Online Appendix for "Disentangling Exploration from Exploitation"

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

November 21, 2025

Abstract. We present additional analyses pertaining to the special case of our environment in which one project is safe. We also show that non-indexability occurs for arbitrary disentanglement levels and that our characterization of the optimal policy is robust to small perturbations in the disentanglement level. Next, we illustrate the tightness of our bound on duration of exploration of the favorable project in good news settings. Further, we provide derivations speaking to the region of maximal under-exploitation in teams. Finally, we illustrate how our framework translates to an agency problem between an agent who collects information and an agent who has executive power to select actions.

Keywords: Exploration and Exploitation, Poisson Bandits, Gittins Index

JEL codes: C73, D81, D83, O35

[*]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 Comparative Statics when One Project is Safe

In this section, we provide details for several observations regarding the special case in which one project is safe.

1.1 Comparison of Good and Bad News Settings

We start by showing that the value of disentanglement is higher under pure good news than under pure bad news, holding other parameters constant.

For any set of parameters, denote by $V_\alpha^G(p, r/\lambda)$ and $V_\alpha^B(p, r/\lambda)$ denote the expected payoffs in pure good news and pure bad news, respectively, when the level of disentanglement is α, the prior that the risky project is good is p and the ratio of discount rate to news arrival rate is r/λ. The value of disentanglement is then $\Delta V^X(p, r/\lambda) = V_0^X(p, r/\lambda) - V_1^X(p, r/\lambda)$ for $X = G, B$.

Proposition OA_1 *The value of disentanglement is greater in good news settings than in bad news settings: $\Delta V^G(p, r/\lambda) > \Delta V^B(p, r/\lambda)$ for all p, r, and λ.*

Proof of Proposition OA_1. We use Propositions A and B in the main text's Appendix. The cutoff probabilities $\bar{p}(0)$ and $\bar{p}(1)$ are the same for any r or λ (corresponding to arrival rate of good or bad news). There are therefore three cases to consider:

1. $p \le \bar{p}(1) \le \bar{p}(0)$**:** In this case,

$$\Delta V^G(p, r/\lambda) = R_L + p\frac{\lambda}{r+\lambda}(R_H - R_L) - R_L = p\frac{\lambda}{r+\lambda}(R_H - R_L),$$

whereas

$$\Delta V^B(p, r/\lambda) = R_L + p\left[\frac{\Omega(\bar{p}(0))}{\Omega(p)}\right]^{r/\lambda}\frac{\lambda}{r+\lambda}(R_H - R_L) - R_L = p\left[\frac{\Omega(\bar{p}(0))}{\Omega(p)}\right]^{r/\lambda}\frac{\lambda}{r+\lambda}(R_H - R_L).$$

Since $p \le \bar{p}(0)$, it follows that $\Omega(p) > \Omega(\bar{p}(0))$ and the result follows.

2. $\bar{p}(1) \le p \le \bar{p}(0)$**:** In this case,

$$\Delta V^G(p, r/\lambda) = R_L + p\frac{\lambda}{r+\lambda}(R_H - R_L) - pR_H - \frac{1-p}{1-\bar{p}(1)}\left[\frac{\Omega(p)}{\Omega(\bar{p}(1))}\right]^{r/\lambda}(R_L - \bar{p}(1)R_H),$$

whereas

$$\Delta V^B(p, r/\lambda) = R_L + p\left[\frac{\Omega(\bar{p}(0))}{\Omega(p)}\right]^{r/\lambda}\frac{\lambda}{r+\lambda}(R_H - R_L) - pR_H - (1-p)\frac{\lambda}{r+\lambda}R_L.$$

Since $\bar{p}(1) \le p \le \bar{p}(0)$, it follows that $\Omega(\bar{p}(1)) \ge \Omega(p) \ge \Omega(\bar{p}(0))$. Therefore, the first three terms of $V^G(p, r/\lambda)$ are greater than the first three terms of $V^B(p, r/\lambda)$. Further-

more, the last term in $V^G(p, r/\lambda)$ is smaller than

$$\frac{1-p}{1-\bar{p}(1)}(R_L - \bar{p}(1)R_H).$$

Now, recall that $\bar{p}(1) = \frac{rR_L}{R_H(r+\lambda)-R_L\lambda}$. Plugging in and reorganizing terms, we get that the bound on the last term of $V^G(p, r/\lambda)$ is

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

and the result follows.

