists an optimal strategy in Γ_{Good} with $T = 0$ or $T = \infty$.

Proof of Claim A.3 Suppose Alex explores project y from the start, i.e., Alex uses the strategy σ_0 . Bailey, facing the same decision problem, uses σ_T with $0 < T \leq \bar{t}_x(p_L, p_H)$. We claim that Alex has a higher expected payoff than Bailey.

Consider Alexis and Baylor, who face a coupled problem. Baylor, like Bailey, explores project x for a period of T or until receiving news. Denote by ω the random time when Baylor either receives news on project x or a period of T has transpired (so that ω is the minimum between T and the arrival time of news on project x , which, conditional on the project's quality, is distributed exponentially with arrival rates of $\lambda_x^b = 0$ or λ_x^g). Like Bailey, after time ω , Baylor switches to exploring project y . Unlike Bailey, at any time $t \geq \omega$, Baylor receives the news Alexis has received at time $t - \omega$ on project y . Alexis, like Alex, explores project y until news. Let τ be the random variable that represents the first arrival of news on project y for Alex (distributed exponentially with parameters λ_y^b or λ_y^g). At any time $t \in [\tau, \tau + \omega]$, Alexis receives the news Baylor has received on project x at time $t - \tau$, after which Alexis receives news independently on project x . Thus, Alexis' and Baylor's information is coupled. Since Alex and Bailey's news arrivals are independent and identical, Alexis receives the same expected payoff as Alex and Baylor receives the same expected payoffs as Bailey. We now show that Alexis receives a weakly higher expected payoff than Baylor.

Conditional on τ , at any moment t such that $0 \leq t \leq \omega + \tau$, Baylor does not learn whether project y is good or bad. Since $\lambda_x^b = 0$, Baylor can only receive good news or no news about project x until such time t . Since $T \leq \bar{t}_x(p_L, p_H)$, in either case, Baylor continues exploiting project x . Alexis, however, exploits project x until time τ , when a switch to project y may be optimal when news about project y is good. Therefore, conditional on ω and τ , up to time $\min\{\omega, \tau\}$, Alexis' and Baylor's expected payoffs coincide, whereas over the period between $\min\{\omega, \tau\}$ and $\omega + \tau$, Alexis' expected payoff is weakly higher than Baylor's. At any moment t such that $t > \omega + \tau$, both Alexis and Baylor know whether project y is good or bad and have explored project x for a period $t - \tau$, receiving the same information ex-ante.³⁰ Therefore, at moments t such that $t > \omega + \tau$, Alexis' and Baylor's expected payoffs are the same. Therefore, Alexis' expected payoff is weakly higher than Baylor's, as required. ■

Proof of Claim 1. We show that the unique optimal strategy is the strategy described in Proposition 3 and that the optimal switching point of exploration is generically unique. It suffices to prove this for an auxiliary problem as in the proof of Proposition 3, in which balanced news from the unfavorable arm y arrives at rate λ_y^g .

Fix all parameters except p_x . Consider the difference in the agent's expected payoffs between a strategy that explores project x for a small duration $\Delta > 0$ and then switches to

³⁰Recall that we assumed the agent explores the ex-ante favorable project x after receiving news on project y , even when having received news on project x as well.

exploring project y forever and a strategy that explores project y forever. This difference is given by

$$\Delta \left((1-p_x) \lambda_x^b \frac{r}{r+\lambda_y^g} p_y R_y - r p_y \frac{\lambda_y^g}{r+\lambda_y^g} (R_y - p_x R_x) \right) + o(\Delta) = \Delta \frac{r}{r+\lambda_y^g} p_y R_y D(p_x) + o(\Delta), \quad (6)$$

with $D(p_x) = \lambda_x^b(1-p_x) - \lambda_y^g(1-p_x R_x/R_y)$. The first term is the benefit from the fact that the agent learns that project x is bad and then, until receiving news from project y , exploits project y (instead of getting a payoff of 0 from project x). The second term corresponds to the costs incurred when project y is good and there is a duration Δ in which the agent would have exploited project y , but exploits project x instead.

In the auxiliary problem, and as long as project x remains favorable, an optimal strategy must explore project x if $D(p_x) > 0$ and cannot switch from exploring project x to exploring project y if $D(p_x) < 0$. Thus, an optimal strategy entails an exploration switch from project x to project y at the point in which $D(p_x) = 0$.

