Two-Level Bandits for Collaborator Selection and Sustained Cooperation

Anonymous Author(s)
Affiliation
email

Abstract

Agentic systems increasingly need to select other agents for delegation, collaboration, and sustained workflows. We study this problem as a two-level bandit: an outer agent selects among uncertain collaborators, while an inner repeated game determines the value of cooperation with the selected agent. We formalize this interaction, develop algorithms for joint collaborator selection, and characterize when cooperation can be sustained. We show that selecting agents based only on advertised capability or compatibility can systematically favor high-capability defectors, while accounting for cooperation dynamics identifies the true best collaborator. We also derive an optimal reservation policy for discounted commitment and validate the framework through simulation. The results connect bandit learning, optimal stopping, and cooperation, providing a framework for collaborator selection in agent marketplaces, delegation systems, and federated AI.

1 Introduction

As autonomous agents are deployed into open ecosystems, a new class of decision appears: an agent must select *another agent* to enter a sustained cooperative relationship with. A planning agent delegates a multi-step task to a third-party tool agent; a federated-learning coordinator admits a client that must keep contributing honest updates; an orchestrator contracts a downstream service for a long-running workflow; agents form a coalition to solve a task neither can complete alone. In each case the decision has the structure of a *sequential commitment*: the agent evaluates candidates of uncertain quality under a bounded interaction budget, and, having committed, must then sustain a cooperative relationship whose value unfolds over many subsequent rounds.

The prevailing engineering practice is to rank candidates by an advertised or statically measured *compatibility* score—a capability profile, a benchmark result, a declared SLA, an aggregated reputation. This paper shows that this practice is not merely approximate but systematically wrong, and that it gets worse with more data.

Two mature theories each capture one half of the problem. The *secretary problem* and its bandit generalizations [Ferguson, 1989, Gilbert and Mosteller, 1966, Robbins, 1952, Lai and Robbins, 1985] describe how to search and when to stop under uncertainty and a deadline. The *iterated Prisoner's Dilemma* (IPD) and the folk theorem [Axelrod and Hamilton, 1981, Fudenberg and Maskin, 1986] describe when self-interested parties will cooperate rather than defect once matched. Yet these are almost always studied in isolation, leaving a gap precisely where the two interact: *which agent is worth choosing depends on the equilibrium the pair will settle into, and how long one can afford to search depends on how much that cooperation is worth.*

We close this gap with a **two-level bandit** model. At the outer level, an agent faces candidate collaborators as arms of a best-arm-identification (BAI) bandit; at the inner level, each engaged collaborator induces a repeated PD whose discounted equilibrium payoff is that arm's reward. The coupling is the crux: the outer reward is not drawn from an exogenous distribution but is the *solution*

Preprint.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

of an inner game. This single change—reward-as-equilibrium— is what makes the problem more than the sum of its parts, and it drives every result below.

Contributions.

- **Model.** A formal two-level bandit coupling outer BAI with an inner discounted IPD, with a clean value decomposition $V(a) = \mu(a) + \mathbb{E}[\bar{u} \mid a]$ (compatibility plus cooperation value) (§3).
- **Algorithm.** A two-level pull oracle that spends a full relationship rollout per outer pull, turning the coupled value V into a directly estimable bandit reward; instantiated as fixed-budget (TwoLevel-UCB, TwoLevel-SE) and fixed-confidence (TwoLevel-LUCB) best-arm identification (§3.1, Algs. 1–3).
- **Theory.** (i) An inner cooperation-sustainability threshold $\gamma^\star = (T-R)/(T-P)$; (ii) a *value-decomposition regret* result showing a compatibility-only (level-1-only) searcher is suboptimal by exactly the cooperation gap, which can be arbitrarily large; (iii) completion of the previously open "Section 4" commitment problem as a discounted optimal-stopping problem, with a monotone reservation-threshold optimal policy (§4).
- **Empirics.** A reproducible suite (six experiments) that validates each claim: an evolutionary characterization of cooperation, a (γ, ϵ) phase diagram matching $\gamma^\star$, two-level BAI vs. compatibility-only search, secretary reservation curves, and the completed commitment-stopping problem (§5).
- **Prior art and novelty.** A differentiation of our coupled formulation from the secretary, bandit, matching-market, IPD, multi-agent orchestration, and trust/reputation literatures (§2).

2 Related Work and Prior Art

Optimal stopping and the secretary problem. The classical secretary problem—maximize the probability of selecting the best of N sequentially revealed candidates with no recall—has the famous $1/e$ solution [Lindley, 1961, Gilbert and Mosteller, 1966, Ferguson, 1989], with value-based and multiple-choice extensions [Chow et al., 1971, Kleinberg, 2005]. These works treat the value of a chosen candidate as *given*; none model a strategic maintenance phase after stopping.