3. $p \geq \bar{p}(0)$**:** In this case,

$$\Delta V^G(p, r/\lambda) = (1-p)\left[\frac{\Omega(p)}{\Omega(\bar{p}(0))}\right]^{r/\lambda}\frac{\lambda}{r+\lambda}R_L - \frac{1-p}{1-\bar{p}(1)}\left[\frac{\Omega(p)}{\Omega(\bar{p}(1))}\right]^{r/\lambda}(R_L - \bar{p}(1)R_H),$$

whereas

$$\Delta V^B(p, r/\lambda) = 0$$

and the result follows directly.

■

As noted in the text, intuitively, disentanglement is valuable only when the agent seeks to exploit the safe project L. This exploitation comes at the cost of reduced exploration of project H, where uncertainty remains. Under these conditions, good news about project H is 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 pure good news settings.

1.2 Marginal Impacts of Disentanglement Level

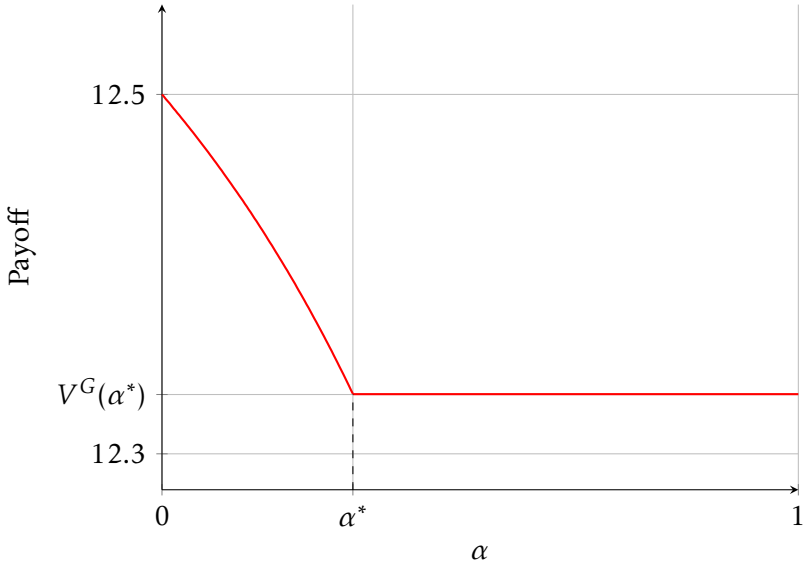
In the main text, we view the level of disentanglement, α, as exogenously given. One could contemplate endogenizing α. For that, it is useful to understand how the value of disentanglement changes with α.

As a preview, increasing α tightens the agent's constraint, which reduces her expected payoffs. However, as we note in the paper, the relationship between expected payoffs and α is neither concave nor convex. To see why, 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 given by:

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

which is strictly decreasing and concave in α. Therefore expected payoffs are neither concave nor convex in α over the interval $[0,1]$. The following figure, depicting payoffs for $p_H = 0.6, R_L = 10, R_H = 15, r = 1$, and $\lambda = 5$, illustrates the shape of the payoffs:



For simplicity, consider the balanced news setting, where payoffs are given by Equation (1) above. If we let $\beta = 1 - \alpha$, so that increasing β increases disentanglement, then in the region in which $\alpha < \alpha^*$ as described above, the derivative (with respect to r/λ) of the marginal payoff (with respect to β) is given by:

$$\frac{\beta - r/\lambda}{(\beta + r/\lambda)^3}.$$

In particular, the marginal payoff (with respect to β) is single-peaked in r/λ for $\alpha < \alpha^*$. When $\alpha > \alpha^*$, payoffs are flat in α and the marginal benefit from decreasing or increasing α is 0.