Now, $D(p_x)$ is linear in p_x with slope $-\lambda_x^b + \lambda_y^g R_x/R_y$. As long as $\lambda_x^b R_y \neq \lambda_y^g R_x$, which is generically the case, $D(p_x)$ has a non-trivial slope and there is only one point in which exploration can switch from project x to project y . ■

Proof of Claim 2. If project L is explored, only good news yields a switch of the exploited project. If project H is explored, absent news, the agent switches her exploited project after $\bar{t}_H(p_L, p_H)$ has passed, when she is indifferent between the expected payoffs of both projects. We now characterize $\bar{t}_H(p_L, p_H)$, where we drop the arguments when there is no risk of confusion.

By definition, after a duration \bar{t}_H of exploring project H , the agent's posterior that project H is good declines to qp_H , where $q = \frac{p_L R_L}{p_H R_H} \in (0, 1)$. Certainly, if the agent receives good news on project H before reaching indifference, the corresponding posterior jumps to 1. The conditional probability that the agent reaches indifference when exploring project H , conditional on project H being good, is therefore $\frac{q(1-p_H)}{1-qp_H}$.³¹ The exponential distribution of news then yields:

$$e^{-\lambda_H^g \bar{t}_H} = \frac{q(1-p_H)}{1-qp_H}.$$

The discount at the indifference time \bar{t}_H is given by $w = e^{-r\bar{t}_H}$. As before, let $\rho_z = \lambda_z^g/(r+\lambda_z^g)$,

³¹The arguments are reminiscent of those used in the proof of Proposition A. Set β to satisfy $p_H = \beta q p_H + (1-\beta)$, so that β is the probability such that, if the agent explores project H , she will reach a time at which she is indifferent between the projects. Set z to satisfy $\beta = p_H z + (1-p_H)$, so that z is the conditional probability that the agent reaches indifference, conditional on project H being good. Simple algebra yields the formula.

$z = L, H$, denote the expected discount at the time \bar{t}_H at which news first arrives when the arrival rate is λ_z^g .

Suppose the agent explores project H indefinitely. As argued in the proof of Proposition 3, her payoff coincides with the payoff of an agent who, after time \bar{t}_H , sees all news from project H —good or bad, at a (balanced) rate λ_H^g . So, a-priori, the agent expects to receive $e_0 = p_H R_H$ up to a time that is exponentially distributed with parameter λ_H^g beyond the indifference time \bar{t}_H . After that time, she receives $e_H = p_H R_H + (1 - p_H)p_L R_L$. The agent's expected payoff is therefore:

$$(1 - w\rho_H)e_0 + w\rho_H e_H = e_0 + w\rho_H(e_H - e_0).$$

Now, suppose the agent explores project L instead. Define, analogously, $e_L = p_L R_L + p_H R_H(1 - p_L)$ to be the expected value from exploring project L upon indifference.

As shown in the proof of Proposition 3, the agent's expected payoff is the same as in the balanced news setting, and equals

$$(1 - \rho_L)e_0 + \rho_L(1 - \rho_H)e_L + \rho_L\rho_H e_H = e_0 + \rho_L(1 - \rho_H)(e_L - e_0) + \rho_L\rho_H(e_H - e_0).$$

Thus, it is optimal to explore project H if and only if

$$\rho_H(w - \rho_L)(e_H - e_0) \geq \rho_L(1 - \rho_H)(e_L - e_0).$$

The statement of the claim then follows. ■

Proof of Proposition 4. The proof follows several claims:

Claim B.1 *There exists an optimal policy with the property that, if it is optimal to explore project H at some point when it is favorable, then, from that point on, it is optimal to explore project H until news is received.*

Proof of Claim B.1 Consider an auxiliary problem Γ_A that coincides with the original problem Γ_{Bad} with the following modification: if the agent explores project H and project H is currently weakly favorable, the agent receives both good news and bad news on project H at a rate of λ_H^b . In particular, the agent has more information in Γ_A than in Γ_{Bad} .

Any strategy described in the statement of the proposition generates the same payoff in Γ_A as it does in Γ_{Bad} . Indeed, if project H is favorable, and the agent explores it, then in both Γ_{Bad} and Γ_A , project H would remain favorable as long as no bad news arrive.