Bandits and best-arm identification. Multi-armed bandits [Robbins, 1952, Lai and Robbins, 1985, Gittins, 1979, Bubeck and Cesa-Bianchi, 2012, Slivkins, 2019] and the pure-exploration / BAI setting in particular—UCB [Auer et al., 2002], successive elimination [Even-Dar et al., 2006], and LUCB [Kalyanakrishnan et al., 2012]—provide our outer-level machinery. In all of these the per-pull reward is an exogenous random variable. In our setting the reward is endogenous: it is the discounted payoff of an inner repeated game, so the "arm distribution" is shaped by strategic equilibrium.

Repeated games and the evolution of cooperation. The IPD, Tit-for-Tat, Win-Stay-Lose-Shift, generous strategies [Axelrod and Hamilton, 1981, Axelrod, 1984, Nowak and Sigmund, 1992, 1993], the folk theorem [Fudenberg and Maskin, 1986], and zero-determinant strategies [Press and Dyson, 2012] characterize *within*-relationship dynamics, and multi-agent RL studies sequential social dilemmas [Leibo et al., 2017]. This literature takes the pairing as given and asks whether cooperation is sustained; it does not model how the partner was *selected* from a population under search costs.

Matching markets. Stable-matching theory [Gale and Shapley, 1962] studies who ends up with whom at a market-clearing equilibrium, typically with known, static preferences and no repeated-game maintenance. We instead study a single agent's sequential search and the post-match cooperative dynamics.

Novelty and differentiation. Our primitives (secretary/BAI and IPD) are individually classical—so, to answer the natural reviewer question directly: yes, the building blocks are known, and the contribution is *not* a new bandit algorithm or a new PD strategy. The novelty is in the *coupling* and its consequences, which do not appear in any of the strands above:

1. **Reward-as-equilibrium.** The outer BAI reward of an arm is the discounted equilibrium payoff of an inner repeated game, not an exogenous draw. This makes arm values depend on the discount factor and on execution noise through $\gamma^\star$, coupling stopping theory to game theory in a way neither field addresses alone.
2. **A provable failure of standard practice.** We show (Prop. 2) that ranking candidates by compatibility—exactly what capability-profile matching, benchmark-score routing, and reputation-scored agent registries do—is suboptimal by the cooperation gap, and we measure that gap empirically (1.60; the highest-compatibility candidate is the worst long-run choice). Worse, the failure is *anti-scaling*: correct-selection probability falls to zero as the budget grows (Exp. 3).
3. **Completing an open problem.** We give a well-posed formulation and full solution of the "α-scaled commitment" stopping problem that the originating work stated but left undefined, and show the classical $1/e$ rule is far from optimal once delay is discounted.

In short, the pieces are old; the object they compose into—a bandit whose arm rewards are game equilibria, together with the resulting stopping policy—is new, and it produces non-obvious, testable predictions.

3 The Two-Level Bandit Model

Outer level: partner selection. Let $A = \{a_1, \ldots, a_N\}$ be candidate partners (arms). Engaging a_i yields a noisy *compatibility* signal with mean $\mu_i = \mathbb{E}[R_i]$. Under a deadline the agent must identify $\arg\max_i \mu_i$ from as few interactions as possible—the fixed-confidence BAI problem [Even-Dar et al., 2006, Kalyanakrishnan et al., 2012].

Inner level: relationship maintenance. Once engaged, partner a_i induces a repeated game with actions $S = \{\mathrm{C}, \mathrm{D}\}$ and the PD stage payoffs

	C	D
C	(R, R)	(S, T)
D	(T, S)	(P, P)

$$T > R > P > S, \quad 2R > T + S.$$

Over rounds $t = 1, 2, \ldots$ each side plays $s_t \in S$ and the agent accrues the discounted payoff $U = \sum_{t \geq 1} \gamma^{t-1} u(s_t)$ with $\gamma \in (0, 1)$. Execution is noisy: each intended action is flipped independently with probability ϵ (a trembling hand). We write $\mathbb{E}[U \mid a_i]$ for the agent's expected discounted payoff when it plays its maintenance strategy π (we use Tit-for-Tat) against partner i.

Over a finite engagement of m rounds it is convenient to work with the *per-round normalized* payoff

$$\bar{u} \;=\; U \Big/ \textstyle\sum_{t=1}^{m} \gamma^{t-1}, \qquad \bar{u} \in [S, T] \;\text{ for every } m, \gamma, \tag{1}$$

which is the discount-weighted average stage payoff. We use U itself when reasoning about equilibrium sustainability (Prop. 1, where the threshold is a statement about discounted sums) and $\bar{u}$ whenever a cooperation value must be added to a compatibility signal, so that the two levels are expressed on a common per-interaction scale.