Importantly, the non-concavity of payoffs with respect to α means that standard first-order conditions cannot automatically be used to optimize the choice of α, even in the presence of costs that are convex in α.

1.3 Comparative Statics with Respect to the Prior

In the main text, we show that the normalized benefit of disentanglement is maximized at intermediate values of the prior. We now illustrate that the benefits of disentanglement are, in fact, single-peaked.

Proposition OA_2 *The value of disentanglement is single-peaked in p. That is, $\Delta\Pi^X(p, r/\lambda)$ is single-peaked in p for all r, λ, and $X = G, B$. Furthermore, for all r and λ,*

- *In good news settings, the peak of the normalized value of disentanglement, $\Delta\Pi^B(p, r/\lambda)$, is in the interval $[\bar{p}(1), \bar{p}(0)]$*
- *In bad news settings, the peak of the normalized value of disentanglement, $\Delta\Pi^B(p, r/\lambda)$, is at $\bar{p}(1)$.*

Proof. Consider good news settings first. For any r and λ, there are three regions of priors p to consider as in the proof of Proposition OA_1:

1. $p \leq \bar{p}(1) \leq \bar{p}(0)$**:** In this case,

$$\Delta V^G(p, r/\lambda) = R_L + p\frac{\lambda}{r+\lambda}(R_H - R_L) - R_L = p\frac{\lambda}{r+\lambda}(R_H - R_L),$$

which is an increasing function of p. Furthermore,

$$\frac{p}{pR_H + (1-p)R_L} = \frac{1}{R_H - R_L + R_L/p}$$

is increasing in p, implying that $\Delta\Pi^G(p, r/\lambda)$ is increasing in p as well.

2. $\bar{p}(1) \leq p \leq \bar{p}(0)$**:** In this case,

$$\Delta V^G(p, r/\lambda) = R_L + p\frac{\lambda}{r+\lambda}(R_H - R_L) - pR_H - \frac{1-p}{1-\bar{p}(1)}\left[\frac{\Omega(p)}{\Omega(\bar{p}(1))}\right]^{r/\lambda}(R_L - \bar{p}(1)R_H),$$

Differentiating and condensing terms, we get:

$$\frac{\partial V^G(p, r/\lambda)}{\partial p} = A + B\Omega(p)^{r/\lambda - 1}\Big[\Omega(p) - (1-p)\frac{r}{\lambda}\Omega'(p)\Big],$$

where $A = -\frac{r}{r+\lambda}R_H - \frac{\lambda}{r+\lambda}R_L < 0$ and $B = \frac{1}{(1-\bar{p}(1))\Omega(\bar{p}(1))^{r/\lambda}}\Big[R_L - \bar{p}(1)R_H\Big] > 0$. Differentiating again, we then get:

$$\frac{\lambda}{Br}\frac{\partial^2 V^G(p, r/\lambda)}{\partial p^2} = \Omega(p)^{r/\lambda - 2}\Big[2\Omega(p)\Omega'(p) - (1-p)(\frac{r}{\lambda} - 1)(\Omega'(p))^2 - (1-p)\Omega(p)\Omega''(p)\Big].$$

Recall that $\Omega(p) = \frac{1-p}{p}$, so that $\Omega'(p) = -\frac{1}{p^2}$ and $\Omega''(p) = \frac{2}{p^3}$. The term in the square

parentheses above then corresponds to:

$$-\frac{2}{p^3}+\frac{2}{p^2}-\frac{(1-p)(\frac{r}{\lambda}-1)}{p^4}-(1-p)\left(\frac{2}{p^4}-\frac{2}{p^3}\right)=-\frac{(1-p)r}{p^4\lambda}-\frac{1}{p^4}+\frac{1}{p^3}<0.$$

Therefore, $V^G(p,r/\lambda)$ is concave within the region. The normalized value of disentanglement, $\Delta\Pi^G(p,r/\lambda)$, is the ratio of a concave function and a positive linear function. Hence, it is quasi-concave and single-peaked.

