Efficient Resource Allocation in Discrete-time Breakthrough Bandit Models*

Audrey Hu†
City University of Hong Kong

Liang Zou‡
University of Amsterdam

January 27, 2025

*The coauthors would like to express their gratitude to David Besanko, Yeon-Koo Che, Claudio Mezzetti, Carlos Oyarzun, Eilon Solan and Zaifu Yang for their invaluable comments, as well as to the seminar participants at the National University of Singapore, Nanyang Technological University, the University of Queensland, the University of Technology Sydney, and the University of York. We are also thankful to the participants of ESEM 2024, EARIE 2024, and the HK Theorists Workshop 2023 for their valuable feedback. Special thanks to Yating Yuan for outstanding research assistance. All remaining errors are our own.

†*Corresponding author.* Address: 9-256, Lau Ming Wai Academic Building, City University of Hong Kong, Hong Kong SAR. Telephone number: +852 34426767. Email: audrey.hu@cityu.edu.hk.

‡Email: zou.uva@gmail.com

Efficient Resource Allocation in Discrete-Time Breakthrough Bandit Models

by Audrey Hu and Liang Zou

Abstract. We study a two-armed bandit model in discrete time, where a team of agents allocates limited resources (e.g., time) per period to achieve a breakthrough under uncertainty. In the cooperative case, we characterize (Pareto) efficient allocation strategies with interior solutions, improving upon the binary strategies prevalent in existing literature. These strategies exhibit two notable features: (i) persistence, where experimentation continues until a breakthrough is achieved, and (ii) adherence to a "Goldilocks principle," whereby each agent's incentives to experiment are maximized at specific team sizes or task difficulties. In the noncooperative case, we identify a symmetric equilibrium of strategic experimentation, characterized by free-riding, inefficiency, and procrastination, with no other equilibria existing. We further demonstrate that Pareto-efficient outcomes can be restored in equilibrium using a dynamic profit-sharing contract that satisfies budget-balancing and limited-liability constraints.

Keywords: Pareto efficiency, Resource allocation, Strategic experimentation, Dynamic profit-sharing contract, Discrete time, Exponential bandit, Goldilocks principle.

JEL Classification: D83, C73, D86, C02,O32

1 Introduction

Optimizing resource allocation under uncertainty is a fundamental challenge across diverse real-world contexts. This paper examines a discrete-time, two-armed bandit model that encapsulates the core economic trade-off between exploiting a "safe" arm with a guaranteed return and exploring a "risky" arm with uncertain but potentially superior rewards. The risky arm is either "good," offering a present value greater than the safe arm, or "bad," devoid of any value. A team of agents, each endowed with a single unit of divisible resources (e.g., time) per period, must decide how to allocate their resources between the two arms. Success is achieved when a "breakthrough" confirms the risky arm is good. If the arm is bad, all exploration time is ultimately wasted.

This "breakthrough bandit" framework serves as a tractable model for numerous real-world problems, such as R&D investment, pharmaceutical trials, mineral exploration, and the search for evidence to support an unproven conjecture. While prior research has extensively analyzed bandit problems, economic applications often restrict attention to binary strategies: allocating all available resources to either the safe or risky arm, combined with an optimal stopping rule in the absence of success (see, e.g., Bergemann and Välimäki (2008)). Consistent with the Gittins index theorem (Gittins and Jones (1974); Gittins (1979)), such strategies simplify analysis while omitting the tradeoffs between marginal benefit and cost—a typical feature characterizing interior solutions.[1] Indeed, many real-world problems naturally involve continuous allocation variables, such as determining annual R&D budgets, balancing research time across projects, or setting experimental prices for new products. In such settings with periodic decisions and concave objective functions, efficient allocation strategies are more likely to involve interior solutions (Rothschild (1974)).

This study addresses the gap in the literature by analyzing a breakthrough bandit model with continuous, interior solutions, assuming the probability of success is an exponential function of the time allocated to the good risky arm—a standard but non-essential assumption that simplifies characterizing the optimal policy. Our key methodological contribution is a transformation approach that incorporates both posterior beliefs and allocation strategies as state variables, allowing for a transparent and tractable characterization of the optimal policy. A pivotal lemma (Lemma 1)

[1] See, e.g., Malueg and Tsutsui (1997) and Bonatti and Hörner (2017) for analyses of interior solutions in continuous time, and Aghion et al. (1991) in discrete time.

simplifies the first-order conditions, providing a useful tool for understanding the optimal trade-off between marginal benefit and cost associated with bandit dynamics.

Under cooperative decision-making, we find that efficient strategies feature continuous allocations in the range (0,1), and experimentation never ceases unless a breakthrough is achieved. This persistence arises from a subtle interaction between Bayesian belief updates and first-order optimality conditions. While continuous exploration does not guarantee discovering the good arm, it ensures that experimentation only diminishes asymptotically over time in the absence of success. Moreover, the optimal allocation follows a "Goldilocks principle," where incentives to experiment peak when the task is neither too easy nor too difficult.

In contrast, when agents act noncooperatively, strategic experimentation under hidden actions leads to a symmetric equilibrium, which is unique among all Nash equilibria. Unlike continuous-time models, which often allow multiple equilibria (e.g., Keller et al. (2005); Bonatti and Hörner (2011); Hörner et al. (2022)), our discrete-time framework ensures uniqueness due to the concavity of payoff functions. However, inefficiencies such as free-riding and procrastination persist, highlighting the need for mechanisms to align individual incentives with efficient outcomes.

To address inefficiencies arising from strategic experimentation, we design an outcome-contingent, dynamic profit-sharing contract that restores Pareto efficiency in equilibrium. In our public-good environment, it is possible to achieve the first-best solution under both budget-balancing and limited-liability constraints. Although players' actions are hidden information and the outcomes depend largely on luck, we show that, at least in theory, there exist fine-tuned profit-sharing rules based on the information available in each period that perfectly align players' incentives with the efficient policies. Additionally, our optimal contract avoids extreme rules like "winner-takes-all," instead preserving the public-good nature of experimentation and ensuring all agents benefit to some extent from a breakthrough.

The remainder of the paper is structured as follows. Section 2 reviews related literature. Section 3 introduces the cooperative model and derives the Pareto-efficient solution. Section 4 extends the analysis to strategic experimentation with hidden actions, characterizing equilibrium behavior. Section 5 presents the dynamic profit-sharing rule that restores efficiency. Section 6 concludes, and the Appendix contains technical or lengthier proofs.

2 Related Literature

This paper builds on the extensive literature on breakthrough bandit models, particularly the continuous-time framework introduced by Keller et al. (2005). Their seminal work assumes conclusive breakthroughs with arrival times following an exponential distribution. In addition to focusing on the discrete-time counterpart of Keller et al. (2005), this study also introduces a two-dimensional transformation approach that offers a tractable solution method for exponential bandit problems, addressing continuous allocation strategies.

The literature on discrete-time bandit problems is vast but has primarily focused on finite control sets.[2] Among the exceptions, Chikte (1980) examined continuous allocation strategies in discrete-time bandit problems but relied on the assumption that the conditional probability mass function is log-supermodular, which implies convexity of the objective function and corner solutions. In contrast, our exponential bandit model exhibits log-submodularity, which gives rise to interior solutions, distinguishing it from Chikte's framework. Aghion et al. (1991) advanced the analysis of experimentation by considering continuous decisions in a general model, but their focus is mainly on the asymptotic properties of beliefs and actions as time tends to infinity. While Aghion et al. (1991) only made a limited attempt at characterizing optimal experimentation strategies, the present study fully characterizes the optimal strategies for the exponential bandit models.

In the context of discrete-time breakthrough bandits, prior research has predominantly explored binary actions and optimal stopping times. Bergemann and Hege (1998, 2005) analyzed innovation financing with linear breakthrough probabilities, deriving binary strategies, and focused on stopping decisions. Rosenberg et al. (2007) extended Keller et al. (2005)'s continuous-time framework to discrete time with unobservable outcomes, showing that equilibria involve time-varying cutoff strategies under binary-action assumptions. Heidhues et al. (2015) examined payoff observability and cheap-talk communication in discrete-time bandits with binary actions, identifying conditions for socially optimal equilibria. Halac et al. (2016) analyzed a principal-agent contracting problem in an exponential bandit setting with binary actions. To the best of our knowledge, this study is the first that provides an interior

[2] For example, Berry and Fristedt (1985)'s book, which includes over 200 annotated references, defines a strategy as a mapping that assigns an integer to each (partial) history of observations, indicating which arm to select at the next stage.

solution for continuous, efficient allocation strategies in discrete-time breakthrough bandit models.

The study is also related to the growing literature on strategic experimentation, which has been extensively studied in the continuous-time framework following the seminal work of Bolton and Harris (1999). For the exponential bandit, Keller et al. (2005) derived the influential "bang-bang" strategy for cooperative solutions, where resources are fully allocated to the risky arm above a belief threshold and withdrawn below it. This strategy has been widely applied in theoretical and practical contexts (e.g., Awaya and Krishna (2021); Thomas (2021); Besanko and Wu (2013); Murto and Välimäki (2011)). Among the earlier works, Choi (1997) explored R&D races under hazard rate uncertainty, while Malueg and Tsutsui (1997) characterized interior solutions by introducing quadratic cost functions. Extensions to inconclusive breakthroughs (Keller and Rady (2010)), career concerns with continuous decision variables (Bonatti and Hörner (2017)), and principal-agent problems (Guo (2016); Halac et al. (2017)) further highlight the adaptability of continuous-time models. However, discrete-time and continuous-time exponential bandits exhibit fundamental differences. For one thing, in discrete time, breakthrough probabilities are exponential and concave in allocations, reflecting diminishing returns to exploration. In continuous time, probabilities are linear in allocations due to their Poisson structure, leading to corner solutions for risk-neutral agents with linear cost functions. In these situations, interior (or mixed) allocations in continuous time arise mainly in noncooperative equilibria, where players are indifferent between arms. While Keller et al. (2005), Bonatti and Hörner (2011), and Hörner et al. (2022) document the multiplicity of equilibria in continuous-time frameworks, we demonstrate that discrete time ensures equilibrium uniqueness.

In the context of contracting under bandit dynamics, our study is related to, among others, the above-cited studies by Bergemann and Hege (1998, 2005), Guo (2016); and Halac et al. (2016) on principal-agent problems, Bonatti and Hörner (2011) on free riding, Halac et al. (2017) on contest design, and Bonatti and Hörner (2017) on career concerns. Our dynamic profit-sharing contract complements Bonatti and Hörner (2011)'s self-enforcing deadline contracts by fully restoring Pareto efficiency. Moreover, our optimal contracting result complements Halac et al. (2017)'s findings regarding the optimality of "public winner-takes-all" or "hidden equal-sharing" contests under certain scenarios. Notably, our contract preserves the public-good nature of experimentation, allowing all agents to share the benefits of breakthroughs.

Finally, this study relates to broader applications of exponential bandits in economics. Sadler (2021) examined optimal tax and subsidy policies to enhance research spillovers in sequential exponential bandits, while Hörner and Skrzypacz (2017) emphasized the versatility of bandit models for studying innovation and strategic experimentation. By addressing interior solutions and introducing dynamic profit-sharing mechanisms, this study deepens our understanding of discrete-time settings and offers practical insights for real-world applications.

3 The Basic Model

Time is discrete, with countable periods $t \in \mathbb{N} = \{1, 2, ...\}$. There are $n \geq 1$ agents, each endowed with one unit of a perfectly divisible resource (referred to as time) per period. Each agent faces an identical two-armed bandit problem. One arm is "safe," and the other is "risky." The safe arm provides a known (expected) return of $\ell > 0$ per period. All agents have the same discount factor $\delta \in (0, 1)$, so if an agent allocates full time to the safe arm indefinitely, he enjoys the present value $L = \ell/(1 - \delta)$ from the safe arm. The risky arm has an unknown type $\omega \in \{0, 1\}$. It is "good" for all agents when $\omega = 1$ or "bad" for all agents when $\omega = 0$. If it is bad, it yields nothing; if it is good, it is worth full-time exploitation, with an expected present value $G = \gamma L \in (L, \infty)$ to each agent, where $\gamma > 1$ measures the relative attractiveness of the good risky arm compared to the safe arm.

A "breakthrough" occurs when any of the agents obtains conclusive evidence that the risky arm is good. Before this happens, the agents have independent probabilities of a breakthrough in any period t. Given that the risky arm is good, when an agent allocates a fraction $a_t \in [0, 1]$ of his time to the risky arm in period t, his breakthrough probability follows an exponential distribution: $1 - e^{-\lambda a_t}$, where $\lambda \in (0, \infty)$ represents how easy it is to achieve a breakthrough. Following standard approaches in bandit problems, we assume memoryless experiments, where the probability of success in each period depends solely on the current allocation.

In this basic model, we assume the agents work cooperatively as a team, jointly choosing a feasible allocation plan that maximizes the sum of their expected payoffs. For the exponential bandit problem under investigation, the probability of a breakthrough depends on the total sum of agents' times allocated to the risky arm and not on how they share it. Thus, without loss of efficiency, we assume that agents share the total allocation of time equally so that the team's problem can be analyzed as a

single (representative) agent's problem, whose decisions are mimicked by all others. With $a_t \in [0,1]$ of his time allocated to the risky arm, the representative agent's conditional probability of a breakthrough now equals $F(a_t) = 1 - e^{-\lambda n a_t}$ given the arm is good.

A feasible allocation plan for the team can be now described as a sequence of contingent allocations (or actions) by each agent: $\alpha = (\alpha_t)_{t=1}^{\infty}$, such that $\alpha_t \in [0,1]$ is measurable with respect to the information available at the start of period t, and the plan stops once a breakthrough occurs.

The agents start with a prior common belief that the risky arm has a probability $\pi_0 \in [0,1]$ to be good. According to Bayes' rule, if the representative agent (henceforth, agent) allocates $\alpha_1 ..., \alpha_{t-1}$ in periods 1 to $t-1$ without a breakthrough, the posterior belief by the end of period t is updated as follows:

$$\pi_t = \begin{cases} 1^+ & \text{if a breakthrough occurs in period } t \\ \frac{\pi_{t-1}(1-F(\alpha_t))}{1-\pi_{t-1}F(\alpha_t)} & \text{if no breakthrough in } t \end{cases} \tag{1}$$

where 1^+ means that if a breakthrough occurs, $\pi_t = 1$, together with the fact that the risky arm is verified (rather than believed) to be good. The sequence $(\pi_t)_{t=0}^{\infty}$ forms a martingale, meaning that the conditional expectation $E(\pi_t|\pi_{t-1})$ equals π_{t-1} for all $\alpha_t \in [0,1]$ and $t \in \mathbb{N}$.