It suffices to show that, under the optimal strategy in Γ_A , once the agent starts exploring project H , she continues doing so until receiving news. Indeed, if the agent explores project H in the auxiliary problem when project H is currently favorable, then the state

variable—her posterior—does not change. By dynamic programming principles, it must be optimal to continue exploring project H until news arrives.

Claim B.2 If project L is favorable at the outset, then, at any point in which project H becomes strictly favorable, it is optimal to explore project H until receiving news.

Proof of Claim B.2 Project H becomes strictly favorable when one of the following occurs. First, upon arrival of bad news about project L or good news about project H , project H becomes favorable and exploring either project is optimal. The second option is that the agent explores project H . Since $\lambda_H^b > \lambda_H^g$, over time, the agent becomes more optimistic about project H . In this case, by Claim B.1, the agent should continue exploring project H .

For the next step of the proof, consider a balanced news setting with arrival rates λ_H^g for project H and λ_L^b for project L that starts with prior probabilities p_H and p_L such that L is favorable. Let \hat{p}_L be such that the agent explores project H if $p_L \geq \hat{p}_L$, holding all other parameters fixed. By Proposition 2,

$$\lambda_H^g(1 - \tilde{p}_H) = \lambda_L^b(1 - \hat{p}_L), \quad (7)$$

where $\tilde{p}_H = \hat{p}_L R_L / R_H$.

Claim B.3 If project L is favorable and $p_L > \hat{p}_L$, then it is optimal to explore project H until news.

Proof of Claim B.3 Consider an auxiliary problem Γ_B , a modification of the original problem Γ_{Bad} in which exploring project L generates balanced news at a rate of λ_L^b . The agent is weakly better off in Γ_B relative to the original problem Γ_{Bad} since she has access to information that arrives at higher rates. Furthermore, exploring project H until news generates the same payoff in Γ_B as it does in Γ : news about project H arrives at the same rate in both problems and, in both, exploiting project H (or project L) forever once project H is observed to be good (or bad) maximizes expected payoffs.

Suppose that project L is favorable and $p_L > \hat{p}_L$. We show that, in Γ_B , the agent optimally explores project H until news. Assume, by way of contradiction, that it is optimal to explore project L in Γ_B .³² Absent news, exploring project L does not alter the agent's beliefs about the projects' quality and, therefore, it must be optimal to explore project L until news. Consider a deviation to first exploring project H for a short interval Δ and then exploring project L until news, where Δ is sufficiently small so that, absent news during the time period Δ , project L remains favorable. We claim that this deviation improves

³²Since news is balanced on project L , if it is optimal to explore project L at any posterior, it is optimal to continue exploring project L as long as news does not arrive.

payoffs. Suppose Alex plays the candidate strategy—exploring project L until news—and Bailey follows the deviation.

Let τ_L and τ_H denote the random variables corresponding to the first arrival time of news on project L and project H , respectively, where arrival rates are those specified in the auxiliary problem Γ_B . Both Alex and Bailey receive news on project $z = L, H$ after exploring project z for a duration τ_z . Thus, Alex's and Bailey's information is coupled. Furthermore, Alex's and Bailey's payoffs from the suggested strategies are the same as before.

The difference between Bailey's and Alex's payoffs is then:

$$p_H \lambda_H^g \Delta \frac{r}{r + \lambda_L^b} (R_H - p_L R_L) - (1 - p_L) \frac{\lambda_L^b}{r + \lambda_L^b} r \Delta p_H R_H + o(\Delta).$$

The first term corresponds to the case in which Bailey receives good news on project H in the initial duration of Δ (occurring with probability $p_H \lambda_H^g \Delta$), while Alex is delayed in learning about project H until receiving news on project L at time τ_L (occurring at a rate of λ_L^b whether project L is good or bad). The expected discounted weight of that duration is $1 - \frac{\lambda_L^b}{r + \lambda_L^b} = \frac{r}{r + \lambda_L^b}$ (see equation (4)). The second term corresponds to project L being bad. In that case, conditional on not receiving news in the first period of Δ (occurring with probability $1 - \lambda_H^g \Delta$), Bailey would be delayed by Δ relative to Alex in learning that project L is bad. Observing that project L is bad would lead either agent to exploit project H , which generates an expected payoff of $p_H R_H \Delta$ (up to $o(\Delta)$ due to updating on project H during the initial period of Δ). The relevant expected discount at τ_L , when Alex learns that project L is bad, is $\frac{\lambda_L^b}{r + \lambda_L^b}$.