3. $p \geq \bar{p}(0)$**:** In this case,

$$\Delta V^G(p,r/\lambda)=(1-p)\left[\frac{\Omega(p)}{\Omega(\bar{p}(0))}\right]^{r/\lambda}\frac{\lambda}{r+\lambda}R_L-\frac{1-p}{1-\bar{p}(1)}\left[\frac{\Omega(p)}{\Omega(\bar{p}(1))}\right]^{r/\lambda}(R_L-\bar{p}(1)R_H).$$

Both terms are decreasing in p, implying that $V^G(p,r/\lambda)$ is decreasing in p. Since the normalizing factor, $pR_H+(1-p)R_L$ is increasing in p, it follows that $\Delta\Pi^G(p,r/\lambda)$ is also decreasing in this range.

Since $V^G(p,r/\lambda)$ is continuous in p, its single-peakedness follows.

Consider bad news settings. For any r and λ, there are again three regions of priors p to analyze:

1. $p \leq \bar{p}(1) \leq \bar{p}(0)$**:** In this case,

$$\Delta V^B(p,r/\lambda)=R_L+p\left[\frac{\Omega(\bar{p}(0))}{\Omega(p)}\right]^{r/\lambda}\frac{\lambda}{r+\lambda}(R_H-R_L)-R_L=p\left[\frac{\Omega(\bar{p}(0))}{\Omega(p)}\right]^{r/\lambda}\frac{\lambda}{r+\lambda}(R_H-R_L),$$

which is increasing in p. Furthermore,

$$\frac{p}{pR_H+(1-p)R_L}=\frac{1}{R_H-R_L+R_L/p}$$

is increasing in p, implying that $\Delta\Pi^B(p,r/\lambda)$ is increasing in p as well.

2. $\bar{p}(1) \leq p \leq \bar{p}(0)$**:** In this case,

$$\Delta V^B(p,r/\lambda)=R_L+p\left[\frac{\Omega(\bar{p}(0))}{\Omega(p)}\right]^{r/\lambda}\frac{\lambda}{r+\lambda}(R_H-R_L)-pR_H-(1-p)\frac{\lambda}{r+\lambda}R_L.$$

Denote by $a=r/\lambda$. The derivative with respect to p is given by:

$$\frac{\partial\Delta V^B(p,r/\lambda)}{\partial p}=\frac{R_H-R_L}{1+a}\left(\frac{1-\bar{p}(0)}{\bar{p}(0)}\right)^a\frac{p^a(1+a-p)}{(1-p)^{a+1}}-(R_H-\frac{1}{1+a}R_L).$$

Now, both terms are positive, with the first term strictly increasing in p and the second term being constant. It therefore suffices to show that the derivative is non-

positive when $p = \bar{p}(0)$:

$$\frac{\partial \Delta V^B(\bar{p}(0), r/\lambda)}{\partial p} = \frac{R_H - R_L}{1+a} \frac{(1+a-\bar{p}(0))}{(1-\bar{p}(0))} - (R_H - \frac{1}{1+a} R_L).$$

Plugging in $\bar{p}(0) = R_L/R_H$ in the first term above, we get:

$$\frac{R_H - R_L}{1+a} \frac{1+a-R_L/R_H}{(R_H - R_L)/R_H} = R_H - \frac{1}{1+a} R_L.$$

It follows that $\frac{\partial \Delta V^B(\bar{p}(0), r/\lambda)}{\partial p} = 0$ and, thus, $\Delta V^B(p, r/\lambda)$ is decreasing in this range. Since the normalizing factor, $pR_H + (1-p)R_L$ is increasing in p, it follows that $\Delta \Pi^B(p, r/\lambda)$ is also decreasing in this range.