If a breakthrough arrives in period t, we assume that the agent enjoys the present value G starting from period $t+1$. Thus, the agent's conditional expected payoff in each period t, given that no breakthrough has occurred yet, equals $(1-\alpha_t)\ell + \delta\pi_{t-1}F(\alpha_t)G$. Given the allocation plan α and the belief updating rule (1), the agent's discounted expected payoff thus equals

$$\sum_{t=1}^{\infty} \delta^{t-1} \prod_{s=1}^{t-1}(1 - \pi_{s-1}F(\alpha_s)) \left[(1-\alpha_t)\ell + \delta\pi_{t-1}F(\alpha_t)G\right] \tag{2}$$

where $\prod_{s=1}^{0}(\cdot) := 1$, and $\prod_{s=1}^{t-1}(1 - \pi_{s-1}F(\alpha_s))$ is the probability of no breakthrough until $t-1$. The representative agent's objective is to choose a feasible allocation plan α to maximize the expected payoff in (2) subject to (1), given the prior π_0, with the understanding that the plan will stop as soon as a breakthrough arrives. By standard arguments (e.g., Karlin (1955)), an optimal solution exists for the agent's problem.

Using dynamic programming, if α attains optimality, then the agent's optimal expected payoff satisfies the functional equation, or Bellman equation, for all $t \in \mathbb{N}$:

$$v(\alpha_t, \pi_{t-1}) = \max_{a \in [0,1]} v(a, \pi_{t-1}) \text{ s.t. } (1) \tag{3}$$

where $v : [0,1]^2 \to [L,G]$ represents the agent's conditional expected payoff given no breakthrough has occurred:

$$v(a,\pi_{t-1}) = (1-a)\ell + \delta\pi_{t-1}F(a)G + \delta(1-\pi_{t-1}F(a))v(\alpha_{t+1},\pi_t) \tag{4}$$

The interpretation is straightforward: suppose there has been no breakthrough so far and the agent's past allocations imply a posterior belief π_{t-1}. If the agent allocates a to the risky arm in period t, he receives an immediate return $(1-a)\ell$ from the safe arm. There is a probability $\pi_{t-1}F(a)$ that a breakthrough will occur by the end of period t, rewarding the agent with G. If no breakthrough occurs, the agent updates his belief and continues to the next period. By the principle of optimality for dynamic programming (e.g., Stokey et al. (1989), Chapter 4), since payoffs are bounded, the sequence α maximizes (2) if and only if α_t maximizes $v(a,\pi_{t-1})$ for all $t \in \mathbb{N}$, subject to (1).

The standard approach to solving (3)-(4) often restricts attention to solutions and policies that are time-invariant functions of the posterior beliefs. Suppose there exists a value function $v^* : [0,1] \to [L,G]$ as the solution to (3)-(4), along with a pure stationary Markov policy function $\varphi : [0,1] \to [0,1]$ that achieves this solution. Then, given any prior π_0, the sequence $\alpha = (\alpha_t)_{t=1}^{\infty}$, defined by $\alpha_t = \varphi(\pi_{t-1})$, represents the optimal allocation plan, attaining the value $v^*(\pi_{t-1}) = v(\varphi(\pi_{t-1}),\pi_{t-1})$.

While this restriction is theoretically sound (Blackwell (1965), Puterman (1994)) and works effectively in specific cases (e.g., Bergemann and Hege (1998, 2005)), it presents significant tractability challenges in more general scenarios. For instance, when allocation strategies involve interior solutions characterized by first-order conditions, the optimal policy often can only be defined implicitly. Furthermore, with belief-updating, the conditions that define the optimal policy function typically depend on and interact with subsequent belief updates in a manner that is analytically intricate and difficult to disentangle. Of course, this restriction is unnecessary. Given the initial state of belief and updating rules, any complete prescription that specifies actions in each period t, accounting for all contingencies, qualifies as a valid strategy (Karlin (1955)). To overcome the potential difficulties in characterizing the optimal policy as a single-dimensional state function, we therefore adopt a two-dimensional transformation approach to solve the program in (3)-(4). Specifically, we define both the allocation α_t and the posterior belief π_{t-1} as state variables, combining the allocation space and the posterior space into a product state space $[0,1]^2$. Our objective is then to derive a time-invariant, continuous and monotone transformation from $[0,1]^2$

to itself, mapping each paired state (α_t, π_{t-1}) to its adjacent paired state.

This transformation, together with a given prior π_0 and a terminal or transversality condition on $\lim_{t\to\infty}(\alpha_t, \pi_{t-1})$, enables us to derive the optimal allocation-belief pair in the first period, (α_1, π_0), using backward induction (Theorem 1). The resulting optimal policy plan is then uniquely characterized by the law of motion implied by the transformation (Theorem 2). Furthermore, denoting $v_t = v(\alpha_t, \pi_{t-1})$, the process $\{(\alpha_t, \pi_{t-1})\}_{t=1}^{\infty}$ uniquely determines the value process $(v_t)_{t=1}^{\infty}$ through the recursive relation:

$$v_t = \delta G + (1-\alpha_t)\ell + \delta(1-\pi_{t-1}F(\alpha_t))\,(v_{t+1} - G) \tag{5}$$

Definition 1 *The process $\{(\alpha_t, \pi_{t-1})\}_{t=1}^{\infty}$ is called an **optimal policy plan** if it satisfies (1) and $v(\alpha_t, \pi_{t-1}) = \max_{a\in[0,1]} v(a, \pi_{t-1})$ for all $t \in \mathbb{N}$. The plan stops from period t onward if $\pi_{t-1} = 1^+$.*

As will be seen, the proposed transformation approach significantly simplifies the analysis, providing a dynamic framework that balances exploration and exploitation in optimal allocation strategies. An additional advantage of this approach is its flexibility, as it naturally accommodates the possibility of an exogenous deadline for experimentation—a practical feature not afforded by pure stationary Markov policies or Gittins indexing policies. While we do not address exogenous deadlines separately, this possibility is naturally embedded as straightforward corollaries in our main theorems.

Before presenting Theorem 1, we need to establish some preparatory results first. By standard arguments, $v(a, \pi)$ is continuously differentiable in both arguments (given F exponential). Therefore, for all $a = \alpha_t$ at which (3) has an interior solution, the first-order condition must hold:

$$\begin{aligned} \frac{\partial}{\partial a} v(\alpha_t, \pi_{t-1}) &= -\ell + \delta\pi_{t-1}F'(\alpha_t)\,[G - v(\alpha_{t+1}, \pi_t)] && (6)\\ &\quad +\delta(1-\pi_{t-1}F(\alpha_t))\frac{\partial v(\alpha_{t+1}, \pi_t)}{\partial \pi_t}\frac{\partial \pi_t}{\partial a} && (7)\\ &= 0, \end{aligned}$$

or else $\alpha_t = 0$ or 1, depending on whether $v(0, \pi_{t-1})$ or $v(1, \pi_{t-1})$ is the greatest over the entire interval $[0, 1]$.

The term in (7) captures the learning effect inherent in the bandit problem. This effect is negative because, by the envelope theorem,

$$\frac{d}{d\pi_t} v(\alpha_{t+1}, \pi_t) = \frac{\partial}{\partial \pi_t} v(\alpha_{t+1}, \pi_t) > 0$$

(holding any allocation fixed, increasing posterior π_t increases the probability of a breakthrough), and $\partial\pi_t/\partial a < 0$, for all $\pi_{t-1} \neq 0, 1$. Myopic solutions, on the other hand, ignore this learning effect and equate the marginal cost ℓ with the current-period marginal benefit $\delta\pi_{t-1}F'(\alpha_t)\left[G - v(\alpha_{t+1}, \pi_t)\right]$. As a result, optimal allocations that account for learning effects are generally lower than the myopic allocations.

We now derive a simpler expression for the partial derivative in (6)-(7), which will prove highly useful. The lemma below holds for a general probability function F. So, we continue to use the more general functional form F for ease of applications.

Lemma 1 *Suppose* $\{(\alpha_t, \pi_{t-1})\}_{t=1}^{\infty}$ *is an optimal policy plan. Then,* $\forall t \in \mathbb{N}$,

$$\frac{\partial}{\partial a} v(a, \pi_{t-1}) = -\ell + \delta\pi_{t-1}F'(a)H_{t+1} \tag{8}$$

where $H_{t+1} \in (0, G - L]$ *is defined recursively by*

$$H_{t+1} = (\gamma - 1 + \alpha_{t+1})\ell + \delta(1 - F(\alpha_{t+1}))H_{t+2} \tag{9}$$

Proof. See Appendix. ■

The proof of this lemma demonstrates that tracking future posteriors is unnecessary. In comparison to (6)-(7), we observe from (8) that the function H_{t+1} captures the overall marginal benefit, including learning effects, of investing time in the risky arm. The simplified form of $\partial v/\partial a$ arises because H_{t+1} does not depend directly on the sequence of posterior beliefs $\{\pi_t, \pi_{t+1}, ...\}$, unlike $v(\alpha_{t+1}, \pi_t)$.

3.1 Basic properties of the optimal plan

From now on in the basic model, we invoke the assumption $F(a) = 1 - e^{-\lambda n a}$. If exploring the risky arm is deemed unprofitable, the agent will choose the safety arm indefinitely so that $v \equiv L$. To avoid this trivial case, we assume

Assumption 1 $\delta\lambda(\gamma - 1) > 1 - \delta$.

Clearly, this assumption fails to hold if γ is close to 1 (the potential reward is too low), if λ is sufficiently low (too difficult to make a breakthrough), or if δ is sufficiently low (too costly to wait for an outcome). Assumption 1 allows us to define a minimum posterior belief:

$$\pi_{\min} = \frac{1 - \delta}{\delta\lambda n(\gamma - 1)} \in (0, 1) \tag{10}$$

Note that:

$$\frac{\partial}{\partial a_t} v(a_t, \pi_{\min})|_{a_t=0} = -\ell + \delta \pi_{\min} \lambda n (G - L) = 0,$$

which implies that $\alpha_t > 0$ if and only if $\pi_{t-1} > \pi_{\min}$.

We partition the bandit environment into two complementary scenarios:

$$\text{Scenario I} \quad : \quad \delta e^{-\lambda n} (1 + \lambda n \gamma) < 1 \tag{11}$$

$$\text{Scenario II} \quad : \quad \delta e^{-\lambda n} (1 + \lambda n \gamma) \geq 1 \tag{12}$$

For simplicity of notation, we will often denote λn by θ when there is no need to distinguish between λ and n individually.

We now introduce a "no stopping" result that sharply contrasts with existing literature on the cooperative exponential bandit problems under binary strategies.

Proposition 1 *Suppose $\pi_0 \in (\pi_{\min}, 1]$, and $\{(\alpha_t, \pi_{t-1})\}_{t=1}^{\infty}$ is an optimal policy plan attaining $(v_t)_{t=1}^{\infty}$, a solution to (3)-(3).*

(i) If $\pi_0 = 1$, then $\alpha_t \equiv \hat{\alpha} \in (0, 1]$ and $v_t \equiv v^*$, given by

$$v^* = G - \frac{(\gamma - 1 + \hat{\alpha})\, \ell}{1 - \delta e^{-\lambda n \hat{\alpha}}},$$

In Scenario I, $\hat{\alpha} \in (0, 1)$ is the unique solution of

$$\delta e^{-\lambda n \hat{\alpha}} (\lambda n (\gamma - 1 + \hat{\alpha}) + 1) - 1 = 0 \tag{13}$$

In Scenario II, $\hat{\alpha} = 1$.

(ii) If $\pi_0 \in (\pi_{\min}, 1)$, then $\alpha_t \in (0, 1]$ and $v_t > L$ for all $t \in \mathbb{N}$.

Consequently, experimentation with the risky arm never stops without a breakthrough.

Proof. (i) Suppose $\pi_0 = 1$. In this case, the Bellman equations in (3)-(4 involve no learning effect, as the belief that the risky arm is good is already certain. This allows for a direct derivation of the solution. Substituting $\pi_{t-1} = 1$ and $\alpha_t = \hat{\alpha}$ for all t, the equations reduce to the following (recall $\theta = \lambda n$):

$$\begin{aligned} v^* &= (1 - \hat{\alpha})\ell + \delta \left(1 - e^{-\theta \hat{\alpha}}\right) G + \delta e^{-\theta \hat{\alpha}} v^* \\ &= \frac{(1 - \hat{\alpha})\ell + \delta \left(1 - e^{-\theta \hat{\alpha}}\right) G}{1 - \delta e^{-\theta \hat{\alpha}}} \\ &= G - \frac{(\gamma - 1 + \hat{\alpha})\, \ell}{1 - \delta e^{-\theta \hat{\alpha}}} \end{aligned}$$

We now differentiate $v(a_t, 1)$ in (4) with respect to a_t, holding future actions fixed at $\hat{\alpha}$:

$$\frac{\partial}{\partial a_t} v(a_t, 1) = -\ell + \delta\theta e^{-\theta a_t} (G - v^*)$$

Noting that $v(a_t, 1)$ is a concave function of a_t, the optimal $\hat{\alpha}$ can then be determined by solving the first-order condition $\frac{\partial}{\partial a} v(\hat{\alpha}, 1) = 0$, or by verifying boundary conditions when $\hat{\alpha}$ reaches its limit $\hat{\alpha} = 0$ or $\hat{\alpha} = 1$.

We can now substitute v^* into the equation and rearrange the terms to obtain:

$$\begin{aligned} \frac{\partial}{\partial a} v(\hat{\alpha}, 1) &= -\ell + \delta\theta e^{-\theta\hat{\alpha}} (G - v^*) \\ &= \left[\delta e^{-\theta\hat{\alpha}} (\theta(\gamma - 1 + \hat{\alpha}) + 1) - 1\right] \frac{\ell}{1 - \delta e^{-\theta\hat{\alpha}}} \end{aligned}$$

Define the term in square brackets by $M(\hat{\alpha})$. This is a continuous and strictly decreasing function:

$$M'(\hat{\alpha}) = -\theta^2 \delta e^{-\hat{\alpha}\theta} (\gamma - 1 + \hat{\alpha}) < 0$$

The condition $\pi_{\min} < 1$ implies $M(0) > 0$, so $\hat{\alpha} > 0$. Under Scenario I, we have $M(1) < 0$, so $M(\hat{\alpha}) = 0$ defines a unique solution $\hat{\alpha} \in (0, 1)$. Under Scenario II, we have $M(1) \geq 0$ so that $\hat{\alpha} = 1$.

