State Sensitivity in an Additive Discovery Game

Yifan Hong (hongyf23@mails.tsinghua.edu.cn)
Department of Industrial Engineering
Tsinghua University

Chen Wang (chenwang@tsinghua.edu.cn)
Department of Industrial Engineering
Tsinghua University

Bonan Zhao (bzhao2@ed.ac.uk)
School of Informatics
University of Edinburgh

Abstract

Successful innovation hinges on balancing exploring new ideas and exploiting existing ones. A rational innovator should be state-sensitive, effectively switching to exploitation when the best available idea reaches some standard. We tackle innovation with a discovery-by-recombination game under additive reward growth, and compare the optimal state-dependent policy with a state-independent policy. Our experiment reveals that participants made state-dependent decisions, exploring more in rounds with early successes, albeit being told of the same true success probability. In contrast, the optimal state-dependent policy switches to exploitation earlier. This suggests that participants' state-sensitivity may be driven by ad-hoc subjective probabilities. Participants also deviated from optimality through excessive exploration, switching multiple times between exploration and exploitation, and their switching points also differed from the theoretical optimum.

Introduction

Innovation has long been a driving force behind the development of human civilization. Ancient people combined natural objects to create tools. In modern science, new ideas often stem from recombining existing knowledge (Uzzi et al., 2013). A central question in this process is how a rational agent should balance exploring new ideas with exploiting established ones. This challenge is particularly relevant for scientists choosing research topics (Huang et al., 2022). Excessive exploration may waste opportunities to reap rewards, while premature exploitation risks suboptimal outcomes.

In this work, we study innovation strategies through recombination within the discovery game framework (Zhao et al., 2024). In a discovery game, an agent sequentially chooses between fusing items (exploration) and extracting rewards (exploitation). Zhao et al. (2024) analyzed the game under exponential reward growth, where the optimal policy is independent of the best available item. However, exponential growth fails to capture the constrained gains typical in competitive innovation environments (Chu & Evans, 2021). We instead adopt additive reward growth, leading to an optimal policy that switches from exploration to exploitation based on the current best item. We compare this state-dependent policy with a state-independent strategy, where a predetermined switching point guides actions regardless of history.

We investigate whether humans are state-sensitive in the discovery game with an online behavioral experiment. We used a forced-choice paradigm with fusion outcome manipulation to control the initial state of participants' decisions. To foreshadow, we found that participants made state-dependent choices, exploring more in luckier rounds—both in their first choice and throughout the game. This state sensitivity contrasts with the optimal strategy, suggesting that ad-hoc updates of subjective probabilities may influence people.

Background

Exploration-exploitation trade-off

Humans constantly face the challenge of balancing between exploring unknowns and exploiting the known rewards (Cohen et al., 2007; Mehlhorn et al., 2015; Sutton & Barto, 1998). Exploration may benefit on the long-term reward, while exploitation emphasizesx short-term returns. Both excessive exploration and premature exploitation can result in suboptimal results. A famous example is the optimal stopping problem (Ferguson, 2006), where one sequentially accepts or rejects the options. Rejected items cannot be revisited while accepting terminates the process.

Optimal policies for the explore-exploit tasks seek a balance between the opposing sides. Previous work shows that people could adopt a threshold-based strategy that depends on the time left and the best options available, as does the optimal policy (Baumann et al., 2020; Sang et al., 2020). However, people often adopt a threshold lower than the optimal one (Baumann et al., 2020). They often switch between exploration and exploitation multiple times, and exhibit sequence-level decision noise (Song et al., 2019). The deviation from optimality is attributed to various factors, including risk aversion (Bhatia et al., 2021) and the representation of the problem (Baumann et al., 2025).

Innovation by recombination

Innovation rarely springs from pure originality (Bramley et al., 2023). Instead, new ideas often emerge through combining existing knowledge (Xiao et al., 2022), which in turn becomes the foundation for future breakthroughs. The game of discovery by recombination formalizes this process as a finite-horizon sequential decision process (Zhao et al., 2024), where players choose between extracting immediate rewards from items or fusing items to create potentially more valuable ones. Initially, all items yield a base reward. Each successful fusion increases the level of the best item and its reward. Unlike traditional multi-armed bandit models (Reverdy et al.,

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

2013), which learn the values of items, the discovery game captures how innovation expands the space of ideas through recombination. The action space of the discovery game can grow infinitely, while both the value of items and the chance of a successful fusion are explicitly known.