3. $p \geq \bar{p}(0)$**:** In this case,

$$\Delta V^B(p, r/\lambda) = \Delta \Pi^B(p, r/\lambda) = 0.$$

It follows that both $\Delta V^B(p, r/\lambda)$ and $\Delta \Pi^B(p, r/\lambda)$ are single-peaked, with a unique maximum at $\bar{p}(1)$. ■

We further conjecture that, for any p, both $\Delta V^X(p, r/\lambda)$ and $\Delta \Pi^X(p, r/\lambda)$ are single-peaked in r/λ for $X = G, B$. Yet, as our preceding analysis shows, in certain ranges of p, the derivatives of these functions with respect to $a = r/\lambda$ involve transcendental expressions in a. Identifying their roots analytically is therefore challenging. Further work could investigate these questions more deeply.

2 Non-indexability for Arbitrary Disentanglement Levels

For simplicity, consider the balanced news setting. Similar arguments shows the lack of index in the good news and bad news environments. Suppose, by contradiction, that for some disentanglement level $\alpha \in [0, 1)$, the optimal exploration policy can be described via an index tailored to each project. We denote by $I_\alpha(p, R, \lambda)$ 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 now that

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

and that $\lambda_1 = 1$.

We now pick λ_2 sufficiently small such that, when the agent has access to projects 1 and 2, she optimally exploits project 2, but explores project 1. This is possible to do since this would be the optimal strategy when $\lambda_2 = 0$ and the payoff from every strategy is continuous in the parameters. That is, $I(p_1, R_1, \lambda_1) > I(p_2, R_2, \lambda_2)$.

Further, we pick λ_3 sufficiently high such that when the agent has access to projects 1 and 3, she explores and exploits project 3. Therefore, $I(p_3, R_3, \lambda_3) > I(p_1, R_1, \lambda_1)$.

Finally, the fact that $p_2 R_2 > R_3$ implies that, regardless of λ_2 and λ_3, when the agent has access to projects 2 and 3, she optimally explores and exploits project 2. This is the optimal policy even when $\alpha = 0$, so it must be the optimal policy when $\alpha > 0$. Thus, $I(p_2, R_2, \lambda_2) > I(p_3, R_3, \lambda_3)$, establishing a cycle, in contradiction.

3 Robustness of the Optimal Policy

In what follows, we discuss the robustness of the optimal policies to less extreme disentanglement levels than those analyzed in the main body of the text.

To gain some intuition, consider first the case of a single risky project. It follows from the explicit form of the optimal strategy in Proposition 1 that the optimal strategy for small $\alpha > 0$ is close to the optimal strategy for $\alpha = 0$ in the following sense. Consider the good news setting and an initial belief $p > \bar{p}(0)$ that the risky project is good. Absent news, the optimal strategy when $\alpha = 0$ is to exploit and explore project H for some time t_0, the time it takes for the belief to reach $\bar{p}(0)$, and then switch to exploiting project L and exploring project H. The optimal strategy when $\alpha > 0$ is to exploit and explore project H for a duration of time t_α, the time it takes for the belief to reach $\bar{p}(\alpha)$, with $t_\alpha > t_0$, and then switch to exploiting project L and exploring project H at a rate of $1 - \alpha$. Similar arguments apply for the case in which the initial belief is $p < \bar{p}(0)$, as well as for bad news settings.

We now sketch a similar robustness result for the two risky projects problem. For concreteness, we focus on good news settings. We only need to consider the optimal strategy until news arrives, since from that point onward, the problem reduces to the case of a single risky project.

A strategy is given by $(\sigma_1, \sigma_2) : [0, \infty) \to [0, 1]$ where $\sigma_1(t)$ and $\sigma_2(t)$ are, respectively, the exploitation and exploration resources project H at time t. The space of strategies is equipped with the weak* topology of $L^\infty([0, \infty), \mu)$, where μ is the probability distribution over $[0, \infty)$ given by $\mu = re^{-rt}t$. By the Banach–Alaoglu theorem, this set is compact. A strategy is feasible if $\sigma_2(t) \geq \alpha\sigma_1(t)$, so the feasibility correspondence from α to strategies has a closed graph. Finally, the payoff function is jointly continuous in α and the strategy (taking into account the continuity in α of the continuation payoff after news). Therefore, by the maximum theorem, for sufficiently small α, the optimal strategy in the α-problem is

arbitrarily close to the optimal strategy for $\alpha = 0$. Arbitrarily close in this topology implies that for every T and $\epsilon > 0$, there is an $\bar{\alpha} > 0$ such that for any $\alpha < \bar{\alpha}$, the optimal strategy with disentanglement level α is at a distance of at most ϵ from the optimal strategy for $\alpha = 0$ during horizon T, except possibly on a subset of times of measure at most ϵ.