(ii) Consider next that $\pi_0 \in (\pi_{\min}, 1)$. It suffices to show that $\pi_{t-1} > \pi_{\min}$ implies $\pi_t > \pi_{\min}$ for all $t \in \mathbb{N}$. We prove this by contradiction. Pick any t such that $\pi_{t-1} > \pi_{\min}$, meaning $\alpha_t > 0$. Suppose $\pi_t \leq \pi_{\min}$. Then, $\alpha_{t+1} = 0$, implying $H_{t+1} = G - L$. From (8) in Lemma 1, α_t, π_{t-1} and π_t must satisfy the following conditions:

$$\begin{aligned} \frac{\partial}{\partial a_t} v(\alpha_t, \pi_{t-1}) &= -\ell + \delta\pi_{t-1}\theta e^{-\theta\alpha_t}(G - L) \geq 0 \\ \frac{\partial}{\partial a_{t+1}} v(0, \pi_t) &= -\ell + \delta\pi_t\theta(G - L) \leq 0 \end{aligned}$$

where the second inequality comes from the assumption $\pi_t \leq \pi_{\min}$. Cancelling terms, these two conditions imply

$$\frac{\pi_{t-1}}{\pi_t} \geq e^{\theta\alpha_t}. \tag{14}$$

However, by (1), for $\alpha_t > 0$, we have

$$\frac{\pi_{t-1}}{\pi_t} = e^{\theta\alpha_t} (1 - \pi_{t-1}) + \pi_{t-1} < e^{\theta\alpha_t} \tag{15}$$

The contradiction between (14) and (15) proves $\pi_t > \pi_{\min}$, and therefore $\alpha_{t+1} > 0$ and $v_{t+1} > L$. ■

The general "no-stop" result in this proposition might appear counterintuitive at first glance. The proof highlights a key reason: for stopping to be optimal at any time $t+1$, two conditions must be met simultaneously. One condition, outlined in (14), arises from the first-order condition for optimality, while the other, in (15), is derived from Bayes' rule. However, stopping in any period t creates a conflict between these conditions, rendering it impossible for both to hold simultaneously.

This result naturally raises an important question: if the agent never stops experimenting with the risky arm, does this imply that the good arm will always be discovered with probability 1? More formally, suppose that in period 0, Nature randomly determines the type of the risky arm $\omega \in \{0,1\}$ to be good, i.e., $\omega = 1$. Without knowing this outcome, will the agent, through persistent experimentation, eventually discover the truth about the risky arm's type?

Our next proposition directly addresses this question.

Let $P_\alpha(\omega) : \{0,1\} \to [0,1]$ denote the conditional probability of an eventual breakthrough under plan $\{(\alpha_t, \pi_{t-1})\}_{t=1}^{\infty}$, given that the type of the risky arm is ω. Obviously, we have $P_\alpha(0) = 0$.

Proposition 2 *Suppose $\pi_0 \in (\pi_{\min}, 1]$ and $\{(\alpha_t, \pi_{t-1})\}_{t=1}^{\infty}$ is optimal.*

(i) If $\pi_0 = 1$, then $P_\alpha(1) = 1$.

For $\pi_0 \in (\pi_{\min}, 1)$, the following results hold.

(ii) $\pi_t \to \pi_{\min}$ and $\alpha_t \to 0$ as $t \to \infty$.

(iii) Given that the risky arm is good, the conditional breakthrough probability equals

$$P_\alpha(1) = \frac{\pi_0 - \pi_{\min}}{\pi_0 (1 - \pi_{\min})} \in (0,1) \tag{16}$$

Consequently, the unconditional breakthrough probability is given by

$$\pi_0 P_\alpha(1) = \frac{\pi_0 - \pi_{\min}}{1 - \pi_{\min}} \in (0,1) \tag{17}$$

(iv) Both probabilities in (16) and (17) are increasing functions of π_0, δ, λ, n, and γ.

Proof. The conditional probability of no breakthrough by period T, given that the risky arm is good, equals $e^{-\theta \sum_{t=1}^{T} \alpha_t}$.

(i) Suppose $\pi_0 = 1$. By Proposition 1(i), $\alpha_t \equiv \hat{\alpha} > 0$. It follows that given $\omega = 1$, the conditional breakthrough probability $1 - e^{-\theta T \hat{\alpha}} \to 1$ as $T \to \infty$.

(ii) Suppose $\pi_0 \in (\pi_{\min}, 1)$. By Bayes' rule, the odds ratios of the posterior beliefs are updated as follows:

$$\begin{aligned} \frac{\pi_T}{1-\pi_T} &= e^{-\theta\alpha_T}\frac{\pi_{T-1}}{1-\pi_{T-1}} \\ &= e^{-\theta\sum_{t=1}^{T}\alpha_t}\frac{\pi_0}{1-\pi_0} \end{aligned} \tag{18}$$

The sequence of no-breakthrough posteriors $(\pi_t)_{t=0}^{\infty}$ forms a decreasing sequence, bounded below by $\pi_{\min}$. Thus, by the Monotone Convergence Theorem, π_t tends to a limit $\pi_\infty \geq \pi_{\min}$. Taking the limit as $T \to \infty$ in (18) yields

$$\frac{\pi_\infty}{1-\pi_\infty} = e^{-\theta\sum_{t=1}^{\infty}\alpha_t}\frac{\pi_0}{1-\pi_0} \tag{19}$$

This implies $\lim_{t\to\infty}\alpha_t = 0$ because the sum of the nonnegative numbers $\sum_{t=1}^{\infty}\alpha_t$ is finite. Consequently, taking the limit in (9), we obtain

$$\begin{aligned} \lim_{t\to\infty} H_{t+1} &= (\gamma - 1 + \lim_{t\to\infty}\alpha_{t+1})\ell + \delta \lim_{t\to\infty} e^{-\theta\alpha_{t+1}} \lim_{t\to\infty} H_{t+2} \\ &= (\gamma-1)\ell + \delta\lim_{t\to\infty} H_{t+2} \end{aligned}$$

Solving this gives

$$\lim_{t\to\infty} H_{t+1} = \frac{\gamma-1}{1-\delta}\ell \quad (= G - L)$$

By Lemma 1, for all t such that $\alpha_t \in (0,1)$, we have

$$\frac{\partial}{\partial a_t} v(\alpha_t, \pi_{t-1}) = -\ell + \delta\pi_{t-1}\theta e^{-\theta\alpha_t} H_{t+1} = 0$$

Thus, as $t \to \infty$, we obtain

$$-\ell + \delta\pi_\infty\theta\frac{\gamma-1}{1-\delta}\ell = 0 \Longrightarrow \pi_\infty = \pi_{\min}$$

Thus, the proof of (ii) is complete.

(iii) Given $\omega = 1$, the agent's breakthrough probability can now be computed from (19), substituting $\pi_{\min}$ for π_∞ :

$$\begin{aligned} P_\alpha(1) &= 1 - e^{-\theta\sum_{t=1}^{\infty}\alpha_t} \\ &= 1 - \frac{\pi_{\min}}{1-\pi_{\min}}\frac{1-\pi_0}{\pi_0} = \frac{\pi_0 - \pi_{\min}}{\pi_0(1-\pi_{\min})} \end{aligned} \tag{20}$$

which gives (16). Multiplying both sides by π_0 gives (17).

(iv) The results are straightforward to verify, noting that $\pi_{\min}$ is a decreasing function of δ, λ, n, and γ. ■

This proposition demonstrates that the good risky arm will be discovered with probability 1 if and only if the agent's prior belief is correct, i.e., $\pi_0 = 1$. This scenario corresponds to the "certain success project" discussed in Bergemann and Hege (2005). The certainty of success arises because, in this case, the posterior belief remains at 1 regardless of any failures in previous periods, thereby encouraging the agent to persist with a constant, positive allocation of time ($\hat{\alpha} > 0$) to the risky arm (Proposition 1). As a result, the probability of failure diminishes to 0 as the number of trials increases indefinitely.

In contrast, when $\pi_0 < 1$, even though the agent never stops experimenting, the lack of a breakthrough leads to repeated downward adjustments of the posterior belief toward $\pi_{\min}$. This, in turn, causes the time allocated to the risky arm (α_t) to gradually decline toward 0. While this decline alone does not necessarily imply that the probability of failure remains above 0 as $t \to \infty$, the result is driven by Bayes' rule in (18), combined with the breakthrough probability expressions in (16) and (17). Proposition 2 thus shows that "adequate learning" that the true state of the risky arm will be uncovered with certainty in the long run is not guaranteed. See, e.g., Aghion et al. (1991) for more about adequate learning in related contexts.

The necessary condition for α to be optimal—that (α_t, π_{t-1}) must tend to $(0, \pi_{\min})$ as $t \to \infty$—serves as a critical transversality condition in deriving the optimal policy plan. This condition ensures consistency with the underlying dynamic programming framework and the natural progression of beliefs and allocations in the absence of a breakthrough.

Part (iv) of Proposition 2 examines the comparative statics of breakthrough probabilities with respect to key parameters π_0, θ, δ, and γ. Before interpreting these results, it is useful to first analyze the comparative statics of the total maximum time that the representative agent plans to allocate to the risky arm.

Proposition 3 *Suppose $\pi_0 \in (\pi_{\min}, 1)$ and $\{(\alpha_t, \pi_{t-1})\}_{t=1}^{\infty}$ is an optimal plan. Define for each agent the maximum time that he plans to allocate to the risky arm by $A = \sum_{t=1}^{\infty} \alpha_t$.*

(i) A is an increasing function of (π_0, δ, γ).

(ii) Fix any (π_0, δ, γ) and consider A as a function of θ. There exists a unique θ^* that maximizes A such that $A'(\theta) > 0$ for $\theta < \theta^*$ and $A'(\theta) < 0$ for $\theta > \theta^*$.

Proof. By (16), we have

$$A = \frac{1}{\theta} \ln \left(\frac{\pi_0}{(1-\pi_0)} \frac{1-\pi_{\min}}{\pi_{\min}} \right)$$

(i) The right-hand side increases in π_0 and decreases in $\pi_{\min}$. Given that $\pi_{\min}$ is a decreasing function of δ and γ, A is an increasing function of (π_0, δ, γ).

(ii) Substituting (10), we get

$$A'(\theta) = -\frac{1}{\theta} \left(A - \frac{\delta(\gamma-1)}{\delta - \theta\delta + \theta\gamma\delta - 1} \right)$$

Therefore, $A' = 0$ implies

$$\begin{aligned} A &= \frac{\delta(\gamma-1)}{\delta - \theta\delta + \theta\gamma\delta - 1} \\ \text{and } A'' &= -\frac{1}{\theta} \left(A' - \frac{\partial}{\partial\theta} \frac{\delta(\gamma-1)}{\delta - \theta\delta + \theta\gamma\delta - 1} \right) \\ &= -\frac{1}{\theta} A^2 < 0 \end{aligned}$$

The statement (ii) is thus confirmed. ■

Since A is directly related to the breakthrough probabilities in (17)-(16), the comparative statics results in Proposition 3(i) and Proposition 2(iv) can be interpreted as follows: In (17), increasing π_0 has two effects. First, there is a direct effect, as a higher initial belief (π_0) increases the probability that the risky arm is good. Second, there is an indirect effect, where a higher π_0 encourages the agent to commit a greater total amount of time (A) to exploring the risky arm. Both effects are positive and contribute to a higher probability of a breakthrough. In (16), only the indirect effect is present, as belief updating through exploration remains the primary driver. Nonetheless, this indirect effect still leads to a higher probability of a breakthrough by inducing a greater allocation of time (A). Similarly, increases in δ or γ make exploration more appealing. This results in a positive, indirect effect on breakthrough probabilities, driven by an increased level of total time allocated to exploration (A).

Regarding the hazard rate, an increase in λ has a direct positive effect on the breakthrough probability. However, its impact on the agent's exploration intensity is ambiguous and may vary, as illustrated in Figure 1. Interestingly, Proposition 3(ii) highlights a new observation: $A(\lambda n)$ follows the *Goldilocks principle*. Specifically, optimal allocation to the risky arm increases with λn when the task is challenging (low λn) but decreases when the task becomes easier (high λn).

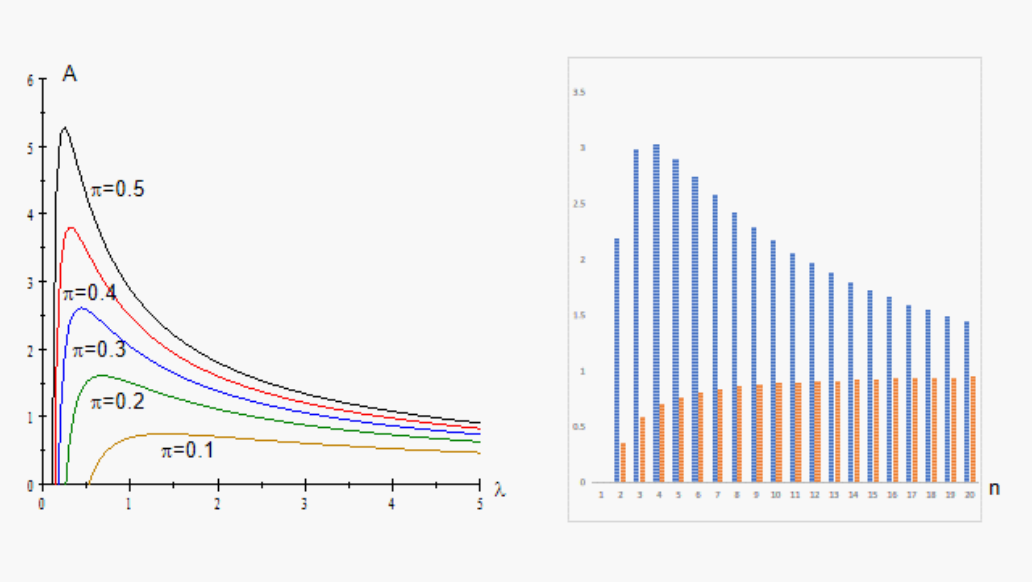

Figure 1: **The Goldilocks Principle in Optimal Allocation.**
Left Panel: The representative agent allocates the maximum total time to experimentation when the task difficulty is moderate—neither too challenging (low λ) nor too easy (high λ). This panel depicts the planned total exploration time from the current period onward (A) as a function of λ and the updated belief π, for parameters $\delta = 0.95$, $\gamma = 2$, and $n = 1$.
Right Panel: The representative agent maximizes total time committed to experimentation when the team size (n) is optimal—neither too small nor too large—as indicated by the taller bars. However, the probability of eventually discovering the good arm increases with team size, as shown by the shorter bars. This figure is generated with parameters $\lambda = 0.1$, $\pi_0 = 0.2$, $\delta = 0.9$ and $\gamma = 5$.

The intuition behind this result lies in the agent's confidence in achieving a breakthrough. When λ or n is high, there is already substantial confidence that a breakthrough will occur as long as the risky arm is good. A further increase in λ or n reduces the marginal need for additional time allocation due to the opportunity cost of exploration. Conversely, when the task is sufficiently difficult, a higher λ or n makes the breakthrough more attainable, motivating the agent to allocate more time to exploration in pursuit of the potential reward.

The finding in Proposition 3(ii) contrasts with Malueg and Tsutsui (1997), who predicted that optimal allocation should always increase with λ. In the binary-choice

literature, Choi (1997) was the first to observe indeterminate comparative static predictions regarding the effect of λ. Choi attributed this ambiguity to hazard rate uncertainty, demonstrating that increasing λ could extend the experiment's stopping time when the prior belief π_0 is low but shorten it when π_0 is high. In contrast, our study provides clearer insights: for any team size n, the total allocation A peaks at a unique $\lambda_n^* \in (0, \infty)$, which is a function of all the exogenous variables.