Coupling. The value of a partnership is the sum of the two levels,

$$V(a_i) \;=\; \underbrace{\mu_i}_{\text{compatibility}} \;+\; \underbrace{\mathbb{E}[\bar{u} \mid a_i]}_{\text{cooperation value}}, \qquad a^\star = \arg\max_{a_i \in A} V(a_i). \tag{2}$$

The agent runs a BAI loop at the outer level in which every "pull" interleaves a compatibility observation with an inner-game rollout, estimating V rather than μ. Because $\mathbb{E}[\bar{u} \mid a_i]$ is the payoff of an equilibrium that exists only when $\gamma \geq \gamma^\star$ (Prop. 1), the outer reward landscape is itself shaped by the inner game.

Note that the choice of scale in Eq. (2) is substantive rather than cosmetic: normalizing only the second summand fixes the exchange rate between a unit of compatibility and a unit of cooperation, and a different normalization would in general induce a different $a^\star$. We adopt the per-round convention throughout, so all reported values of V are on the stage-payoff scale $[S, T]$ plus compatibility.

Algorithm 1 TWOLEVELPULL: one outer pull of candidate partner a_i

Require: partner a_i with compatibility mean μ_i and unknown inner strategy; agent maintenance strategy π (we use Tit-for-Tat); inner horizon m; discount γ; execution noise ϵ; compatibility noise σ_c
Ensure: one unbiased sample of $V(a_i)$ from Eq. (2)

$\pi_{\text{self}} \leftarrow \text{FRESH}(\pi)$, $\pi_{\text{ptnr}} \leftarrow \text{FRESH}(a_i)$ ▷ cold start: no history carried across pulls
$U \leftarrow 0$, $Z \leftarrow 0$, $h_{\text{self}} \leftarrow \emptyset$, $h_{\text{ptnr}} \leftarrow \emptyset$
for $t = 1$ **to** m **do**
  $s_t \leftarrow \pi_{\text{self}}(h_{\text{self}}, h_{\text{ptnr}})$, $s'_t \leftarrow \pi_{\text{ptnr}}(h_{\text{ptnr}}, h_{\text{self}})$
  flip s_t with prob. ϵ; flip s'_t with prob. ϵ ▷ trembling hand, independent per side
  $U \leftarrow U + \gamma^{t-1} u(s_t, s'_t)$, $Z \leftarrow Z + \gamma^{t-1}$
  append s_t to h_{self}, s'_t to h_{ptnr}
end for
$\bar{u} \leftarrow U/Z$ ▷ per-round normalized; $\bar{u} \in [S, T]$ for any m, γ
$c \leftarrow \mu_i + \eta$ with $\eta \sim \mathcal{N}(0, \sigma_c^2)$ ▷ noisy compatibility signal
return $c + \bar{u}$

3.1 The proposed algorithm

Eq. (2) defines the estimand; Algs. 1–3 define how we estimate it. The single new ingredient is the *pull oracle* TWOLEVELPULL (Alg. 1): one outer pull spends an entire m-round relationship rollout, so the outer bandit observes an unbiased draw of the coupled value $V(a_i)$ rather than of the compatibility mean μ_i alone. Given that oracle, any off-the-shelf BAI policy applies unchanged; we report UCB1 and successive elimination in the fixed-budget regime (Alg. 2) and LUCB in the fixed-confidence regime (Alg. 3). The *level-1-only* baseline throughout is the same policy driven by the cheaper oracle that returns $\mu_i + \eta$ and skips the rollout entirely.

Three design choices are load-bearing and we state them explicitly.

(i) Cold-start rollouts. Each pull instantiates a *fresh* copy of both the agent's maintenance strategy π and the partner's strategy, so inner history does *not* persist across pulls of the same arm. Successive pulls of arm i are therefore i.i.d., which is what licenses the standard concentration radii used below. Substantively, this models repeatedly getting to know someone from scratch rather than resuming an existing relationship; carrying history over would induce within-arm dependence and require a bandit-with-state analysis we do not attempt here.

(ii) Discount-normalized inner reward. A pull returns the per-round normalized payoff $\bar{u} = U / \sum_{t=1}^{m} \gamma^{t-1}$ rather than the raw discounted sum U. This keeps $\bar{u} \in [S, T]$ regardless of m and γ, so V is comparable across horizons and discount factors and the confidence radii need no rescaling when either is varied.

(iii) Budget units. One unit of budget is one *engagement* — m PD rounds plus one compatibility observation — not one PD round. The level-1-only baseline is charged the same unit for a strictly cheaper observation, so the accounting if anything favours the baseline; the gap we report in Sec. 5.3 is thus a conservative estimate of the cost of ignoring the inner level.