Reorganizing terms implies that the payoff difference is:

$$\Delta \frac{r}{r + \lambda_L^b} p_H R_H \left(\lambda_H^g (1 - \hat{p}_H) - \lambda_L^b (1 - p_L) \right) + o(\Delta) > 0,$$

where the inequality follows from our assumption that $p_L > \hat{p}_L$. The conclusion of Claim B.3 then follows using Claim B.1.

Claim B.4 *If project L is favorable and $p_L \leq \hat{p}_L$, it is optimal to explore project L for some period, and then explore project H until news.*

Proof of Claim B.4 Consider an optimal strategy, and let T be the first time such that, according to this strategy, if no news arrives up to time T , either project H becomes favorable or the posterior that project L is good reaches \hat{p}_L . By Claims B.2 and B.3, if no news arrives by time T , it is optimal to explore project H . We claim that, before time T , it is optimal to explore project L for some period and then switch to exploring project H .

Suppose, toward a contradiction, that the claim is violated. Then, there must be a sufficiently small Δ , a fraction $\beta > 0$, and times $t' < t'' < T$ with $t' - \Delta > 0$ and $t'' + \Delta < T$,

such that the agent optimally explores project H for an amount of time $\beta\Delta$ in $I' = [t' - \Delta, t']$ and explores project L for an amount of time $\beta\Delta$ in $I'' = [t'', t'' + \Delta]$.

We now show that swapping the order of these $\beta\Delta$ exploration resources between the intervals I' and I'' improves the agent's expected payoff. Indeed, suppose Alex plays the candidate strategy and Bailey performs the swap, and their news are coupled as follows:

1. All news coming from exploration that was not interchanged, which we call *regular news*, are the same for Alex and Bailey.
2. The *additional news on project L* that Bailey receives from the additional $\beta\Delta$ exploration during I' is received by Alex during I''
3. The *additional news on project H* that Alex receives from the additional $\beta\Delta$ exploration during I' is the news received by Bailey during I'' .

We need to show that Bailey's payoff is higher than Alex's. We will, in fact, show that this is the case even if Bailey does not play optimally: we assume that if Bailey receives additional good news about project L , Bailey ignores this news and switches to exploring project H only when either regular good news arrives about project L or when Alex receives the additional good news about project L (in which case Alex also switches to only exploring project H).

Until time T , Alex and Bailey both exploit project L unless they received bad news from project L or good news from project H . They gain different payoffs at time $t \in [t', t'']$ only if they receive no regular news up to time t and either

1. bad news on project L is received by Bailey over I' , in which case Bailey exploits project H , while Alex exploits project L ; or
2. good news on project H is received only by Alex over I' , in which case Alex exploits project H , while Bailey exploits project L .

Therefore, the difference in expected payoffs is

$$\beta\Delta \int_{t'}^{t''} re^{-rt} \rho(t) \left[\lambda_L^b (1 - p_L(t)) p_H(t) R_H - \lambda_H^g p_H(t) (R_H - p_L(t) R_L) \right] dt + O(\Delta),$$

where $\rho(t)$ is the probability that there were no regular news until time t ; the probabilities $p_H(t)$ and $p_L(t)$ are, respectively, the conditional probabilities that projects H and L are good given this event; and $\tilde{p}_H(t) = p_L(t) R_L / R_H$. Rearranging terms, this payoff difference equals:

$$\beta\Delta \int_{t'}^{t''} re^{-rt} \rho(t) p_H(t) R_H \left[\lambda_L^b (1 - p_L(t)) - \lambda_H^g (1 - \tilde{p}_H(t)) \right] dt + o(\Delta) > 0,$$

where the inequality follows from the fact that $p_L(t) < \hat{p}_L$ for every $t < t''$.

Claim B.5 If project H is favorable, it is optimal to explore project L for some period, and then explore project H until news.

Proof of Claim B.5 Suppose $R_L > p_H R_H \geq p_L R_L$. From Claim B.1, once the agent starts exploring project H, it is optimal to do so until news. Towards a contradiction, suppose the agent explores project L until news. Absent news, at any time $t > \bar{t}_L(p_L, p_H)$, project L becomes favorable. Claims B.3 and B.4 then lead to a contradiction.