Similar non-monotonic comparative static results have been documented in other contexts, such as in Halac et al. (2016) regarding the optimal stopping time and in Bobtcheff and Levy (2017) concerning the optimal trigger time for investment. Learning offers one explanation for the negative effect of increasing λ when the task is relatively easy, as posterior beliefs decline more rapidly with higher λ. However, in the present setting with continuous strategies, we can show that the pattern in Figure 1 persists even in a single-period problem. This indicates that learning alone does not fully account for the observed non-monotonicity.

A more fundamental reason for this effect lies in the expected marginal benefit of allocation: it increases with λ when the exploration task is difficult (low λ) but decreases when the task becomes easy (high λ). Thus, the pattern observed in Figure 1 reflects an inherent structural property of the optimization problem, influenced by but not limited to the exponential probability function. We term this phenomenon the "Goldilocks principle," capturing the idea that optimal allocation is maximized when the exploration task is "just right" in difficulty.

The Goldilocks principle also applies to team size. As illustrated in Figure 1 (right, tall bars), there exists an optimal team size, $n = 4$, that maximizes each agent's total time commitment to the risky arm. For $n < 4$, increasing the team size "boosts morale," leading each team member to allocate more time to the risky arm. However, for $n \geq 4$, further increases in team size lead to a "taking-it-easy" effect, causing each member to reduce their committed time for experimentation.

Despite this decline in individual commitment, the overall effect of increasing n on the probability of an eventual breakthrough remains positive. As team size approaches infinity, the good risky arm will be discovered with probability 1, as indicated by the short bars in Figure 1 (right).

3.2 Optimal solution

We are ready now to present the main result of the basic model.

Theorem 1 *There exists a unique solution* $(v_t)_{t=1}^{\infty}$ *to (3)-(4), attained by a unique optimal policy plan* $\{(\alpha_t, \pi_{t-1})\}_{t=1}^{\infty}$. *For* $\pi_0 \in [0, \pi_{\min}]$, $(\alpha_t, \pi_{t-1}) \equiv (0, \pi_{\min})$ *and* $v_t \equiv L$. *For* $\pi_0 = 1$, $(\alpha_t, \pi_{t-1}) \equiv (\hat{\alpha}, 1)$ *and* $v_t \equiv v^*$, *as given in Proposition 1. For* $\pi_0 \in (\pi_{\min}, 1)$, *there exists a time-invariant, continuous and increasing backward transformation* $(\phi, \xi) : [0,1]^2 \to [0,1]^2$ *such that*

(i) in Scenario I, (11), for all $t \in \mathbb{N}$:

$$\alpha_t = \phi(\alpha_{t+1}, \pi_t) := \frac{1}{\lambda n} \ln\left(1 + \delta\lambda n(\gamma - 1 + \alpha_{t+1}) - \frac{1-\delta}{\pi_t}\right) \in (0,1) \quad (21)$$

$$\pi_{t-1} = \xi(\alpha_{t+1}, \pi_t) := \frac{\pi_t}{\pi_t + e^{-\lambda n \phi(\alpha_{t+1}, \pi_t)}(1 - \pi_t)} \in (\pi_{\min}, \pi_0) \quad (22)$$

(ii) in Scenario II, (12), there is a unique switching period $\tau \in \{0\} \cup \mathbb{N}$ *such that* $\alpha_t = 1$ *for all* $t \leq \tau$[3] *and the remaining plan* $\{(\alpha_t, \pi_{t-1})\}_{t=\tau+1}^{\infty}$ *satisfies (21)-(22).*

(iii) The sequence $(v_t)_{t=1}^{\infty}$ *satisfies (5), with the inputs from* $\{(\alpha_t, \pi_{t-1})\}_{t=1}^{\infty}$, *under both Scenarios I and II.*

Proof. See Appendix. ■

Theorem 1 demonstrates that, in any period t, the pair of optimal allocation and belief (α_t, π_{t-1}) is uniquely determined by the subsequent state via the mapping (ϕ, ξ) defined in (21)-(22). This mapping is time-invariant, continuous and increasing, offering a remarkably simple algorithm for determining the optimal allocation plan given any prior π_0 (see Figure 2).

As seen in the proof of this theorem, we introduce a novel approach featuring a form of backward *recursion* to derive the transformation (21)-(22). This approach consists of three steps:

1. *Initial Assumption*: We begin by considering an arbitrarily large but finite T, assuming the experiment stops after period T. In this context, the posterior belief in the last period is $\pi_{T-1} = (1 + \varepsilon_T)\pi_{\min}$ with $\varepsilon_T > 0$ but sufficiently small. Crucially, we treat π_{T-1} as a free variable rather than a function of the preceding state or history of the past allocations.

2. *Backward Optimization:*. Using backward induction while ensuring consistency with Bayes' rule for posteriors, we derive an optimal sequence $(\alpha_t, \pi_{t-1}|\varepsilon_T)_{t=1}^{T}$ parameterized by ε_T.

[3] If $\tau = 0$, without any consequence we define $\alpha_0 = 1$.

3. *Existence and Uniqueness*: We show the existence of a unique $\varepsilon_T(\pi_0)$ such that the sequence $(\alpha_t, \pi_{t-1}|\varepsilon_T(\pi_0))_{t=1}^T$ has the initial prior equal to π_0. Finally, by taking the limit as $T \to \infty$ and applying the transversality condition $(\alpha_T, \pi_{T-1}) \to (0, \pi_{\min})$, which implies $\varepsilon_T(\pi_0) \to 0$, we establish the existence and uniqueness of the optimal policy plan.

Since the mapping (21)-(22) is bijective, the functional relationship between any two adjacent pairs of states can also be expressed by a forward-moving law of motion, given the optimal period-1 state (α_1, π_0).

Theorem 2 *The optimal policy plan characterized in Theorem 1,* $\{(\alpha_t, \pi_{t-1})\}_{t=\tau+1}^{\infty}$ *after the possible switching period* $\tau \in \{0\} \cup \mathbb{N}$*, obeys the law of motion*

$$\alpha_{t+1} = \frac{1}{\delta\lambda n}\left(e^{\lambda n \alpha_t} + \frac{1-\delta}{\pi_t} - 1\right) - (\gamma - 1) \in (0,1) \tag{23}$$

$$\pi_t = \frac{\pi_{t-1} e^{-\lambda n \alpha_t}}{1 - \pi_{t-1}(1 - e^{-\lambda n \alpha_t})} \in (\pi_{\min}, \pi_0) \tag{24}$$

given the initial interior optimal state $(\alpha_{\tau+1}, \pi_\tau)$.

Proof. This is a straightforward corollary of Theorem 1, given the equivalence between (21)-(22) and (23)-(24). ■

Compared to Theorem 1, the forward motion described in Theorem 2 offers the advantage of being more practical to implement, for example, by an automated system or robot once the agent has specified the initial interior state. The agent may begin by fully allocating time to exploring the risky arm. As time progresses without a breakthrough, the agent will switch to a reduced allocation strategy at a specific time τ, allocating less than the full amount. Unlike the "bang-bang" strategy, where the allocation drops abruptly from 1 to 0 at a cutoff posterior belief, Theorem 2 predicts more gradual downward adjustments in response to the absence of breakthroughs.

It is important to emphasize that at the switching time τ, the pair of state variables $(\alpha_{\tau+1}, \pi_\tau)$ must satisfy optimality conditions. If not, the forward motion law described in (23)-(24) would lack an optimality foundation. Thus, applying Theorem 2 requires first deriving $(\alpha_{\tau+1}, \pi_\tau)$ using the backward recursion algorithm outlined in Theorem 1.

Figure 2 depicts the dynamics of the paired state (π_{t-1}, α_t) for Scenario I (left) and Scenario II (right). Consistent with the Goldilocks principle, when λ or n is

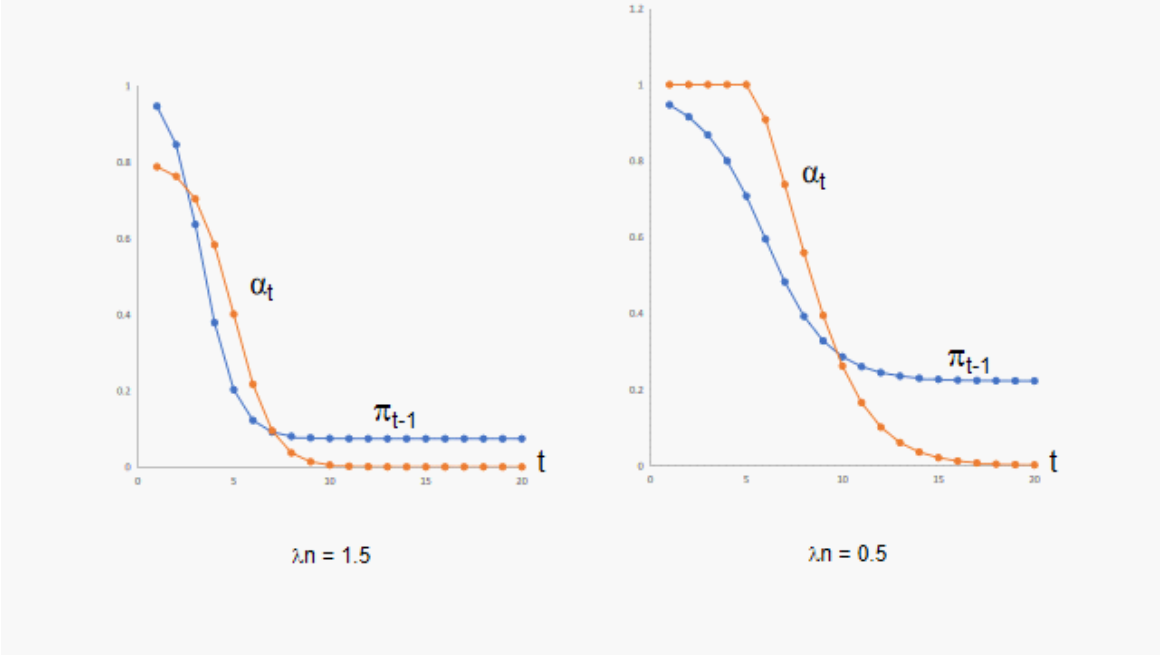

Figure 2: Both figures assume $\delta = 0.9$, $\gamma = 2$, starting with $\pi_0 = 0.95$. They differ only in λn.

relatively high ($\lambda n = 1.5$), expecting an easy breakthrough the agent does not allocate full time to the risky arm. But after a few periods of failure, the posterior quickly drops toward $\pi_{\min}$ indicating that the risky arm is most likely bad. As a result, α_t quickly drops toward 0. When λn is reasonably encouraging but not sufficiently low ($\lambda n = 0.5$), the agent allocates full time to the risky arm, switching to a more moderate allocation strategy after some periods of failure as the posterior declines gradually towards $\pi_{\min}$.

The basic model presented in this section offers broad applicability within the context of exponential bandit theory. We demonstrate its usefulness through examining the consequences of strategic experimentation under noncooperative behavior in Section 4, and exploring optimal contracting solutions for noncooperative agents in Section 5.

4 Strategic Experimentation

In this section, we apply the basic model to strategic experimentation, assuming $n \geq 2$ homogeneous agents (or players). Each agent $i \in \{1, 2, ..., n\}$ allocates in each

period t a fraction of his time, denoted $a_{i,t} \in [0,1]$, to the risky arm and $1-a_{i,t}$ to the safe arm. Again, a breakthrough occurs when at least one agent obtains conclusive evidence that the risky arm is good, in which case each agent will enjoy an expected present value $G = \gamma L \in (L, \infty)$ from the risky arm.

Conditional on that the risky arm is good ($\omega = 1$), each agent i independently has a breakthrough probability $1 - e^{\lambda a_{i,t}}$ in any period t, so that the probability of a breakthrough equals $F(\mathbf{a}_t) = 1 - e^{\lambda \sum_{i=1}^{n} a_{i,t}}$, given $\mathbf{a}_t = (a_{1,t}, ..., a_{n,t}) \in [0,1]^n$. We assume that a breakthrough by any agent is publicly observable, ending everyone's need for further experimentation. However, each agent's allocation decision is private information, so that $a_{i,t}$ is unknown to agents other than i (see Bonatti and Hörner (2011)).

We model the agents as playing a dynamic noncooperative game under incomplete information due to hidden actions. They start with a common prior belief that the probability of $\omega = 1$ is $\pi_0 \in [0,1]$. While a (pure) strategy for player i could be defined by a pure stationary Markov policy as a function of the common belief, this approach is cumbersome and unnecessary, as discussed in the basic model. We thus define player i's strategy by a feasible allocation plan $\kappa_i = (\kappa_{i,t})_{t=1}^{\infty}$ such that each $\kappa_{i,t} \in [0,1]$ is measurable with respect to the information available at the start of period t. The strategy terminates once the game concludes upon a breakthrough. A strategy profile is then defined as the collection of all players' strategies $\kappa = (\kappa_1, ..., \kappa_n)$. In each period t, let $\boldsymbol{\kappa}_t = (\kappa_{1,t}, ..., \kappa_{n,t})$ represent the vector of players' allocations, and let $\kappa_{-i,t}$ denote the allocation vector excluding player i's allocation $\kappa_{i,t}$. The total allocation of all players in period t is denoted by $K_t = \sum_{i=1}^{n} \kappa_{i,t}$, with $K_{-i,t} = K_t - \kappa_{i,t}$ representing the total allocation by all players except player i.

Given strategy profile κ and that no breakthrough has occurred till $t-1$, the players' common belief is updated as follows:

$$\eta_t = \begin{cases} 1^+ & \text{if breakthrough occurs in } t \\ \frac{\eta_{t-1} e^{-\lambda K_t}}{1 - \eta_{t-1}(1 - e^{-\lambda K_t})} & \text{if no breakthrough in } t \end{cases} \tag{25}$$

where $\eta_0 = \pi_0$. Following this updating rule, player i's expected payoff is expressed as

$$\sum_{t=1}^{\infty} \delta^{t-1} \left(\prod_{s=1}^{t-1} (1 - \eta_{s-1}(1 - e^{-\lambda K_s})) \left[(1 - \kappa_{i,t})\ell + \delta \eta_{t-1}(1 - e^{-\lambda K_t}) G \right] \right) \tag{26}$$

Each player i seeks to maximize his expected payoff (26) by choosing a policy plan

$(\kappa_{i,t}, \eta_{t-1})_{t=1}^{\infty}$ (see Definition 1), taking as given the other players' strategy profile κ_{-i} and the common belief process $(\eta_{t-1})_{t=1}^{\infty}$.