Because $\bar{u} \in [S, T]$ and μ_i is bounded, V has range $[\, S,\ \max_i \mu_i + T \,]$ rather than $[0, 1]$; the LUCB radius β carries $v_{\max}$ explicitly for this reason.[1]

4 Theory

4.1 When does the inner relationship cooperate?

Proposition 1 (Cooperation-sustainability threshold)**.** *Against a grim-trigger partner, mutual cooperation is a subgame-perfect equilibrium of the discounted IPD if and only if*

$$\gamma \ \geq \ \gamma^{\star} \ = \ \frac{T - R}{T - P}.$$

[1]The successive-elimination radius r_i in Alg. 2 is stated in its usual unit-range form, as implemented. Scaling it by $v_{\max}$ would make the two policies' radii consistent; since successive elimination is used only in the fixed-budget regime, where the radius controls the elimination schedule and not a correctness guarantee, this affects how aggressively arms are pruned but not the validity of the reported regret.

Algorithm 2 Fixed-budget two-level BAI (TWOLEVEL-UCB and TWOLEVEL-SE)

Require: arms $A = \{a_1, \dots, a_N\}$; budget B engagements; exploration constant c_{ucb}; inner parameters $(\pi, m, \gamma, \epsilon, \sigma_c)$
Ensure: recommended partner $\hat{a}$

procedure UPDATE(i)
  $x \leftarrow \text{TWOLEVELPULL}(a_i)$, $n_i \leftarrow n_i + 1$, $\hat{V}_i \leftarrow \hat{V}_i + (x - \hat{V}_i)/n_i$, $t \leftarrow t + 1$
end procedure

$\hat{V}_i \leftarrow \text{TWOLEVELPULL}(a_i)$, $n_i \leftarrow 1$ for all i; $t \leftarrow N$ ▷ initialize with one pull per arm
Variant A: UCB1.
while $t < B$ **do**
  $i \leftarrow \arg\max_j \left[\hat{V}_j + \sqrt{c_{\text{ucb}} \log \max(t, 2)/n_j}\right]$
  UPDATE(i)
end while
Variant B: successive elimination.
$\mathcal{A} \leftarrow \{1, \dots, N\}$
while $t < B$ **and** $|\mathcal{A}| > 1$ **do**
  for all $i \in \mathcal{A}$ **do**
    UPDATE(i) **, break if** $t \geq B$ ▷ one round-robin sweep of the survivors
  end for
  $r_i \leftarrow \sqrt{\log\big(4Nn_i^2 + 1\big)/(2n_i)}$ for $i \in \mathcal{A}$
  $b \leftarrow \arg\max_{i \in \mathcal{A}} \hat{V}_i$
  $\mathcal{A} \leftarrow \{\, i \in \mathcal{A} \,:\, \hat{V}_i + r_i \;\geq\; \hat{V}_b - r_b \,\} \cup \{b\}$ ▷ drop arms whose UCB falls below the leader's LCB
end while
while $t < B$ **do** UPDATE($\arg\max_{i \in \mathcal{A}} \hat{V}_i$) ▷ spend leftover budget on the leader

return $\hat{a} \leftarrow a_j$ with $j = \arg\max_i \hat{V}_i$

Algorithm 3 Fixed-confidence two-level BAI (TWOLEVEL-LUCB)

Require: arms A; confidence parameter δ; value range $v_{\max} \geq \sup_i V(a_i)$; inner parameters $(\pi, m, \gamma, \epsilon, \sigma_c)$
Ensure: $\hat{a}$ with $\Pr[\hat{a} = a^\star] \geq 1 - \delta$, and the number of engagements consumed

$\hat{V}_i \leftarrow \text{TWOLEVELPULL}(a_i)$, $n_i \leftarrow 1$ for all i; $t \leftarrow N$
$\beta(n) \leftarrow v_{\max}\sqrt{\log\big(4Nn^2/\delta\big)/(2n)}$ ▷ anytime Hoeffding radius
loop
  $b \leftarrow \arg\max_i \hat{V}_i$ ▷ empirical leader
  $j^\dagger \leftarrow \arg\max_{i \neq b} \left[\hat{V}_i + \beta(n_i)\right]$ ▷ most optimistic challenger
  if $\hat{V}_b - \beta(n_b) \;\geq\; \hat{V}_{j^\dagger} + \beta(n_{j^\dagger})$ **then**
    break ▷ leader separated at level $1 - \delta$
  end if
  for all $i \in \{b, j^\dagger\}$ **do**
    $x \leftarrow \text{TWOLEVELPULL}(a_i)$, $n_i \leftarrow n_i + 1$, $\hat{V}_i \leftarrow \hat{V}_i + (x - \hat{V}_i)/n_i$, $t \leftarrow t + 1$
  end for
end loop
return $\hat{a} \leftarrow a_j$ with $j = \arg\max_i \hat{V}_i$; t

For the classical payoffs $(T, R, P, S) = (5, 3, 1, 0)$, $\gamma^\star = 0.5$.