4 Bound on Switching Time under Good News

Proposition 3 in the main text illustrates that if the agent initially explores a favorable project x, then if, absent news, she switches to exploring the other (initially unfavorable) project y, she does so at a time $T \leq \bar{t}_x(p_L, p_H)$. While we do not have an analytical characterization for T, we now illustrate that, depending on parameters, it can be as close to 0 or as close to $\bar{t}_x(p_L, p_H)$ as we wish.

To see an example in which $T < \bar{t}_x(p_L, p_H)$ for a favorable project x, consider a case where $p_L R_L < \tilde{p}_H R_H < R_L$: it is optimal to exploit project H with prior $\tilde{p}_H$, but learning that project L is good would lead to a switch in exploitation. Suppose the arrival rate λ_L^g is sufficiently high so that the agent strictly prefers exploring project L (and, by the Proposition, continues doing so until news). Now, keeping all other parameters fixed, consider a prior $p_H > \tilde{p}_H$ such that $p_H R_H > R_L > p_L R_L$. If $\lambda_H^b > 0$, then it must be optimal to explore project H at p_H since learning about project L would not induce a change in which project is exploited. The agent will then switch when the posterior declines to at least $\tilde{p}_H$. Since $p_H R_H > \tilde{p}_H R_H > p_L R_L$, the agent will switch strictly before hitting indifference. In particular the agent must switch at some time T strictly below $\bar{t}_H(p_L, p_H)$.

To see an explicit example where T indeed approaches $\bar{t}_x(p_L, p_H)$ for a favorable project x, consider a balanced news setting with news arrival rates λ_L and λ_H such that, when $\hat{p}_H R_H = p_L R_L$, the agent is also indifferent between exploring project H and exploring project L (as determined by the condition given in Proposition 2). From the characterization in Proposition 2, this indifference at $\hat{p}_H$ implies that the agent strictly prefers exploring project H when $p_H > \hat{p}_H$ and exploring project L when $p_H < \hat{p}_H$. Now, modify project H so that it is a good news project that is "almost balanced" by setting $\lambda_z^b = \lambda_z - \epsilon'$ (and $\lambda_z^g = \lambda_z$), for $z = L, H$ and $\epsilon' > 0$. For any $\epsilon > 0$, pick $\delta > 0$. We can find $\epsilon' > 0$ sufficiently small such that the following hold. First, the agent optimally starts exploring project H when the prior that project H is good is $p_H = \hat{p}_H + \delta$. Second, the agent stops exploring project H at a prior sufficiently close to $\hat{p}_H$ that the time it takes to switch is at least $\bar{t}_H(p_L, p_H) - \epsilon$.[1]

[1] Notice that when the posterior that project H is good reaches $\hat{p}_H$, the agent must be exploring project L. Indeed, if the agent continues exploring project H beyond $\hat{p}_H$, her posterior that project H is good would drop so that project H becomes unfavorable, implying that the agent would never switch, in contradiction to Proposition 3.

5 Maximal Under-Exploitation by Teams

Here, we provide the derivations underlying the size of the maximal under-exploitation region discussed in the main text.