At the start of any period t, a player's information set consists of his private information regarding past allocations and the public news of whether a breakthrough has occurred. If a breakthrough occurs, the game ends. If no breakthrough occurs, the no-breakthrough news reveals nothing about any player's private past allocations. Consequently, the process of common beliefs along any equilibrium path remains unaffected by deviations from equilibrium strategies. This fact simplifies the analysis of perfect Bayesian equilibria in this game, allowing them to be analyzed as Nash equilibria without requiring additional refinements. In this framework, each player selects a personal policy plan, treating the other players' strategies and the updated common belief as given.

Extending the basic model (see, e.g., Definition 1), we define $\left(\boldsymbol{\kappa}_t, \eta_{t-1}\right) \in [0,1]^{n+1}$ as the state in period t where $\boldsymbol{\kappa}_t$ represents the vector of players' allocations and η_{t-1} the common belief at the start of the period.

Definition 2 *We say that the process* $\left\{\left(\boldsymbol{\kappa}_t, \eta_{t-1}\right)\right\}_{t=1}^{\infty}$ *constitutes a Perfect Bayesian Equilibrium (PBE, or simply Nash equilibrium) if:*

1. *The common belief* η_t *is updated according to (25).*
2. *For each player* i*, given the strategies of all other players,* κ_{-i}*, the process* $(\kappa_{i,t}, \eta_{t-1})_{t=1}^{\infty}$ *is an optimal policy plan.*

The optimality of the policy plan ensures sequential rationality. In the probability-0 events where a player—perhaps by mistake—deviates from the equilibrium path, we may assume that they will update their private belief and adjust subsequent allocations accordingly. But since such deviations are suboptimal and unobservable by other players, the common equilibrium beliefs remain unaffected. This completes the formal description of the Perfect Bayesian Equilibrium (PBE).

Consequently, if $\left\{\left(\boldsymbol{\kappa}_t, \eta_{t-1}\right)\right\}_{t=1}^{\infty}$ is an equilibrium, then for all $i = 1, ..., n$ and $t \in \mathbb{N}$:

$$v_i(\boldsymbol{\kappa}_t, \eta_{t-1}) = \max_{a \in [0,1]} v_i(a, \kappa_{-i,t}, \eta_{t-1}) \text{ s.t. } (25) \tag{27}$$

where $v_i(\cdot, \kappa_{-i,t}, \eta_{t-1}) : [0,1] \to [L, G]$ represents agent i's conditional expected payoff given no breakthrough has occurred:

$$\begin{aligned} v_i(a, \kappa_{-i,t}, \eta_{t-1}) &= (1-a)\ell + \delta\eta_{t-1}\left(1 - e^{-\lambda(a+K_{-i,t})}\right) G & (28) \\ &\quad +\delta(1-\eta_{t-1}\left(1 - e^{-\lambda(a+K_{-i,t})}\right))v_i(\boldsymbol{\kappa}_{t+1}, \eta_t) & (29) \end{aligned}$$

Following the approach in the basic model, we now define a cutoff belief level under strategic experimentation:

$$\eta_{\min} = \frac{1-\delta}{\delta\lambda\left(\gamma-1\right)} = n\pi_{\min}$$

Let κ_{-i} be given and suppose κ_i is player i's optimal response. Then, in Lemma 1 with $F(a) = 1-e^{-\lambda(a+K_{-i,t})}$, for all $t \in \mathbb{N}$ the marginal effect of a one-stage deviation from $\kappa_{i,t}$ is given by

$$\frac{\partial}{\partial a} v_i(a, \kappa_{-i,t}, \eta_{t-1}) = -\ell + \delta\eta_{t-1}\lambda e^{-\lambda(a+K_{-i,t})} H_{i,t+1} \tag{30}$$

where $H_{i,t+1} \in (0, G-L]$, and is defined recursively by

$$H_{i,t+1} = (\gamma - 1 + \kappa_{i,t+1})\ell + \delta e^{-\lambda K_{t+1}} H_{i,t+2} \tag{31}$$

By the "one-stage deviation principle," for $\left\{\left(\boldsymbol{\kappa}_t, \eta_{t-1}\right)\right\}_{t=1}^{\infty}$ to be an equilibrium each agent must be deterred from deviating for one period and then following the equilibrium strategy (see, e.g., Athey and Segal (2013)).

Theorem 3 *Suppose* $\eta_0 \in (\eta_{\min}, 1]$.

(i) If there exists any equilibrium $\left\{\left(\boldsymbol{\kappa}_t, \eta_{t-1}\right)\right\}_{t=1}^{\infty}$, *then the equilibrium must be symmetric, in that* $\kappa_{i,t} \equiv \kappa_t \in [0,1]$ *for all* $t \in \mathbb{N}$.

For all $\eta_0 \in (\eta_{\min}, 1)$,

(ii) the game, in any equilibrium, never stops without a breakthrough, i.e., $\kappa_t > 0$ *and* $\eta_{t-1} > \eta_{\min}$ *for all* $t \in \mathbb{N}$, *such that* $\lim_{t\to\infty}\left(\boldsymbol{\kappa}_t, \eta_{t-1}\right) = (\mathbf{0}, \eta_{\min})$.

(iii) there exists a unique symmetric equilibrium, which is characterized by a unique switching time $\tau \in \{0\} \cup \mathbb{N}$ *such that* $\kappa_t = 1$ *for all* $t \leq \tau$, *and the sequence of the joint states* $\left\{\left(\boldsymbol{\kappa}_t, \eta_{t-1}\right)\right\}_{t=\tau+1}^{\infty}$ *satisfies the backward transformational relation:*

$$\kappa_t = \frac{1}{\lambda n}\ln\left(1 + \delta\lambda\left(\gamma - 1 + \kappa_{t+1}\right) - \frac{1-\delta}{\eta_t}\right) \in (0,1) \tag{32}$$

$$\eta_{t-1} = \frac{\eta_t}{\eta_t + e^{-\lambda n\kappa_t}(1-\eta_t)} \in (\eta_{\min}, \eta_0) \tag{33}$$

(iv) If $\eta_0 = 1$, *then* $\kappa_t \equiv \hat{\kappa}$ *such that* $\hat{\kappa} = 1$ *when* $\delta e^{-\lambda n}\left(1 + \lambda\gamma\right) \geq 1$ *and otherwise,* $\hat{\kappa} \in (0,1)$ *is the unique solution of*

$$\delta e^{-\lambda n\hat{\kappa}}\left(\lambda(\gamma - 1 + \hat{\kappa}) + 1\right) - 1 = 0 \tag{34}$$

Proof. See Appendix. ■

The equilibrium uniqueness result in Theorem 3 stands in contrast to the findings of multiple (asymmetric) equilibria reported in prior studies (e.g., Keller et al. (2005); Bonatti and Hörner (2011)). The proof of Theorem 3 identifies two key factors underlying this result, both stemming from the discrete-time exponential bandit framework:

1. Strict Concavity of Payoffs: Each player's expected payoff is a strictly concave function of their action, ensuring the existence of a unique pure allocation plan necessary for an optimal response.

2. Dependence on Aggregate Allocations: A player's marginal expected payoff in equilibrium depends solely on the total sum of all players' allocations. As a result, starting with a homogeneous population of agents, every player faces the same equilibrium trade-off. This symmetry in trade-offs leads all players to choose the same optimal response, resulting in a symmetric equilibrium.

Consistent with previous studies, Theorem 3 indicates that the probability of an eventual breakthrough in the strategic equilibrium is equivalent to what a single agent can achieve. In the initial periods, the strategic players typically allocate a joint effort K_t that is strictly lower than the single-agent allocation. Over time, as no breakthrough occurs, the strategic common beliefs surpass the single-agent posterior belief. Once this difference becomes sufficiently large, the joint allocations of the strategic players exceed those of the single agent (see Figure 3).

Following Bonatti and Hörner (2011), the more flattened shape of the equilibrium joint-allocation path can be interpreted as a "procrastination effect." This effect arises from each player's tendency to free ride more as the number of players increases. The result in Theorem 3 concerning the "certain success" case, where $\eta_0 = 1$, provides a framework to highlight this tendency, as formalized in the following proposition.

Proposition 4 *Suppose $\eta_0 = 1$. Denote by $\hat{\kappa}(n)$ the constant symmetric equilibrium allocation in each period by each player when the team size is n, and assume $\delta e^{-\lambda}(1+\lambda\gamma) \leq 1$. Then, $n\hat{\kappa}(n)$ is a strictly decreasing sequence:*

$$\hat{\kappa}(1) > 2\hat{\kappa}(2) > ... > n\hat{\kappa}(n) > (n+1)\hat{\kappa}(n+1)$$

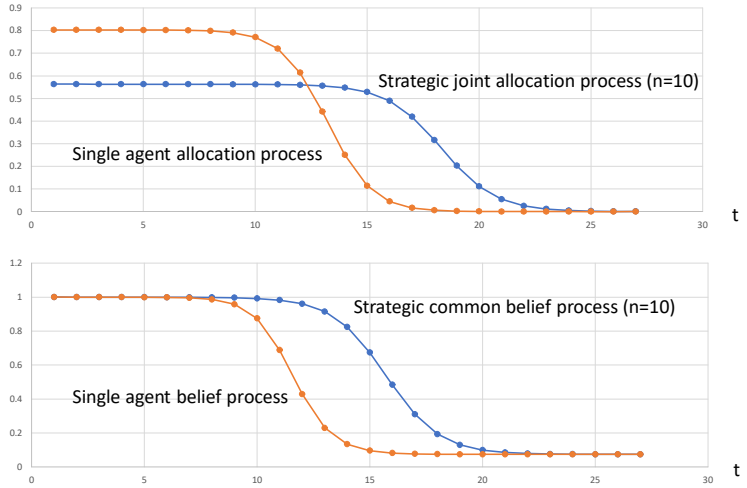

Figure 3: **Single-agent optimal policy plan vs. multi-player strategic equilibrium path.** Both processes entail the same probability of an eventual breakthrough, although their dynamic behaviors differ. The figures assume $\delta = 0.9$, $\lambda = 1.5$, $\gamma = 2$, and the prior belief $\pi_0 = 0.999$.

Proof. Denote $K(n) = n\hat{\kappa}(n)$. From (34), $K(n)$ is the solution of

$$M(K(n), n) := \delta e^{-\lambda K(n)} \left(\lambda(\gamma - 1 + \frac{K(n)}{n}) + 1 \right) - 1 = 0, \quad \forall n$$

where the auxiliary function M satisfies $M(0, n) > 0$ (by Assumption 1) and a single-crossing condition from above:

$$\begin{aligned} M(K, n) &= 0 \text{ implies} \\ \frac{\partial}{\partial K} M(K, n) &= -\delta\lambda e^{-\lambda K} \left(\lambda(\gamma - 1 + \frac{K}{n}) + 1 \right) + \delta e^{-\lambda K} \lambda \frac{1}{n} \\ &= -\lambda \left(1 - \frac{\delta e^{-\lambda K}}{n} \right) < 0 \end{aligned}$$

Since $M(K, n)$ decreases in n, we have

$$\begin{aligned} M(K(n), n) &= 0 \\ &\Rightarrow \quad M(K(n), n+1) < 0 = M(K(n+1), n+1) \\ &\Rightarrow \quad K(n) > K(n+1) \end{aligned}$$

■

This proposition confirms our intuition that given any time horizon $T < \infty$, the probability of a breakthrough before T decreases as the number of players n increases when agents engage in strategic experimentation. Consequently, this probability is maximized when experimentation is performed by a single agent:

$$1 - e^{-\lambda \hat{\kappa}(1) T} > ... > 1 - e^{-\lambda n \hat{\kappa}(n) T}, \quad \forall T$$

even though the long-run probabilities converge to the same limit of 1 as T tends to infinity.

5 Optimal contracting

To address the inefficiency caused by free riding, this section explores ways to improve Pareto efficiency under the assumption that agents can make monetary transfers among themselves. We propose a dynamic profit-sharing agreement based on publicly observable and verifiable outcomes. In the present context, the observable outcome is whether a breakthrough occurs in a given period and, if so, which agent(s) achieved it.

If a contract can be designed that incentivizes each agent to allocate more time to experimentation, and if the resulting expected payoff exceeds that of the free-riding equilibrium for all agents, then it will be in every agent's interest to participate in such an agreement.

Definition 3 *Let the agents who achieve a breakthrough in a given period be referred to as "winners" and those who do not as "losers." A budget-balanced dynamic profit-sharing rule for strategic experimentation is a sequence of payment rules* $\vartheta = (\vartheta_{k,t} : k = 0, 1, ..., n;\ t \in \mathbb{N})$ *where* $\vartheta_{k,t} \in [0, 1]$ *represents the percentage of* G *that each loser must pay to each of the* k *winners if the first breakthrough occurs in period* t *when* $0 \leq k \leq n$ *agents achieve a breakthrough in that period. Whenever* $k > 0$ *in a given period, the contract terminates starting from the subsequent period.*

In addition, the profit sharing rule further satisfies the limited liability constraint if for all $t \in \mathbb{N}$, $0 \leq k\vartheta_{k,t} \leq 1$.

Among the n agents, for $k = 0, 1, ..., n$, if each agent allocates α_t of their time to the risky arm, then, based on the Bernoulli distribution, the probability of exactly k breakthroughs in a period t equals

$$\Phi(k, n, p) := \frac{n!}{k!(n-k)!}(1 - e^{-\lambda\alpha_t})^k e^{-(n-k)\lambda\alpha_t}$$

where $p = 1 - e^{-\lambda\alpha_t}$ and $\frac{n!}{k!(n-k)!}$ represents the binomial coefficient.

Now consider an arbitrary player and suppose all other players allocate α_t in period t. Let $R_t(a)$ denote this player's conditional expected reward under the sharing rule $(\vartheta_{k,t})_{k=0}^n$ in period t, assuming he chooses action a, given that at least one breakthrough will occur among the players in period t. The conditional expected reward for this player is then given by:

$$R_t(a) = \frac{1 - e^{-\lambda a}}{1 - e^{-\lambda(a+(n-1)\alpha_t)}} \sum_{k=0}^{n-1} \Phi(k, n-1, p)\left(1 + (n-1-k)\vartheta_{k,t}\right) G \tag{35}$$

$$+ \frac{e^{-\lambda a}}{1 - e^{-\lambda(a+(n-1)\alpha_t)}} \sum_{k=1}^{n-1} \Phi(k, n-1, p)\left(1 - k\vartheta_{k,t}\right) G \tag{36}$$

The term in (35) specifies the player's expected reward in the event he wins. Conditional on being a winner, the reward depends on how many among the other $n-1$ players also achieve a breakthrough in period t. For example, if k others also win, then there will be $n-1-k$ losers. Each loser will pay the winning player $\vartheta_{k,t}G$,

so, including his own gain of G, the winning player will receive $(1+(n-1-k)\vartheta_{k,t})\,G$ reward in total.