Proof sketch. Cooperating forever yields $R/(1-\gamma)$. The most profitable one-shot deviation earns T now and triggers permanent punishment worth $\gamma P/(1-\gamma)$. Cooperation is sustainable iff $R/(1-\gamma) \geq T + \gamma P/(1-\gamma)$, i.e. $R \geq (1-\gamma)T + \gamma P$, which rearranges to $\gamma \geq (T-R)/(T-P)$. □

Thus a patient relationship (γ large) supports cooperation; an impatient one does not. Execution noise $\epsilon > 0$ shrinks the cooperative region because accidental defections trigger costly retaliation (Exp. 2).

4.2 Why compatibility-only search is suboptimal

Let $i^\star_\mu = \arg\max_i \mu_i$ be the best arm by compatibility and $i^\star = \arg\max_i V(a_i)$ the best by total value.

Proposition 2 (Value-decomposition regret)**.** *Any policy that identifies the compatibility-optimal partner $i^\star_\mu$ (the level-1-only objective) incurs simple regret in true value*

$$\Delta \;=\; V(a_{i^\star}) - V(a_{i^\star_\mu}) \;=\; \big(\mu_{i^\star} + \mathbb{E}[U \mid a_{i^\star}]\big) - \big(\mu_{i^\star_\mu} + \mathbb{E}[U \mid a_{i^\star_\mu}]\big) \;\geq\; 0,$$

and, on the raw discounted scale*, Δ can be made arbitrarily large: for any $M > 0$ there is an instance with $\Delta > M$, obtained by pairing a high-μ defector ($\mathbb{E}[U] \to P/(1-\gamma)$) against a slightly-lower-μ cooperator ($\mathbb{E}[U] \to R/(1-\gamma)$). On the per-round scale of Eq.* (1) *the same construction gives $\Delta \to \max_i \mu_i - \max_{j \neq i} \mu_j + (R - P)$, i.e. the gap is bounded by $R - P$ but does not vanish for any $\gamma \geq \gamma^\star$.*

Proof sketch. $\Delta \geq 0$ by optimality of $i^\star$ for V. For the lower bound, place all compatibility mass on a defector and note the inner gap $\mathbb{E}[U \mid \text{coop}] - \mathbb{E}[U \mid \text{defect}] \to (R - P)/(1 - \gamma)$, which diverges as $\gamma \to 1$; choosing μ close between the two arms leaves Δ dominated by this inner gap. □

This formalizes a practical failure mode: systems that rank partners by a compatibility score alone (as capability-matching routers and reputation-scored agent registries do) provably lose the entire cooperation gap. Experiment 3 exhibits an instance with $\Delta = 1.60$ in which the *most* compatible candidate is the *worst* long-run choice; that figure is the per-round quantity, against a theoretical ceiling of $R - P = 2$ for these payoffs, so the constructed instance realizes 80% of the worst case attainable at this scale.

4.3 Completing the commitment problem ("Section 4")

The originating work introduced, but did not define, a model in which candidates of unknown quality $\alpha_i \in \mathbb{N}$ are encountered in a fixed order and play is scaled by α until mutual cooperation (C, C) is reached. We give it a well-posed form. Reaching (C, C) is a *commitment* that terminates search and pays a terminal reward $R\,\alpha_i$; each additional candidate evaluated discounts the eventual reward by $\beta \in (0, 1)$, capturing the cost of delay—in an agentic deployment, the latency, token spend, and forgone throughput incurred while the orchestrator is still shopping rather than executing. The objective is

$$\max_{\tau} \; \mathbb{E}[\beta^\tau \, R\, \alpha_\tau]\,, \tag{3}$$

a finite-horizon discounted optimal-stopping problem with no recall.

Proposition 3 (Monotone reservation threshold)**.** *The optimal policy for* (3) *commits at position i iff $\alpha_i \geq \theta_i := c_{i+1}/R$, where $c_{i+1} = \beta\,\mathbb{E}[V_{i+1}]$ is the continuation value obtained by backward induction $V_i(\alpha) = \max\{R\alpha,\ c_{i+1}\}$, $c_N = 0$. The thresholds are non-increasing in i: standards relax as the deadline approaches.*

Proof sketch. Optimality of a reservation rule is immediate from the Bellman recursion for (3). Monotonicity follows because $\mathbb{E}[V_{i+1}]$ is non-increasing in i (fewer remaining draws cannot raise the continuation value), so $c_{i+1} = \beta\mathbb{E}[V_{i+1}]$ and hence θ_i are non-increasing. □

A striking consequence, confirmed in Exp. 5, is that the classical $1/e$ rule— optimal for maximizing $\mathbb{P}(\text{best rank})$—is badly suboptimal here, because it waits for the top rank while the discount β^τ annihilates the reward of a late commitment.