Define

$$\Delta(\alpha, n) = \underline{p}_{\alpha}(n) - p_{\alpha}^{SW}(n) = \lambda(n-1)R_L(R_H - R_L)\frac{\alpha\Big[r + n\lambda(1-\alpha)\Big]}{D_1(\alpha)D_2(\alpha)}, \qquad n > 1,$$

where The two positive linear denominators are

$$D_1(\alpha) = (r + n\lambda)R_H - n\lambda\alpha R_L,$$
$$D_2(\alpha) = \Big[r + \lambda\Big(n - (n-1)\alpha\Big)\Big]R_H - \lambda\alpha R_L.$$

Substitute the individual derivatives to obtain

$$\frac{\partial\Delta}{\partial\alpha} = \frac{\lambda(n-1)R_L(R_H - R_L)(n\lambda + r)}{\Big[D_1(\alpha)D_2(\alpha)\Big]}\left[\frac{r + n\lambda(1-2\alpha)}{\alpha\,[\,r + n\lambda(1-\alpha)\,]} - \left(\frac{n\lambda R_L}{D_1(\alpha)} + \frac{\lambda[(n-1)R_H + R_L]}{D_2(\alpha)}\right)\right].$$

Define the term in square brackets in the equation above as

$$\Psi(\alpha) = \frac{r + n\lambda\Big(1 - 2\alpha\Big)}{\alpha\,[\,r + n\lambda(1-\alpha)\,]} - \left[\frac{n\lambda R_L}{D_1(\alpha)} + \frac{\lambda\Big[(n-1)R_H + R_L\Big]}{D_2(\alpha)}\right]$$

Since $D_1(\alpha), D_2(\alpha) > 0$, the sign of $\frac{\partial\Delta}{\partial\alpha}$ is the same as the sign of $\Psi(\alpha)$. The bracket on the right is strictly positive and increasing in α, while the first fraction starts positive is decreasing in α and is dominant for small values of α: it becomes arbitrarily large as $\alpha \to 0$. Hence $\Psi(\alpha)$ is positive for small values of α and can change signs at most once on $(0, 1]$. We now evaluate $\Psi(\alpha)$ at $\alpha = 1$ to obtain

$$\Psi(1) = \frac{r - n\lambda}{r} + \frac{n\lambda R_L}{(r + n\lambda)R_H - n\lambda R_L} + \frac{\lambda\Big[(n-1)R_H + R_L\Big]}{(r + \lambda)R_H - \lambda R_L},$$

so that $\Psi(1) = 0$ when r equals $\hat{r} := \frac{\lambda\Big(R_H - R_L\Big)\sqrt{n}}{R_H}$. This implies that $\Psi(1)$ is negative whenever $r < \hat{r}$ and non-negative when $r \geq \hat{r}$. This allows us to conclude that, for $r \geq \hat{r}$, $\Delta(\alpha)$ is increasing for all $\alpha \in (0, 1]$, whereas, for $r < \hat{r}$, $\Delta(\alpha)$ increases until a single peak α^* and then decreases.

6 Introducing Agency Frictions

We now illustrate an additional application of our framework. Specifically, we examine a particular form of agency problem in the context of exploration. To illustrate the core

ideas, we focus on a setting with a single risky project in a pure good-news setting: one project yields a known reward, whereas the other produces payoffs at rate λ only if successful. This discussion aims to highlight the richness of our basic framework.[2]

Recall that, with full disentanglement ($\alpha = 0$), in the benchmark single-agent environment with one risky project, the optimal strategy is to continuously explore the risky project while exploiting whichever project yields the higher expected reward at any point in time.

We consider a setting with two agents—a *Doer* and an *Observer*—who make decisions independently. The Doer has the authority to act but lacks the time, ability, or expertise to monitor the outcomes of the risky project. The Observer, by contrast, possesses the time, ability, or expertise to gather information but cannot act on it directly. This setup reflects a range of real-world scenarios, including R&D and management teams in technology or pharmaceutical firms, hiring committees and department chairs in academia, and advisory bodies and decision-makers in politics.[3]

For simplicity, we assume that the Doer has access to the information collected by the Observer, including her observations, and that the Observer exerts effort at zero cost.

The agents share a common prior p about the probability that the risky project is successful. However, their payoffs differ: the Doer values the safe and risky projects at R_L^D and R_H^D, respectively, while the Observer values them at R_L^O and R_H^O. We assume that, at any time, both agents' strategies are measurable with respect to the information available by that time. We focus on Markov Perfect Equilibria (MPE). When the Doer and Observer share identical preferences, the setting reduces to the single-agent problem analyzed in the paper.