The term in (36) specifies the player's expected reward (which may be negative) in the event he loses. Conditional on being a loser, his reward also depends on how many among the other $n-1$ players make a breakthrough in period t. For instance, if k others win, then the losing player will pay each of the k winning player $\vartheta_{k,t}G$ so that, subtracted from his own gain of G, the player will receive $(1-k\vartheta_{k,t})\,G$ in total. It is important to note that when the player is a loser, k starts from 1 since at least one other player must win for this condition to apply.

Lemma 2 *Under any payment rule ϑ defined in Definition 3, when everyone follows the cooperative strategy $\{(\alpha_t,\pi_{t-1})\}_{t=1}^{\infty}$ as outlined in the basic model, then $R_t(\alpha_t)\equiv G$.*

Proof. Conditional on a breakthrough, the total gain for the n players is nG. The payment rules defined in Definition 3 are symmetric, in that no player has any advantage or disadvantage when they contribute the same allocation α_t in every period t. Therefore, given that the rules are (strictly) budget balancing, each player's conditional expected reward equals G. ■

Now suppose all other players adhere to the representative agent's cooperative strategy $\{(\alpha_t,\pi_{t-1})\}_{t=1}^{\infty}$, where π_{t-1} is now interpreted as the common belief. Let us analyze the impact of a one-stage deviation on an arbitrary player's payoff, denoted by $\hat{v}$. In (4), substituting $1-e^{-\lambda(a+(n-1)\alpha_t)}$ for $F(a)$ yields

$$\begin{aligned}
&\hat{v}(a,\pi_{t-1})\\
=\;&(1-a)\ell+\delta\pi_{t-1}\left(1-e^{-\lambda(a+(n-1)\alpha_t)}\right)R_t(a)+\delta\left(1-\pi_{t-1}\left(1-e^{-\lambda(a+(n-1)\alpha_t)}\right)\right)\hat{v}(\alpha_{t+1},\hat{\pi}_t)\\
=\;&(1-a)\ell+\delta\pi_{t-1}\left(1-e^{-\lambda(a+(n-1)\alpha_t)}\right)G+\delta\left(1-\pi_{t-1}\left(1-e^{-\lambda(a+(n-1)\alpha_t)}\right)\right)\hat{v}(\alpha_{t+1},\hat{\pi}_t)\\
&+\delta\pi_{t-1}\left(1-e^{-\lambda(a+(n-1)\alpha_t)}\right)(R_t(a)-G) \qquad (37)
\end{aligned}$$

where $\hat{\pi}_t=\pi_t$ when $a=\alpha_t$, and in general

$$\hat{\pi}_t=\begin{cases}1^+ & \text{if breakthrough occurs in } t\\ \frac{\pi_{t-1}e^{-\lambda(a+(n-1)\alpha_t)}}{1-\pi_{t-1}(1-e^{-\lambda(a+(n-1)\alpha_t)})} & \text{if no breakthrough in } t\end{cases} \qquad (38)$$

By Lemma 2, for $a=\alpha_t$, the last term in (37) vanishes, so

$$\hat{v}(\alpha_t,\pi_{t-1})=(1-\alpha_t)\ell+\delta\pi_{t-1}\left(1-e^{-\lambda n\alpha_t}\right)G+\delta\left(1-\pi_{t-1}\left(1-e^{-\lambda n\alpha_t}\right)\right)\hat{v}(\alpha_{t+1},\pi_t)$$

which implies $\hat{v}(\alpha_t, \pi_{t-1}) = v(\alpha_t, \pi_{t-1})$, the optimal solution for the cooperative case, for all $t \in \mathbb{N}$.

Theorem 4 *Suppose the players agree on a contract involving a dynamic profit-sharing rule ϑ from the outset, defined as*

$$\vartheta_{k,t} = 0, \ \forall k \neq 1 \text{ and } \vartheta_{1,t} = \frac{1-\delta}{\delta\gamma\pi_{t-1}\lambda n e^{-\lambda(n-1)\alpha_t}} \tag{39}$$

Under this rule:

1. *ϑ satisfies budget balancing and limited liability, ensuring $\vartheta_{1,t} \in (0, \frac{1}{\gamma}(\gamma-1))$,*
2. *The contract induces the Pareto optimal solution in equilibrium.*

Proof. By Lemma 1, the cooperative solution is characterized by

$$\frac{\partial}{\partial a} v(\alpha_t, \pi_{t-1}) = -\ell + \delta\pi_{t-1} n\lambda e^{-n\lambda\alpha_t} H_{t+1} \begin{cases} = 0 & \text{if } \alpha_t \in (0,1) \\ \geq 0 & \text{if } \alpha_t = 1 \end{cases} \tag{40}$$

where

$$H_{t+1} = (\gamma - 1 + \alpha_{t+1})\ell + \delta e^{-n\lambda\alpha_t} H_{t+2} \leq G - L \tag{41}$$

In (35)-(36), substitute $\vartheta_{k,t} = 0$ for all $k \neq 1$, write ϑ_t for $\vartheta_{1,t}$, and multiply both sides by $1 - e^{-\lambda(a+(n-1)\alpha_t)}$. The player's ex ante expected reward equals

$$\begin{aligned}
& \left(1 - e^{-\lambda(a+(n-1)\alpha_t)}\right) R_t(a) \\
= \ & (1 - e^{-\lambda a}) \sum_{k=0}^{n-1} \Phi(k, n-1, p) \left(1 + \Phi(0, n-1, p)(n-1)\vartheta_t\right) G \\
& + e^{-\lambda a} \sum_{k=0}^{n-1} \Phi(k, n-1, p) \left(1 - \Phi(1, n-1, p)\vartheta_t\right) G - e^{-\lambda(a+(n-1)\alpha_t)} G \\
= \ & (1 - e^{-\lambda a}) e^{-(n-1)\lambda\alpha_t} (n-1)\vartheta_t G - e^{-\lambda a} (n-1) (1 - e^{-\lambda\alpha_t}) e^{-(n-2)\lambda\alpha_t} \vartheta_t G \\
& + (1 - e^{-\lambda(a+(n-1)\alpha_t)}) G \\
= \ & (n-1)\vartheta_t G e^{-(n-2)\lambda\alpha_t} \left[(1 - e^{-\lambda a}) e^{-\lambda\alpha_t} - e^{-\lambda a}(1 - e^{-\lambda\alpha_t})\right] + (1 - e^{-\lambda(a+(n-1)\alpha_t)}) G
\end{aligned}$$

from which we confirm $R_t(\alpha_t) = G$, and derive

$$\begin{aligned}
& \left(1 - e^{-\lambda(a+(n-1)\alpha_t)}\right) (R_t(a) - G) \\
= \ & (n-1)\vartheta_t G e^{-(n-2)\lambda\alpha_t} \left[(1 - e^{-\lambda a}) e^{-\lambda\alpha_t} - e^{-\lambda a}(1 - e^{-\lambda\alpha_t})\right]
\end{aligned}$$

Differentiating both sides yields

$$\frac{\partial}{\partial a}\left[\left(1-e^{-\lambda(a+(n-1)\alpha_t)}\right)\left(R_t(a)-G\right)\right]=(n-1)\vartheta_t G e^{-(n-2)\lambda\alpha_t}\lambda e^{-\lambda a}$$

Now, for $F(a)=1-e^{-\lambda(a+(n-1)\alpha_t)}$, we can apply Lemma 1 and obtain from differentiating (37):

$$\begin{aligned}\frac{\partial}{\partial a}\hat{v}(a,\pi_{t-1}) &= -\ell+\delta\pi_{t-1}\lambda e^{-\lambda(a+(n-1)\alpha_t)}H_{t+1}\\ &\quad +\delta\pi_{t-1}(n-1)e^{-(n-2)\lambda\alpha_t}\lambda e^{-\lambda a}\vartheta_t G\end{aligned}$$

Clearly, $\frac{\partial^2}{\partial a^2}\hat{v}(a,\pi_{t-1})<0$, so that the first-order condition is both necessary and sufficient for an interior solution. Now, in light of (40), let us define

$$O_{t+1}=\frac{\ell}{\delta\pi_{t-1}n\lambda e^{-n\lambda\alpha_t}}\begin{cases}=H_{t+1} & \text{if } \alpha_t\in(0,1)\\ \leq H_{t+1} & \text{if } \alpha_t=1\end{cases} \tag{42}$$

Define the profit sharing rule as

$$\vartheta_t=e^{-\lambda\alpha_t}\frac{O_{t+1}}{G} \tag{43}$$

Then, at $a=\alpha_t$,

$$\begin{aligned}\frac{\partial}{\partial a}v(\alpha_t,\pi_{t-1}) &= -\ell+\delta\pi_{t-1}\lambda e^{-\lambda n\alpha_t}H_{t+1}+\delta\pi_{t-1}(n-1)e^{-(n-1)\lambda\alpha_t}\lambda\vartheta_t G\\ &\geq -\ell+\delta\pi_{t-1}\lambda e^{-\lambda n\alpha_t}O_{t+1}+\delta\pi_{t-1}(n-1)e^{-(n-1)\lambda\alpha_t}\lambda\vartheta_t G\\ &= 0\end{aligned}$$

where the inequality holds as an equality for all $\alpha_t\in(0,1)$. Since $O_{t+1}\leq H_{t+1}\leq G-L$, then the sharing rule in (43) is bounded between 0 and $(G-L)/G=(\gamma-1)/\gamma$, satisfying limited liability. Substituting O_{t+1} in (43) gives (39).

We have thus shown that if the agent deviates from α_t in any period t and then moves back to the equilibrium path by choosing $\alpha_{t+1},\alpha_{t+2},...$in the future, he will not find it optimal. This establishes that the allocation-belief sequence $(\alpha_t,\pi_{t-1})_{t=1}^{\infty}$ is the player's best policy plan and therefore $(\boldsymbol{\alpha}_t,\pi_{t-1})_{t=1}^{\infty}$ forms a symmetric PBE. ■

As established in the team moral hazard literature, when each agent's contribution to the outcome is observable, numerous approaches exist for designing a first-best contract (e.g., Holmstrom (1982)). Theorem 4 introduces one of the simplest profit-sharing rules for our bandit setting, which is both budget balancing and satisfies the

limited liability constraint. In this agreement, the lucky winner must be the sole agent who achieves a breakthrough. When limited liability is not a concern, it can be shown that the sharing rule:

$$\vartheta_{0,t} = \vartheta_{n,t} = 0 \text{ and } \vartheta_{k,t} \equiv \vartheta_t = \frac{(1-\delta)}{\delta\gamma\pi_{t-1}n\lambda e^{-\lambda\alpha_t}} \text{ for } 0 < k < n$$

also induces the Pareto optimal equilibrium.

Notably, since $\vartheta_{1,t} \in (0,1)$, the optimal contract in Theorem 4 allows all losers to share in the breakthrough reward (see Figure 4). This result differentiates the optimal contract in our discrete-time exponential bandit model from existing studies, such as the optimal contests designed by Halac et al. (2017) in continuous time, where losers receive no reward. The pursuit of Pareto optimality in our contract design is also a key reason for this distinction.

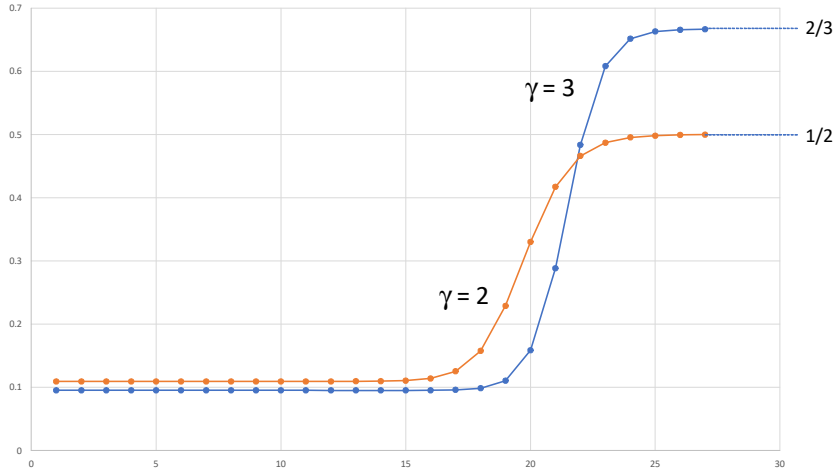

Figure 4: **Dynamic Profit Sharing Rule**. The figure depicts the percentage $\vartheta_{1,t}$ of every loser's reward G to be paid to the winner. This percentage increases with t when the game continues without success, and is capped by $(\gamma-1)/\gamma$ (< 1) as $t \to \infty$. Here, it is assumed that $\delta = 0.9$, $\lambda = 0.15$, $n = 10$, and $\pi_0 = 0.999$.

6 Conclusion

We conclude with reflections on the practical relevance of our findings and their implications. The non-stopping result emphasizes the optimality of persistently pursuing a breakthrough while gradually reducing investment rather than abruptly terminating exploration after repeated failures. This insight aligns with the adage "persistence is the mother of success," providing a rational foundation for traits often associated with success—optimism, determination, and resilience. However, this result is derived under the idealized assumption that resource allocations can be infinitesimally small. In practice, projects incur minimum fixed costs, such as office space, wages, and managerial attention, which often lead to the abandonment of efforts after extended failures. Consequently, the non-stopping conclusion should be interpreted qualitatively, serving as a guiding principle rather than a prescriptive rule.

The Goldilocks principle, which identifies an optimal level of task difficulty that maximizes incentives for exploration, has clear implications for both individual and organizational decision-making. For researchers, this suggests that productivity peaks when a project's difficulty is "just right"—neither too easy nor too challenging. Similarly, for principals hiring agents to undertake experimentation tasks, aligning the agent's abilities with the task's complexity optimally motivates effort. Furthermore, our finding that an optimal team size maximizes individual incentives underscores the importance of carefully selecting team configurations for collaborative endeavors.

The transformations characterizing bandit dynamics in Theorems 1 and 2 provide practical methods for computing optimal allocation plans and could serve as building blocks for extending the model to richer frameworks. For instance, the dynamic profit-sharing rules derived in Theorem 4 offer valuable insights for designing patents and contests, particularly in scenarios where breakthroughs generate significant spillover effects.

Despite its contributions, the current model omits many important features. Extending the framework to broader contexts remains a promising avenue for advancing both the theoretical and practical understanding of dynamic decision-making under uncertainty.