The reward growth assumptions critically influence the optimal policy. Zhao et al. (2024) assumed exponential reward growth, where the item reward scales exponentially with its level. In this case, both the marginal value of fusion and its opportunity cost are linear in the reward, rendering the optimal policy dependent solely on the time left. As we show in the following section, this property does not hold under additive reward growth. When each successful fusion adds a constant to the reward, the marginal value of exploration is constant, while the opportunity cost grows with the item level, making the optimal policy depend on the highest item level. In the next section, we analyze the optimal state-dependent policy and compare it with its state-independent counterpart in more details.

The Additive Discovery Game

Following Zhao et al. (2024), we formalize "ideas" as items and "innovation" as an attempt to fuse existing items for new ones. Players can also extract immediate rewards from an existing item. A successful fusion produces a new, more rewarding item, which players can exploit in future rounds by extracting rewards from it.

Markov decision process (MDP) model

The discovery game is modeled as a Markov decision process (MDP) $G = (T, M, A, \mathbf{P}, \mathbf{R})$, where $T \in \mathbb{N}_+$ is the horizon, M is the set of possible items, A is the set of actions, $\mathbf{P}$ is the transition function, and $\mathbf{R}$ is the reward function. In each round $t \in [T]$, player observes the available items $M_t \subseteq M$, selects an action $a_t \in A$, receives a reward $r_t = \mathbf{R}(s_t, a_t)$, and transitions to the next state according to $\mathbf{P}(s_{t+1}|s_t, a_t)$. Below, we detail the key components of the MDP.

State At time period $t \in [T]$, the state $s_t \in S \subset \mathbb{N}$ tracks the highest item level, which indicates the number of successful fusions leading to that item. Initial items start at level 0. This simplification of M_t relies on the fact that, for a rational player, it is sufficient to focus only on the most rewarding items when making decisions.

Action The action space $A = \{\texttt{extract}, \texttt{fuse}\}$ is also simplified by removing strictly dominated actions. At time period t, the agent at state s_t can either `extract` from the best item in M_t (of level s_t) or `fuse` it with another item.

Transition When the player selects a_t, the system transitions from s_t to s_{t+1} according to $\mathbf{P}$. The highest-level item remains unchanged unless the player chooses `fuse` and the fusion succeeds. A successful fusion increases the level of the best item by one. We assume the probability of successful fusion is homogeneous: `fuse` succeeds with probability p, $p \in (0,1)$. Formally,

$$\begin{aligned} \mathbf{P}(s_{t+1} = s_t | s_t, a_t = \texttt{extract}) &= 1 \\ \mathbf{P}(s_{t+1} = s_t | s_t, a_t = \texttt{fuse}) &= 1 - p \\ \mathbf{P}(s_{t+1} = s_t + 1 | s_t, a_t = \texttt{fuse}) &= p \end{aligned} \tag{1}$$

Reward The reward function $\mathbf{R}$ describes the instant reward of each action at each state. `Fuse` leads to 0 instant reward. `Extract` collects a state-related reward. Each initial item of level-0 yields a base reward r. Each successful fusion increases the item reward by a constant amount $\tilde{r} \in \mathbb{R}$ from the best original items. One can `extract` $r_k = r + k\tilde{r}$ from a level-k item. For simplicity, we let the constant $\tilde{r}$ equal the base reward r. This assumption does not affect the structure of optimal policy. Formally,

$$\begin{aligned} \mathbf{R}(s_t, \texttt{extract}) &= r(s_t) \triangleq (s_t + 1) \cdot r \\ \mathbf{R}(s_t, \texttt{fuse}) &= 0 \end{aligned} \tag{2}$$

Optimal state-dependent policy

Given the formal specification of MDP, an optimal policy $\pi^* : [T] \times S \to A$ can be derived through backward induction. Bellman optimality equation suggests

$$q^*(s_t, a_t) = \mathbf{R}(s_t, a_t) + \sum_{s_{t+1} \in S} \mathbf{P}(s_t, a_t, s_{t+1}) \max_{a_{t+1} \in A} q^*(s_{t+1}, a_{t+1}) \tag{3}$$

At the last period T, $q^*(s_T, \texttt{extract}) = r(s_T)$, $q^*(s_T, \texttt{fuse}) = 0$. Working backward gives the optimal state-action q-value function (see fig. 1 for an example). At each time period t, the optimal policy follows the action with the highest q-value

$$\pi^*(s_t) = \arg\max_{a \in A} q^*(s_t, a). \tag{4}$$

The optimal policy has a special structure of "switching once", as stated in theorem 1. It never returns to exploration once it starts exploitation. Moreover, the timing of this switch is not fixed—it depends on the current best item.