5 Experiments

All experiments are fully reproducible from a single seed; payoffs are $(T, R, P, S) = (5, 3, 1, 0)$ so $\gamma^\star = 0.5$. Code, results (`JSON`), and figures accompany the paper.

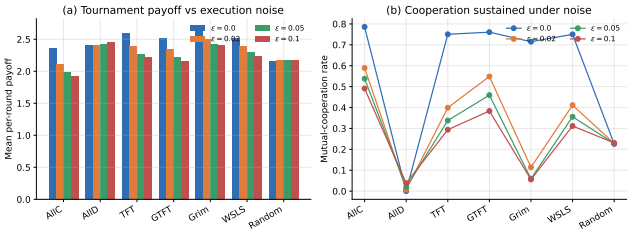


Figure 1: **Inner IPD tournament.** (a) Mean per-round payoff by strategy across execution noise. (b) Mutual-cooperation rate. The defector wins raw payoff by exploiting naive types while destroying cooperation, motivating outer-level screening.

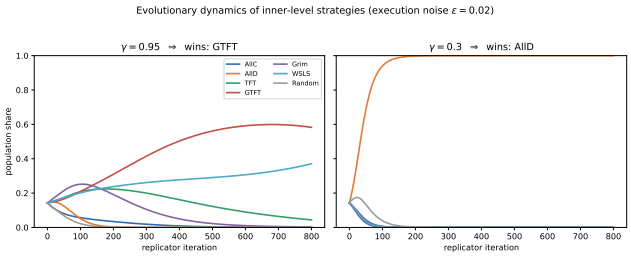


Figure 2: **Evolutionary dynamics.** Replicator trajectories of inner strategies. Left ($\gamma = 0.95 > \gamma^\star$): reciprocal cooperation dominates and defection dies out. Right ($\gamma = 0.30 < \gamma^\star$): defection takes over. The transition matches the analytic threshold $\gamma^\star = 0.5$.

5.1 Experiment 1: inner strategy tournament and evolutionary dynamics

We run a round-robin among AllC, AllD, TFT, Generous-TFT (g=0.3), Grim, Win-Stay-Lose-Shift (WSLS), and Random, with $\gamma = 0.95$, 200 rounds, 300 repetitions, over execution noise $\epsilon \in \{0, 0.02, 0.05, 0.10\}$ (Fig. 1). Averaged over *all* opponents, the pure defector tops raw payoff at ϵ=0.05 (2.43 vs. Grim 2.42, WSLS 2.29), because it exploits unconditional and random partners—its mutual-cooperation rate is only 0.018 against GTFT's 0.459. This is not a contradiction but a *motivation*: defection pays against naive partners, so the outer level must screen them out.

The canonical way to show cooperation is nonetheless favored under repetition is population dynamics. Building the pairwise discounted-payoff matrix and running the replicator dynamics (Fig. 2) gives a sharp dichotomy: at $\gamma = 0.95$ Generous-TFT dominates and the defector's share collapses to 0.000, and a rare defector *cannot* invade a TFT population (fitness gap -1.36); at $\gamma = 0.30 < \gamma^\star$ the defector sweeps the population (1.000) and *can* invade TFT (fitness gap $+0.78$). Cooperation is evolutionarily stable exactly in the patient regime predicted by Prop. 1.

5.2 Experiment 2: cooperative-equilibrium phase diagram

For a grid over (γ, ϵ) we let a boundedly rational agent adopt the best-response policy (from {AllD, AllC, TFT, GTFT, WSLS}) against a TFT partner, and record the realized mutual-cooperation rate (Fig. 3). The empirical cooperation onset at zero noise is $\gamma \approx 0.55$, essentially the analytic $\gamma^\star = 0.50$ (the small gap is finite-horizon truncation of the discounted sum). Increasing execution noise ϵ steadily shrinks the cooperative region, quantifying how fragile trust is to accidental defection.