7 Appendix

Proof of Lemma 1. Let us transform the program (3)-(4) into a mathematically equivalent form:

$$v(\alpha_t, \pi_{t-1}) - G = \max_{a \in [0,1]} \left(v(a, \pi_{t-1}) - G \right) \text{ s.t. } (1)$$

where

$$\begin{aligned} & v(a, \pi_{t-1}) - G \\ = \; & (1-a)\ell - (1-\delta)G + (1 - \pi_{t-1}F(a))\,\delta\,[v(\alpha_{t+1}, \pi_t) - G] \end{aligned} \tag{44}$$

To simplify notation, denote

$$h(\alpha_t) = (\gamma - 1 + \alpha_t)\ell \quad (= (1-\delta)G - (1-\alpha_t)\ell) \tag{45}$$

$$q_t = \pi_{t-1}F(\alpha_t) \tag{46}$$

Then, expanding (44) yields

$$v(\alpha_t, \pi_{t-1}) - G = -h(\alpha_t) + (1-q_t)\,\delta\,[v(\alpha_{t+1}, \pi_t) - G] \tag{47}$$

$$= -h(\alpha_t) - \sum_{s=1}^{\infty} \delta^s \left(\prod_{r=0}^{s-1} (1 - q_{t+r}) \right) h(\alpha_{t+s}) \tag{48}$$

The term in large brackets in (48) represents the conditional probability of no breakthrough over the next s periods. This can be re-written as:

$$\prod_{r=0}^{s-1} (1 - q_{t+r}) = 1 - \pi_{t-1} + \pi_{t-1} \prod_{r=0}^{s-1} (1 - F(\alpha_{t+r})) \tag{49}$$

Define H_{t+1} by

$$H_{t+1} = h(\alpha_{t+1}) + \sum_{s=1}^{\infty} \delta^s \prod_{r=1}^{s} (1 - F(\alpha_{t+r}))h(\alpha_{t+s+1}) \tag{50}$$

It can be readily checked that (50) and (9) are equivalent. Substituting (49) and (50) into (48), and replacing α_t with a, we get

$$v(a, \pi_{t-1}) = \delta G + (1-a)\ell - (1 - \pi_{t-1})\,C_{t+1} - \pi_{t-1}(1 - F(a))\delta H_{t+1} \tag{51}$$

where $C_{t+1} = \sum_{s=1}^{\infty} \delta^s h(\alpha_{t+s})$. Both C_{t+1} and H_{t+1} are functions of the optimally planned actions from period $t+1$ onward, so the envelope theorem implies (8). Since

H_{t+1} is the summation of positive terms, it is positive. Further, comparing (8) to (6)-(7), and noting that the term in (7) is (weakly) negative, we have

$$H_{t+1} \leq G - v(\alpha_{t+1}, \pi_t) \leq G - L$$

for all $t \in \mathbb{N}$. ■

Proof of Theorem 1. Let $1 < T < \infty$ be arbitrarily large, and assume that exploration stops in period $T+1$, so that $v_{T+1} = L$ and $H_{T+1} = G - L$. Let $\pi_{T-1} = (1+\varepsilon_T)\pi_{\min}$ represent the belief in period T, where $0 < \varepsilon_T < \bar{\varepsilon} := \frac{\pi_0}{\pi_{\min}} - 1$.

The method we adopt follows a form of *backward recursion*. We start by treating π_{T-1} (or ε_T) as a free variable, rather than a function of the prior belief and the history of the past allocations. Using backward induction, while ensuring the posteriors remain consistent with Bayes' rule, we will derive an optimal sequence $(\alpha_t, \pi_{t-1}|\varepsilon_T)_{t=1}^T$ parameterized by ε_T and T. In deriving this sequence, we also obtain a sequence of expected payoffs $(v_t|\varepsilon_T)_{t=1}^T$ in which $v_t|\varepsilon_T := v(\alpha_t, \pi_{t-1}|\varepsilon_T)$, such that

$$\alpha_t|\varepsilon_T \in \arg\max_{a\in[0,1]} v(a, \pi_{t-1}|\varepsilon_T)$$

and $\pi_{t-1}|\varepsilon_T$ is determined by Bayes' rule in (22), recalling $\theta = \lambda n$:

$$\pi_{t-1}(\alpha_t, \pi_t|\varepsilon_T) = \frac{\pi_t|\varepsilon_T}{\pi_t|\varepsilon_T + e^{-\theta\alpha_t|\varepsilon_T}(1 - \pi_t|\varepsilon_T)} \tag{52}$$

Next, we determine a unique ε_T such that $\pi_0|\varepsilon_T$ equals the initial belief π_0. Finally, we employ a limit argument to establish the existence and uniqueness of the optimal plan $(\alpha_t)_{t=1}^\infty$ and optimal sequence of expected payoffs $(v_t)_{t=1}^\infty$ conditional on no breakthrough.

To simplify notation, we will suppress the parameter ε_T in the following derivations unless it is needed for clarity.

Parts (i) and (iii). Consider **Scenario I**. We take four steps to prove the conclusions.

Step 1. To begin with, fix π_{T-1} and consider the problem $\max_{a\in[0,1]} v(a, \pi_{T-1})$. Given Proposition 2(i), we may assume $\alpha_T \in (0,1)$ for a sufficiently large T. By Lemma 1,

$$\frac{\partial}{\partial a} v(\alpha_T, \pi_{T-1}) = -\ell + \delta\pi_{T-1}\theta e^{-\theta\alpha_T}(G - L) = 0$$

yields

$$\alpha_T = \frac{1}{\theta}\ln\frac{\pi_{T-1}}{\pi_{\min}} = \frac{1}{\theta}\ln(1+\varepsilon_T) \tag{53}$$

where we used (10) and the fact that $G - L = (\frac{\gamma-1}{1-\delta})\ell$, and the assumption $\pi_{T-1} = (1+\varepsilon_T)\pi_{\min}$. Since $v(a, \pi_{T-1})$ is strictly concave in a, then α_T is unique. We have thus obtained the paired state (α_T, π_{T-1}), and

$$\begin{aligned} v_T &= (1-\alpha_T)\ell + \delta\pi_{T-1}\left(1 - e^{-\theta\alpha_T}\right)(G-L) + \delta L \\ H_T &= (\gamma - 1 + \alpha_T)\ell + \delta e^{-\theta\alpha_T}(G-L) \end{aligned}$$

These outputs from period T can then be used as inputs for period $T-1$. Notice that α_T and π_{T-1} are continuous and increasing functions of ε_T.

Step 2. By backward induction, suppose, hypothetically, that for $1 < t < T$, the paired states $(\alpha_{t+1}, \pi_t), (\alpha_{t+2}, \pi_{t+1}), ..., (\alpha_T, \pi_{T-1})$ have been uniquely determined, being continuous and increasing functions of ε_T, and

$$\alpha_{t+1} = \arg\max_{a\in[0,1]} v(a, \pi_t|\varepsilon_T) \in (0,1)$$

Associated with these states are the hypothetically known $v_{t+1}, ..., v_T$ and $H_{t+1}, ..., H_T$, and the initial conditions $v_{T+1} = L$ and $H_{T+1} = G - L$.

By (8), $\alpha_{t+1} \in (0,1)$ implies

$$\frac{\partial}{\partial a} v(\alpha_{t+1}, \pi_t) = -\ell + \delta\pi_t\theta e^{-\theta\alpha_{t+1}} H_{t+2} = 0$$

or equivalently,

$$H_{t+2} = \frac{\ell}{\delta\pi_t\theta e^{-\theta\alpha_{t+1}}} \tag{54}$$

Using (9), we derive

$$\frac{\partial}{\partial a} v(\alpha_t, \pi_{t-1}) = -\ell + \delta\pi_{t-1}\theta e^{-\theta\alpha_t}\left((\gamma - 1 + \alpha_{t+1}))\,\ell + \delta e^{-\theta\alpha_{t+1}} H_{t+2}\right)$$

Substituting H_{t+2} from (54), we get

$$\frac{\partial}{\partial a} v(\alpha_t, \pi_{t-1}|\varepsilon_T) = -\ell + \delta\pi_{t-1}\theta e^{-\theta\alpha_t}\left((\gamma - 1 + \alpha_{t+1})) + \frac{1}{\pi_t\theta}\right)\ell \tag{55}$$

By substituting (1) and (45), and after rearranging terms, we obtain

$$\frac{\partial}{\partial a} v(\alpha_t, \pi_{t-1}|\varepsilon_T) = \left[\delta\pi_{t-1} M(\alpha_t, \alpha_{t+1}) - (1-\delta)\right]\ell \tag{56}$$

where

$$M(\alpha_t, \alpha_{t+1}) := e^{-\theta\alpha_t}\left(1 + \theta\left(\gamma - 1 + \alpha_{t+1}\right)\right) - 1. \tag{57}$$

If $\alpha_t = 1$, then

$$0 \leq \frac{\partial}{\partial a} v(1, \pi_{t-1}|\varepsilon_T) = \left[\delta \pi_{t-1} M(1, \alpha_{t+1}) - (1-\delta)\right] \ell \leq \left[\delta \pi_{t-1} M(1,1) - (1-\delta)\right] \ell$$

which implies

$$\delta \pi_{t-1} M(1,1) = \delta \pi_{t-1} \left(e^{-\theta}(1+\theta\gamma) - 1\right) \geq 1 - \delta \tag{58}$$

However, since $\pi_{t-1} < \pi_0 < 1$, the above inequality is ruled out under **Scenario I**. Thus, we must have $\alpha_t \in (0,1)$, and by induction, this holds for all $t \leq T$. This fact, combined with (55), implies

$$\frac{\partial}{\partial a} v(\alpha_t, \pi_{t-1}) = -\ell + \delta \pi_{t-1} \theta e^{-\theta \alpha_t} \left((\gamma - 1 + \alpha_{t+1})) + \frac{1}{\pi_t \theta} \right) \ell = 0 \tag{59}$$

for all $t \leq T$. The Bayes rule (52) now gives π_{t-1}. Substituting (52) into (59), and solving for α_t, we obtain

$$\begin{aligned} \alpha_t|\varepsilon_T &= \frac{1}{\theta} \ln \left(1 + \delta\theta \left(\gamma - 1 + \alpha_{t+1}|\varepsilon_T\right) - \frac{1-\delta}{\pi_t|\varepsilon_T} \right) \\ &= \arg\max_{a \in [0,1]} v(a, \pi_{t-1}|\varepsilon_T) \end{aligned} \tag{60}$$

Using α_t and the backward induction hypothesis, we obtain also the unique values of

$$\begin{aligned} v_t &= (1-\alpha_t)\ell + \delta \pi_{t-1} \left(1 - e^{-\theta \alpha_t}\right) (G - v_{t+1}) + \delta v_{t+1} \\ H_t &= (\gamma - 1 + \alpha_t)\ell + \delta e^{-\theta \alpha_t} H_{t+1}. \end{aligned}$$

From (52) and (60), it is clear that π_{t-1} has positive partial derivatives with respect to α_t and π_t, and α_t has positive partial derivatives with respect to α_{t+1} and π_t. Thus, both α_t and π_{t-1} are continuous, increasing functions of (α_{t+1}, π_t). It follows that (α_t, π_{t-1}) are continuous and increasing in ε_T.

Step 3. From Step 2, we have thus derived a unique sequence of states $(\alpha_t, \pi_{t-1}|\varepsilon_T)_{t=1}^T$ and the associated optimal payoff sequence $(v_t|\varepsilon_T)_{t=1}^T$. The first-period paired state $(\alpha_1, \pi_0|\varepsilon_T)$ are continuous and increasing functions of ε_T. We now show that a unique $\varepsilon_T(\pi_0)$ exists such that $\pi_0|\varepsilon_T(\pi_0) = \pi_0$, consistent with the given prior.

To see this, notice that if $\varepsilon_T = \bar{\varepsilon} := \frac{\pi_0}{\pi_{\min}} - 1$, then $\pi_{T-1} = (1+\varepsilon_T)\pi_{\min} = \pi_0$. Bayes' rule implies then $\pi_0|\bar{\varepsilon} > \pi_{T-1} = \pi_0$. If $\varepsilon_T = 0$, then $\pi_{T-1} = \pi_{\min}$ and $\alpha_T = 0$, implying $\pi_{T-2} = \pi_{\min}$. In this case, by backward induction, $\alpha_t \equiv 0$ and $\pi_{t-1} \equiv \pi_{\min}$ for all $t = 1, ..., T$, which imply $\pi_0|0 = \pi_{\min} < \pi_0$.

Since $\pi_0|\varepsilon_T$ is continuous and increasing in ε_T, by the Intermediate Value Theorem there exists a unique $\varepsilon_T(\pi_0) \in (0, \bar{\varepsilon})$ such that $\pi_0|\varepsilon_T(\pi_0) = \pi_0$. Consequently,

we arrive at a unique sequence $(\alpha_t, \pi_{t-1}|\varepsilon_T(\pi_0))_{t=1}^T$ that solves the program in (3)-(4), satisfying (52) and (60) given any prior $\pi_0 \in (\pi_{\min}, 1)$ and stopping time $T+1$.

Step 4. Now consider $T \to \infty$. In light of Proposition 2(i), $\lim_{T\to\infty} \varepsilon_T(\pi_0) = 0$ so that, by continuity in $\varepsilon_T(\pi_0)$, for each $t < T$ the following limits exist:

$$\begin{aligned}
\lim_{T\to\infty} (\alpha_t, \pi_{t-1}|\varepsilon_T(\pi_0)) &= \lim_{\varepsilon_T(\pi_0)\to 0} (\alpha_t, \pi_{t-1}|\varepsilon_T(\pi_0)) =: (\alpha_t, \pi_{t-1}) \\
\lim_{T\to\infty} v(\alpha_t, \pi_{t-1}|\varepsilon_T(\pi_0)) &= \lim_{\varepsilon_T(\pi_0)\to 0} v(\alpha_t, \pi_{t-1}|\varepsilon_T(\pi_0)) =: v(\alpha_t, \pi_{t-1})
\end{aligned}$$

It is easily seen that since, fixing any t, the equations in (52) and (60) hold for all $\varepsilon_T = \varepsilon_T(\pi_0)$, then taking limit on both sides as $T \to \infty$ maintains the equations.

Finally, let $v^*(\pi_0)$ denote the optimal expected payoff, unconstrained by any exogenous deadline, in period 1 (if exists). Since $L \leq v^*(\pi_0) \leq G$, we have

$$0 \leq v^*(\pi_0) - v_1|\varepsilon_T(\pi_0) \leq \delta^{T+1}(G - L) \tag{61}$$

Thus $\lim_{T\to\infty} v_1|\varepsilon_T(\pi_0) = v^*(\pi_0)$, which gives the unconstrained optimal expected payoff. Part (iii) under Scenario I is thus also established.