If $p_H R_H \geq R_L$, news on project L cannot generate a switch in the agent's exploited project and exploring project L indefinitely is dominated. The claim then follows directly from Claim B.1.

The proposition follows from Claims B.3, B.4, and B.5. ■

References

- Audibert, J.-Y., S. Bubeck, and R. Munos (2010). Best arm identification in multi-armed bandits. In *COLT*, pp. 41–53.
- Bardhi, A., Y. Guo, and B. Strulovici (2020). Early-career discrimination: Spiraling or self-correcting? *mimeo*.
- Bergemann, D. and U. Hege (1998). Venture capital financing, moral hazard, and learning. *Journal of Banking & Finance* 22(6-8), 703–735.
- Bergemann, D. and J. Valimaki (2006). Bandit problems.
- Bolton, P. and C. Harris (1999). Strategic experimentation. *Econometrica* 67(2), 349–374.
- Bubeck, S., R. Munos, and G. Stoltz (2011). Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science* 412(19), 1832–1852.
- Carnehl, C. and J. Schneider (2023). On risk and time pressure: When to think and when to do. *Journal of the European Economic Association* 21(1), 1–47.
- Che, Y.-K. and J. Hörner (2018). Recommender systems as mechanisms for social learning. *The Quarterly Journal of Economics* 133(2), 871–925.
- Che, Y.-K. and K. Mierendorff (2019). Optimal dynamic allocation of attention. *American Economic Review* 109(8), 2993–3029.
- Crawford, G. S. and M. Shum (2005). Uncertainty and learning in pharmaceutical demand. *Econometrica* 73(4), 1137–1173.
- Currie, J. M. and W. B. MacLeod (2020). Understanding doctor decision making: The case of depression treatment. *Econometrica* 88(3), 847–878.
- Damiano, E., H. Li, and W. Suen (2020). Learning while experimenting. *The Economic Journal* 130(625), 65–92.
- Dickstein, M. J. et al. (2021). *Efficient provision of experience goods: Evidence from antidepressant choice*.
- Eliaz, K., D. Fershtman, and A. Frug (2024). On optimal scheduling. *American Economic Journal: Microeconomics*.
- Georgiadis-Harris, A. (2024). Preparing to act. *mimeo*.
- Gittins, J., K. Glazebrook, and R. Weber (2011). *Multi-armed bandit allocation indices*. John Wiley & Sons.
- Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 41(2), 148–164.
- Gittins, J. C. and D. M. Jones (1979). A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika* 66(3), 561–565.
- Guo, Y. (2016). Dynamic delegation of experimentation. *American Economic Review* 106(8), 1969–2008.
- Hörner, J. and L. Samuelson (2013). Incentives for experimenting agents. *The RAND Journal of Economics* 44(4), 632–663.
- Jovanovic, B. (1979). Job matching and the theory of turnover. *Journal of Political Economy* 87(5, Part 1), 972–990.
- Keller, G. and S. Rady (2015). Breakdowns. *Theoretical Economics* 10(1), 175–202.
- Keller, G., S. Rady, and M. Cripps (2005). Strategic experimentation with exponential bandits. *Econometrica* 73(1), 39–68.
- Liang, A. and X. Mu (2020). Complementary information and learning traps. *The Quarterly*

- Journal of Economics* 135(1), 389–448.
- Liang, A., X. Mu, and V. Syrgkanis (2022). Dynamically aggregating diverse information. *Econometrica* 90(1), 47–80.
- Maćkowiak, B., F. Matějka, and M. Wiederholt (2023). Rational inattention: A review. *Journal of Economic Literature* 61(1), 226–273.
- Miller, R. A. (1984). Job matching and occupational choice. *Journal of Political Economy* 92(6), 1086–1120.
- Robbins, H. (1952). Some aspects of the sequential design of experiments.
- Rothschild, M. (1974). A two-armed bandit theory of market pricing. *Journal of Economic Theory* 9(2), 185–202.
- Sims, C. A. (2003). Implications of rational inattention. *Journal of monetary Economics* 50(3), 665–690.
- Strulovici, B. (2010). Learning while voting: Determinants of collective experimentation. *Econometrica* 78(3), 933–971.
- Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3-4), 285–294.
- Wald, A. (1947). Foundations of a general theory of sequential decision functions. *Econometrica*, 279–313.
- Zhuo, R. (2023). Exploit or explore? an empirical study of resource allocation in research labs. *mimeo*.