5.3 Experiment 3: two-level BAI vs. compatibility-only search

This experiment instantiates Algs. 1–3 with $N = 8$ arms, inner horizon $m = 100$, $\gamma = 0.95$, $\epsilon = 0.05$, compatibility noise $\sigma_c = 0.5$, $c_{\text{ucb}} = 2$, and—for the fixed-confidence run—$\delta = 0.05$ and $v_{\max} = 6$. The level-1-only baseline replaces TwoLevelPull with the compatibility-only oracle $\mu_i + \eta$. We construct eight candidate partners whose compatibility μ is *anti-aligned* with their cooperativeness (Table 1): the most compatible partner, P0 (μ=1.9), is a pure defector. Ground-truth per-round cooperation values $\mathbb{E}[U]$ (from 4000 rollouts each) yield total values $V = \mu + \mathbb{E}[U]$ whose maximizer

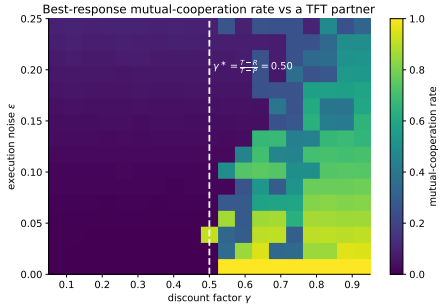


Figure 3: **Phase diagram of cooperation.** Best-response mutual-cooperation rate over (γ, ϵ). The dashed line is the analytic threshold $\gamma^\star = (T-R)/(T-P) = 0.5$; cooperation emerges to its right and erodes as noise grows.

Table 1: Experiment 3 instance. Compatibility μ is anti-aligned with cooperation value $\mathbb{E}[U]$. Level-1-only search selects P0 (highest μ); the true best by V is P1. $\Delta = V_{\text{P1}} - V_{\text{P0}} = 1.60$.

Partner	Inner type	μ	$\mathbb{E}[U]$ (per round)	$V = \mu + \mathbb{E}[U]$
P0	AllD (high-compat trap)	1.9	1.06	2.96
P1	**Generous-TFT**	**1.7**	**2.85**	**4.55**
P2	TFT	1.5	2.52	4.02
P3	Random(0.5)	1.3	2.22	3.52
P4	Generous-TFT(0.1)	1.1	2.68	3.78
P5	AllC	0.9	3.03	3.93
P6	AllD	0.7	1.05	1.75
P7	Random(0.8)	0.5	2.75	3.25

is P1 (Generous-TFT, V=4.55), while the compatibility maximizer is the defector P0 (V=2.96). The value-decomposition regret of a compatibility-only searcher is therefore $\Delta = 1.60$ (Prop. 2), and it *persists no matter how much data it collects*. In contrast, two-level UCB and successive elimination drive simple regret in V to zero as the interaction budget grows, and correct-partner identification approaches 1 (Fig. 4). Fixed-confidence two-level LUCB identifies the true best partner with empirical probability 1.00 at δ=0.05, using on average $\approx 3.6 \times 10^4$ interactions.

5.4 Experiment 4: finite-horizon search under a hard budget

We treat the two-level value V of randomly drawn candidates as the secretary's observation and compare the known-distribution optimal reservation policy (Prop. 3 specialized to undiscounted values), the classical $1/e$ rule, and random choice (Fig. 5). At N=40 the optimal policy captures expected value 4.82 (of a maximum ≈ 5) versus 4.16 for the $1/e$ rule and 3.00 at random ($\mathbb{E}[V]$=3.0). The $1/e$ rule's probability of selecting the very best falls from 0.415 to 0.371 as N grows to 80, tracking the theoretical $1/e \approx 0.368$. The optimal reservation value *decreases* monotonically with the number of candidates already evaluated—4.83 at the start of the budget down to 3.00 at exhaustion—giving a concrete admission-control schedule for a deployed selector: hold a high bar while budget remains, and relax it as the budget is consumed.

5.5 Experiment 5: the completed commitment-stopping problem

We simulate the discounted commitment problem (3) with N=30 candidates of integer quality $\alpha \in \{1, \ldots, 10\}$ and search discount β=0.9 (Fig. 6). The backward-induction optimal policy achieves expected discounted reward 22.17, and a *single* stationary quality threshold $\theta^\star$=7 nearly matches it at 22.21—a practical, memoryless rule. "Commit immediately" earns 16.50. The classical $1/e$ rule earns only 2.22: by holding out for the best rank it commits late, and the discount β^τ destroys the reward. The optimal reservation quality relaxes from 6.65 at the first candidate to 4.95 near the deadline, mirroring Prop. 3. The design lesson is sharp: *when delay is costly, optimizing the probability of finding the single best collaborator is the wrong objective; a moderate, stationary commitment threshold is both near-optimal and trivial to implement*—a one-line admission rule in an orchestrator, rather than a search that must track ranks.

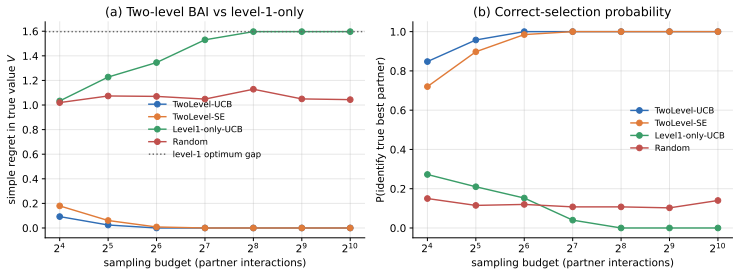


Figure 4: **Two-level best-arm identification.** (a) Simple regret in true value V vs. interaction budget. Two-level policies converge to 0; the compatibility-only searcher plateaus at the cooperation gap $\Delta = 1.60$ (dotted). (b) Probability of identifying the true best partner.