We begin with examples illustrating outcomes in the absence of commitment, then highlight the value of commitment. In particular, we show how the Observer can tailor the exploration strategy to manipulate the Doer's behavior.

Supppose that the initial disagreement is due to the Observer always preferring the risky project: $R_L^O = 0$. If $pR_H^D > R_L$, then there is an equilibrium in which the Observer explores only the safe project and the Doer exploits the risky project indefinitely. When $pR_H^D < R_L^D < R_H^D$, however, the only way for the Observer to convince the Doer to exploit the risky project is to prove it is successful. Therefore, in any MPE, the Observer explores the risky project till a success.

Assume that the Observer prefers the safe project, even if the risky project is successful:

[2]The assumption of good news and one safe project are made for expositional simplicity. A fuller treatment, including more general information structures or two risky projects would certainly be interesting as well, but would require consideration of many additional cases.

[3]For related environments in the context of sequential sampling, see Chan, Lizzeri, Suen, and Yariv (2018) and Henry and Ottaviani (2019).

$R_L^O \geq R_H^O$. In contrast, at the outset, the Doer's expected flow payoffs are higher for the risky project: $p_H R_H^D > R_L^D$.

If the Observer never explores the risky project, the Doer will optimally exploit it indefinitely. In equilibrium, the Observer explores the risky project to lower the Doer's posterior enough to induce a switch to the safe project, even in the absence of news. Let $\bar{t}(p_H)$ be the time required for the posterior to fall to $p_M = R_L^D/R_H^D$, the Doer's myopic threshold.

Proposition OA_3 (Initial Disagreement: Observer Prefers Safe) *In equilibrium, the Observer explores the risky project until the posterior reaches the Doer's myopic cutoff. The exploration time, $\bar{t}(p_H)$, is increasing in p_H and R_H^D, and decreasing in R_L^D.*

Intuitively, the Observer needs to lower the posterior belief to the point at which the Doer becomes indifferent between exploiting the risky and the safe project indefinitely. At that point, the Observer switches to exploring the safe project: exploring the risky project bears risk of generating good news that would lead the Doer to switch back to exploiting it. As a result, the time $\bar{t}(p_H)$ the Observer spends inspecting the risky project, absent good news, corresponds to the duration that the risky project is exploited in equilibrium. This time $\bar{t}(p_H)$ depends solely on the Doer's payoffs and independent of the Observer's payoffs. The comparative statics follow directly. The outcome is inefficient from the Doer's perspective: exploration ceases even though uncertainty about the risky project remains. The Doer would prefer continued exploration of the risky project even after time $\bar{t}(p_H)$.

Consider now the case in which both the Observer and the Doer prefer the risky project ex ante: $pR_H^D > R_L^D$ and $pR_H^O > R_L^O$. However, suppose the Observer's relative benefit from the risky project is much larger than the Doer's: $pR_H^O - R_L^O > pR_H^D - R_L^D$. For simplicity, assume $R_L^O = 0$ so that the Observer prefers the risky project regardless of its success.

Now, the Observer has an incentive to prevent the Doer from learning, which could shift her exploitation to the safe project.

Proposition OA_4 (Initial Agreement) *There exists an equilibrium in which the Observer never explores the risky project. Furthermore, in any equilibrium, the Doer exploits the risky project throughout.*

Certainly, the Observer could also inspect the risky project for a sufficiently short time so that the Doer does not change which project she exploits, which gives rise to equilibrium multiplicity. For all such profiles, the resulting behavior is the same, with the risky project being exploited.

When $R_L^O > 0$, a similar result holds, but under more limited circumstances. In particular, the Observer needs to be sufficiently impatient or learning needs to be sufficiently slow.

References

Chan, J., A. Lizzeri, W. Suen, and L. Yariv (2018). Deliberating collective decisions. *The Review of Economic Studies 85*(2), 929–963.

Henry, E. and M. Ottaviani (2019). Research and the approval process: The organization of persuasion. *American Economic Review 109*(3), 911–955.