Parts (ii) and (iii). Consider now **Scenario II**. Define

$$\bar{\pi} = \frac{(1-\delta)}{\delta\left(e^{-\theta}(1+\theta\gamma) - 1\right)}$$

Take any $t \geq 1$, we only need to show that $\alpha_{t+1} = 1$ implies $\alpha_t = 1$. Replacing t with $t+1$ in (56)-(58), $\alpha_{t+1} = 1$ implies $\pi_t \geq \bar{\pi}$, and therefore $\pi_{t-1} > \bar{\pi}$. Further, $\alpha_{t+1} = 1$ implies

$$\frac{\partial}{\partial a} v(1, \pi_t) = -\ell + \delta\pi_t\theta e^{-\theta} H_{t+2} \geq 0$$

or, equivalently,

$$H_{t+2} \geq \frac{\ell}{\delta\pi_t\theta e^{-\theta}}. \tag{62}$$

Similar to the derivation of (56), now using (62) we get

$$\begin{aligned}
\frac{\partial}{\partial a} v(1, \pi_{t-1}) &\geq \left[\delta\pi_{t-1} M(1,1) - (1-\delta)\right]\ell \\
&= \left[\delta\pi_{t-1}\left(e^{-\theta}(1+\theta\gamma) - 1\right) - (1-\delta)\right]\ell \\
&> 0
\end{aligned}$$

where the last inequality holds because $\pi_{t-1} > \bar{\pi}$. Thus, $\alpha_t = 1$, and by induction, $\alpha_1 = ... = \alpha_{t+1} = 1$.

If $\alpha_1 \in (0,1)$, then $\tau = 0$. If $\alpha_1 = 1$, then there exists a unique $\tau = \max\{s \in \mathbb{N} : \alpha_s = 1\}$ such that $\alpha_t \in (0,1)$ for all $t > \tau$. Part (i) of the proposition then implies that the sequence $(\alpha_t, \pi_{t-1})_{t=\tau+1}^{\infty}$ satisfies (21)-(22), which in turn establishes the conclusion in part (iii) under Scenario II. ■

Proof of Theorem 3. Parts (i)-(ii). Suppose $\left\{\left(\boldsymbol{\kappa}_t, \eta_{t-1}\right)\right\}_{t=1}^{\infty}$ is an equilibrium. Consider any period t preceded with no breakthrough, with common belief η_{t-1} given at the start of the period. We prove the conclusions through five steps.

Step 1. We first show that $\eta_{t-1} \leq \eta_{\min}$ implies $\kappa_{i,t} = 0$ for all i. This conclusion is immediate, following from $H_{i,t+1} \leq G - L$, that $\eta_{t-1} \leq \eta_{\min}$ implies

$$\begin{aligned} \frac{\partial}{\partial a} v_i(a, \kappa_{-i,t}, \eta_{t-1})|_{a=0} &= -\ell + \delta \eta_{t-1} \lambda e^{-\lambda K_{-i,t}} H_{i,t+1} \\ &\leq -\ell + \delta \eta_{\min} \lambda (G - L) = 0, \end{aligned}$$

and the fact that $v_i(a, \kappa_{-i,t}, \eta_{t-1})$ is concave in a for all i.

Step 2. Now suppose $\eta_{t-1} > \eta_{\min}$. We show all players choosing to stop in period t cannot be an equilibrium. This is because if $\kappa_{i,t} = 0$ is optimal for all i, then $K_t = 0$ and the concavity of v_i implies the following holds for all i :

$$\frac{\partial}{\partial a} v_i(a, \kappa_{-i,t}, \eta_{t-1})|_{(a,\kappa_{-i,t})=\mathbf{0}} = -\ell + \delta \eta_{t-1} \lambda H_{i,t+1} \leq 0 \tag{63}$$

If this were true, then the players would face the same problem in period $t+1$ given no new information. So the above inequality will continue to hold for all i in period $t+1$, and onwards by induction, implying $H_{i,t+1} = H_{i,t+2} ... = G - L$. But then

$$\begin{aligned} \frac{\partial}{\partial a} v_i(a, \kappa_{-i,t}, \eta_{t-1})|_{K_t=0} &= -\ell + \delta \eta_{t-1} \lambda (G - L) \\ &> -\ell + \delta \eta_{\min} \lambda (G - L) = 0 \end{aligned}$$

contradicting (63).

Step 3. As in the proof of Proposition 1, we show $\eta_{t-1} > \eta_{\min}$ implies $\eta_t > \eta_{\min}$. By contradiction, suppose $\eta_{t-1} > \eta_{\min} \geq \eta_t$. Then, Step 1 implies $K_{t+1} = 0$ and therefore $H_{i,t+1} = H_{i,t+2} ... = G - L$. Now, $K_t > 0$ and $\eta_t \leq \eta_{\min}$ imply

$$\begin{aligned} \frac{\partial}{\partial a} v_i(a, \kappa_{-i,t}, \eta_{t-1})|_{a=\kappa_{i,t}} &= -\ell + \delta \eta_{t-1} \lambda e^{-\lambda K_t} (G - L) \geq 0 \\ \frac{\partial}{\partial a} v_i(a, \boldsymbol{\kappa}_{-i,t+1}, \eta_t)|_{K_{t+1}=0} &= -\ell + \delta \eta_t \lambda (G - L) \leq -\ell + \delta \eta_{\min} \lambda (G - L) = 0 \end{aligned}$$

But then, by the Bayes rule, we derive a contradiction:

$$e^{\lambda K_t} \leq \frac{\eta_{t-1}}{\eta_t} = e^{\lambda K_t}\left(1-\eta_{t-1}\right)+\eta_{t-1} < e^{\lambda K_t}.$$

We conclude therefore that $K_{t+1} > 0$.

Step 4. As in Proposition 2, similar arguments establish that $\eta_t \to \eta_{\min}$ and $K_t \to 0$ as $t \to \infty$. Thus, for t sufficiently large, we must have $K_t < 1$, implying that no player would choose action 1, and $K_t, K_{t+1} > 0$ imply at least one player in each period, say i and j, will choose $\kappa_{i,t} > 0$ and $\kappa_{j,t+1} > 0$ respectively. We show by contradiction that it cannot be an equilibrium strategy that i chooses $\kappa_{i,t+1} = 0$ (which would imply asymmetric actions).

To see this, we have

$$\begin{aligned}\frac{\partial}{\partial a}v_i(\boldsymbol{\kappa}_t,\eta_{t-1}) &= -\ell+\delta\eta_{t-1}\lambda e^{-\lambda K_t}H_{i,t+1}\\ &= -\ell+\delta\eta_{t-1}\lambda e^{-\lambda K_t}\left(\left(\gamma-1+\kappa_{i,t+1}\right)\right)\ell+\delta e^{-\lambda K_{t+1}}H_{i,t+2}\big)\end{aligned} \quad (64)$$

and

$$\frac{\partial}{\partial a}v_i(\boldsymbol{\kappa}_{t+1},\eta_t) = -\ell+\delta\eta_t\lambda e^{-\lambda K_{t+1}}H_{i,t+2}\begin{cases}=0 & \text{if } \kappa_{i,t+1}>0\\ \leq 0 & \text{if } \kappa_{i,t+1}=0\end{cases} \quad (65)$$

Substituting (65) into (64), and rearranging terms, it can be shown as in the proof of Theorem 1, (55)-(57), that

$$\frac{\partial}{\partial a}v_i(\boldsymbol{\kappa}_t,\eta_{t-1}) = M(\kappa_{i,t+1})\begin{cases}=0 & \text{if } \kappa_{i,t+1}>0\\ \leq 0 & \text{if } \kappa_{i,t+1}=0\end{cases} \quad (66)$$

where $M(\cdot)$ is defined by

$$M(\kappa_{i,t+1}) = \left(-1+\delta\eta_{t-1}\lambda e^{-\lambda K_t}\left(\left(\gamma-1+\kappa_{i,t+1}\right)+\frac{1}{\eta_t\lambda}\right)\right)\ell, \quad (67)$$

satisfying $M' > 0$.

Now, notice that the no-breakthrough part of (25) can be written equivalently as

$$\eta_{t-1}e^{-\lambda K_t} = \frac{\eta_t}{1+\eta_t\left(e^{\lambda K_t}-1\right)}$$

Substituting into (66) and rearranging terms yield

$$\begin{aligned}M(\kappa_{i,t+1}) &= \left(-1+\delta\frac{\eta_t}{1+\eta_t\left(e^{\lambda K_t}-1\right)}\left(\lambda\left(\gamma-\left(1-\kappa_{i,t+1}\right)\right)+\frac{1}{\eta_t}\right)\right)\ell\\ &= \left(1-e^{\lambda K_t}+\delta\lambda\left(\gamma-\left(1-\kappa_{i,t+1}\right)\right)-\frac{1-\delta}{\eta_t}\right)\frac{\eta_t\ell}{1+\eta_t\left(e^{\lambda K_t}-1\right)}\end{aligned} \quad (68)$$

This deduction holds also for player j, replacing $M(\kappa_{i,t+1})$ with $M(\kappa_{j,t+1})$. Notice, though, since by assumption $\kappa_{j,t+1} > 0$, the equality part of (66) holds for j. Now, if i chooses $\kappa_{i,t+1} = 0$, we must have

$$0 = \frac{\partial}{\partial a} v_i(\boldsymbol{\kappa}_t, \eta_{t-1}) \leq M(0) \text{ and } 0 \geq \frac{\partial}{\partial a} v_j(\boldsymbol{\kappa}_t, \eta_{t-1}) = M(\kappa_{j,t+1})$$

where the first inequality derives from $\kappa_{i,t} \in (0,1)$ and (66), and the second inequality derives from $\kappa_{j,t} \in [0,1)$ and $\kappa_{j,t+1} \in (0,1)$. But these two inequalities are contradictory, because $M' > 0$ and $\kappa_{j,t+1} > 0$ imply $M(\kappa_{j,t+1}) > M(0)$.

Step 5. The above results establish that at least one player, say i, will choose $\kappa_{i,t}, \kappa_{i,t+1} \in (0,1)$ for sufficiently large t. This implies $M(\kappa_{i,t+1}) = 0$ so that we derive from (67) that

$$\kappa_{i,t+1} = \frac{1}{\delta\lambda}\left(e^{\lambda K_t} + \frac{1-\delta}{\eta_t} - 1\right) + 1 - \gamma$$

Since the right-hand side is independent of i, and our pick of i is arbitrary, we conclude that $\kappa_{t+1} := \kappa_{i,t+1}$ is the optimal allocation strategy for all $i = 1, ..., n$. In other words, the equilibrium is necessarily symmetric.

Part (iii). Once we can focus on a symmetric equilibrium, the structure of the problem reduces to the basic model. The same arguments as in the proof of Theorem 1 establish the conclusion.

Part (iv). The conclusion can be proved by taking limit as $\eta_t \to 1$ and $\kappa_t \to \hat{\kappa}$ in (68), or by a similar analysis as in the proof of Proposition 1(i). ■

References

Aghion, P., Bolton, P., Harris, C., and Jullien, B. (1991). Optimal learning by experimentation. *The Review of Economic Studies*, 58(4):621–654.

Athey, S. and Segal, I. (2013). An efficient dynamic mechanism. *Econometrica*, 81(6):2463–2485.

Awaya, Y. and Krishna, V. (2021). Startups and upstarts: disadvantageous information in R&D. *Journal of Political Economy*, 129(2):534–569.

Bergemann, D. and Hege, U. (1998). Venture capital financing, moral hazard, and learning. *Journal of Banking & Finance*, 22(6-8):703–735.

Bergemann, D. and Hege, U. (2005). The financing of innovation: Learning and stopping. *The RAND Journal of Economics*, 36(4):719–752.

Bergemann, D. and Välimäki, J. (2008). Bandit problems. In Durlauf, S. and Blume, L., editors, *The New Palgrave Dictionary of Economics*, pages 336–340. Macmillan Press.

Berry, D. A. and Fristedt, B. (1985). *Bandit problems: Sequential Allocation of Experiments*. Springer Dordrecht.

Besanko, D. and Wu, J. (2013). The impact of market structure and learning on the tradeoff between R&D competition and cooperation. *The Journal of Industrial Economics*, 61(1):166–201.

Blackwell, D. (1965). Discounted dynamic programming. *The Annals of Mathematical Statistics*, 36(1):226–235.

Bobtcheff, C. and Levy, R. (2017). More haste, less speed? Signaling through investment timing. *American Economic Journal: Microeconomics*, 9(3):148–86.

Bolton, P. and Harris, C. (1999). Strategic experimentation. *Econometrica*, 67(2):349–374.

Bonatti, A. and Hörner, J. (2011). Collaborating. *American Economic Review*, 101(2):632–663.

Bonatti, A. and Hörner, J. (2017). Career concerns with exponential learning. *Theoretical Economics*, 12(1):425–475.

Chikte, S. D. (1980). Optimal sequential selection and resource allocation under uncertainty. *Advances in Applied Probability*, 12(4):942–957.

Choi, J. P. (1997). Herd behavior, the "penguin effect," and the suppression of informational diffusion: An analysis of informational externalities and payoff interdependency. *The RAND Journal of Economics*, 28(3):407–425.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177.

Gittins, J. C. and Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In Vincze, I., Gani, J., and Sarkadi, K., editors, *Progress in Statistics*, pages 241–266. North-Holland Pub. Co.

Guo, Y. (2016). Dynamic delegation of experimentation. *American Economic Review*, 106(8):1969–2008.

Halac, M., Kartik, N., and Liu, Q. (2016). Optimal contracts for experimentation. *The Review of Economic Studies*, 83(3):1040–1091.

Halac, M., Kartik, N., and Liu, Q. (2017). Contests for experimentation. *Journal of Political Economy*, 125(5):1523–1569.

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

Holmstrom, B. (1982). Moral hazard in teams. *The Bell Journal of Economics*, 13(2):324–340.

Hörner, J., Klein, N., and Rady, S. (2022). Overcoming free-riding in bandit games. *The Review of Economic Studies*, 89(4):1948–1992.

Hörner, J. and Skrzypacz, A. (2017). Learning, experimentation and information design. *Advances in Economics and Econometrics*, 1:63–98.

Karlin, S. (1955). The structure of dynamic programing models. *Naval Research Logistics Quarterly*, 2(4):285–294.

Keller, G. and Rady, S. (2010). Strategic experimentation with poisson bandits. *Theoretical Economics*, 5(2):275–311.

Keller, G., Rady, S., and Cripps, M. (2005). Strategic experimentation with exponential bandits. *Econometrica*, 73(1):39–68.

Malueg, D. A. and Tsutsui, S. O. (1997). Dynamic R&D competition with learning. *The RAND Journal of Economics*, 28(4):751–772.

Murto, P. and Välimäki, J. (2011). Learning and information aggregation in an exit game. *The Review of Economic Studies*, 78(4):1426–1461.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition.

Rosenberg, D., Solan, E., and Vieille, N. (2007). Social learning in one-arm bandit problems. *Econometrica*, 75(6):1591–1611.

Rothschild, M. (1974). A two-armed bandit theory of market pricing. *Journal of Economic Theory*, 9(2):185–202.

Sadler, E. (2021). Dead ends. *Journal of Economic Theory*, 191:105–167.

Stokey, N. L., Lucas, R. E., and Prescott, E. C. (1989). *Recursive Methods in Economic Dynamics*. Harvard University Press.

Thomas, C. D. (2021). Strategic experimentation with congestion. *American Economic Journal: Microeconomics*, 13(1):1–82.