Theorem 1. *The optimal state-dependent policy switches at most once from exploration (`fuse`) to exploitation (`extract`). The switching point is state-dependent.*

$$\pi^*(s_t) = \begin{cases} \texttt{fuse}, & t \leqslant T - \frac{s_t + 1}{p}. \\ \texttt{extract}, & \textit{otherwise}. \end{cases} \tag{5}$$

Proof sketch. Here, we provide the intuition for the proof. A complete proof is presented in the appendix. We first prove the "switch-once" property by construction. A policy generates a sequence of actions for any transition outcomes. In expectation, any action sequence that switches back from exploitation to exploration is dominated by one that swaps the later exploration with exploitation. Thus, the optimal policy must satisfy the "switch-once" property.

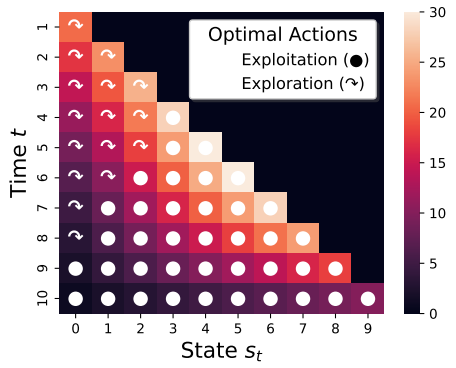

Figure 1: Value function and the optimal state-dependent policy for $p = 0.6$, $T = 10$. The switching point from exploration to exploitation depends linearly on the state. In this example, the optimal state-independent $\tau^* = 4$.

Given the "switch-once" property, the switching point can be derived by comparing exploration and exploitation at each time period. It is better to explore if the value of exploration exceeds its opportunity cost

$$[p(s_t+2)+(1-p)(s_t+1)]r\cdot(T-t) \geqslant (s_t+1)r\cdot(T-t+1)$$
$$\Leftrightarrow t \leqslant T - \frac{(s_t+1)}{p}$$

Optimal state-independent policy

Evaluating the opportunity cost for every step can be computationally extensive, and thus decision-makers could alternatively follow a pre-determined plan throughout the game. Given a game G, the state-independent policy $\pi_\tau : [T] \rightarrow A, \tau = 0, 1, ..., T$ explores for τ periods, and then switches to exploitation.

$$\pi_\tau(s_t) = \begin{cases} \texttt{fuse}, & t \leqslant \tau \\ \texttt{extract}, & \tau < t \leqslant T \end{cases} \tag{6}$$

The optimal state-independent policy chooses τ^* to maximize

$$V_\tau \triangleq \mathbb{E}_\tau \left[\sum_{t=1}^{T} \mathbf{R}(s_t, a_t) \right] = (T-\tau)\mathbb{E}[(1+s_{\tau+1})] \tag{7}$$

where $s_{\tau+1} \sim B(\tau, p)$ is a Binomial random variable whose distribution depends on τ. Optimality is reached at

$$\tau^* = \frac{Tp-1}{2p} = \frac{T}{2} - \frac{1}{2p}. \tag{8}$$

Equation 8 suggests that the optimal switching point is just before $\frac{T}{2}$, the middle of the horizon. This simple rule can be applied without tracking the current best item, and still reaches optimality if played in many independent games.

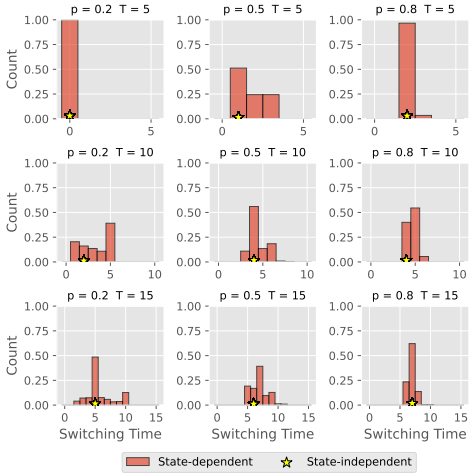

Figure 2: Switching time period of optimal state-dependent and state-independent policy.

Policy comparison

To illustrate the influence of state-sensitivity, we run the discovery game for $N^S = 1000$ independent rounds with game parameters $p = \{0.2, 0.5, 0.8\}$ and $T = \{5, 10, 15\}$. The switching point of the optimal state-dependent policy is distributed around the state-independent one (fig. 2). It explores more and switches later when p is lower or the horizon T is longer. The difference in rewards collected by the two policies is not significant. The optimal state-dependent policy yields a higher expected reward but also greater variance when p is lower (see appendix, table 2).