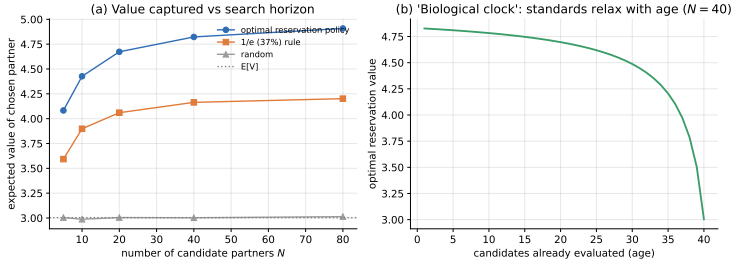


Figure 5: **Search under a deadline.** (a) Expected value captured vs. horizon N for the optimal reservation policy, the $1/e$ rule, and random. (b) The optimal reservation value relaxes as the evaluation budget is consumed (N=40), yielding an admission-control schedule.

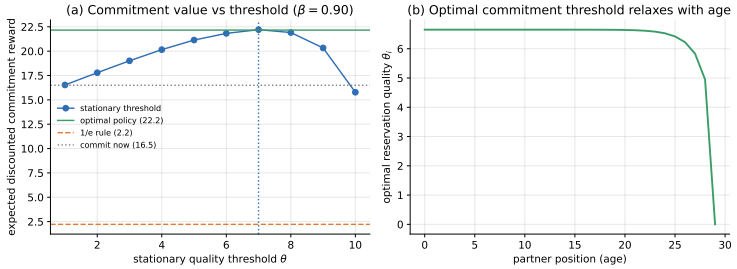


Figure 6: **Commitment as discounted optimal stopping (completed "Section 4").** (a) Expected discounted reward vs. stationary threshold θ; the optimal policy (solid) is nearly matched by $\theta^\star$=7, while the $1/e$ rule (dashed) collapses once delay is discounted. (b) The optimal reservation quality relaxes as the budget is consumed.

6 Conclusion

We recast collaborator selection in agentic systems as a two-level bandit in which outer arm rewards are the equilibria of an inner iterated Prisoner's Dilemma, gave a pull oracle and three selection algorithms for it, proved a cooperation threshold, a value-decomposition regret bound, and the optimality of a monotone reservation policy for a completed commitment-stopping problem, and validated all three empirically. The practical conclusion for systems that pair autonomous agents is uncomfortable: the signal these systems currently rank on—advertised capability or aggregated reputation—is not a noisy proxy for what they want, but a systematically misleading one, and collecting more of it makes the selection worse rather than better. Recovering the right choice requires spending real interaction to observe cooperation, and stopping on a confidence criterion rather than a score. More broadly, *search and cooperation must be reasoned about jointly*: which agent to choose depends on the game the pair will play, and how long one can afford to search depends on how much that game is worth.

References

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

Robert Axelrod. *The Evolution of Cooperation*. Basic Books, 1984.

Robert Axelrod and William D. Hamilton. The evolution of cooperation. *Science*, 211(4489): 1390–1396, 1981.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Yuan Shih Chow, Herbert Robbins, and David Siegmund. *Great Expectations: The Theory of Optimal Stopping*. Houghton Mifflin, 1971.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.

Thomas S. Ferguson. Who solved the secretary problem? *Statistical Science*, 4(3):282–289, 1989.

Drew Fudenberg and Eric Maskin. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3):533–554, 1986.

David Gale and Lloyd S. Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

John P. Gilbert and Frederick Mosteller. Recognizing the maximum of a sequence. *Journal of the American Statistical Association*, 61(313):35–73, 1966.

John C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 655–662, 2012.

Robert Kleinberg. A multiple-choice secretary algorithm with applications to online auctions. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 630–631, 2005.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 464–473, 2017.

Dennis V. Lindley. Dynamic programming and decision theory. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 10(1):39–51, 1961.

Martin A. Nowak and Karl Sigmund. Tit for tat in heterogeneous populations. *Nature*, 355(6357): 250–253, 1992.

Martin A. Nowak and Karl Sigmund. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, 364(6432):56–58, 1993.

William H. Press and Freeman J. Dyson. Iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent. *Proceedings of the National Academy of Sciences*, 109(26):10409–10413, 2012.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends in Machine Learning*, 12(1–2):1–286, 2019.