Most distinctively, the state-dependent policy extends exploration when the best item is not rewarding enough and terminates earlier otherwise. In contrast, the state-independent policy averages over these scenarios, resulting in switching points that are approximately at the mean of the state-dependent policy's switching points.

Empirical Evaluations

We investigate whether people make similar state-sensitive decisions in an online behavioral experiment. The experiment is pre-registered at https://aspredicted.org/w72s-rc2f.pdf.

Experiment

To focus on cases where the state-dependent and state-independent optimal policies make different predictions, we used a forced-choice paradigm to manipulate the state at which participants start to make autonomous decisions. We designed state-time pairs (t_0, s_{t_0}) where t_0 is at the state-independent optimal switching point τ^*, but s_{t_0} can either reach the state-dependent switching threshold or not yet. If people are state-sensitive, they should choose to exploit at t_1 when s_{t_0} exceeds the threshold, and keep exploring if s_{t_0} has not yet reached the threshold. A state-independent agent, however, would always switch to exploit at t_1 regardless of the value of s_{t_0}.

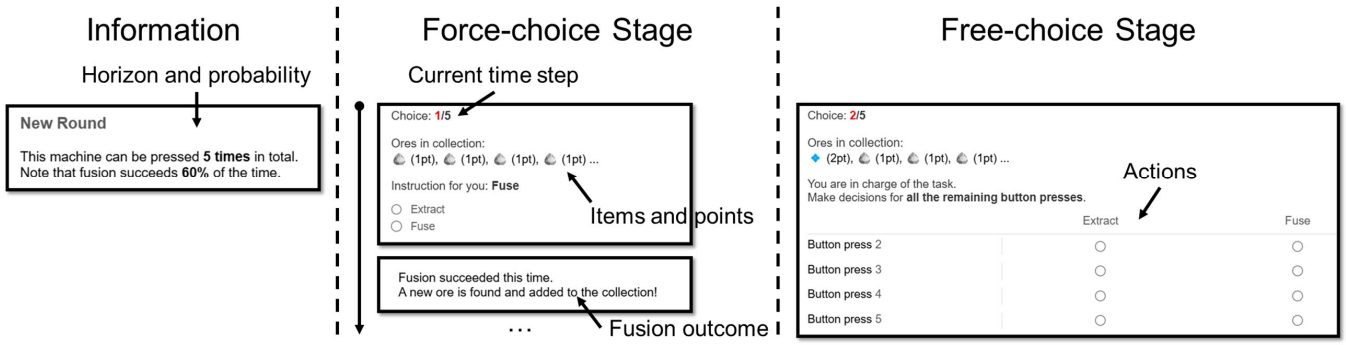

Figure 3: Experiment interface implemented in a survey system. The screenshots display one round of discovery game in the experiment. Arrows and bold text are for illustration only and not shown to participants. Each participant experiences 4 rounds with forced-choices and free-choices ("lucky" or "unlucky" type), and another 4 rounds with free-choices only ("free" type).

Participants 200 participants are recruited through Prolific Academic (86 females, $M_{\text{age}} = 35 \pm 12$). The median task completion time was 16 minutes. The sample size was determined by a power analysis aiming to obtain .95 power to detect an effect size of .3 at .05 alpha level. As preregistered, thirty-one participants were excluded because they chose `fuse` for all the trials within a round. Participants were paid both for their time and a performance-based bonus. The experiment was approved by the Tsinghua University Science and Technology Ethics Committee (project no. THU-04-2025-003). All participants gave informed consent before undertaking the experiment.

Procedure After reading the instructions and passing a comprehension quiz, each participant completed eight rounds of independent discovery games (fig. 3), choosing between `fuse` and `extract`. A successful fusion builds upon the best item, increasing its value by one unit. Extraction does not change the item values. The eight rounds were arranged in two blocks, and each block had four rounds of the same game type (see the **Design** section). Each round was divided into two stages, a forced-choice stage ($t = 1, ..., t_0 - 1$) and a free-choice stage ($t = t_0, ..., T$). During the forced-choice stage, participants were instructed to select `fuse`, and—unknown to the participants—the outcomes were predetermined by the experimenters. Participants began making autonomous decisions at t_0, marking the start of the free-choice stage. The participants are asked to make choices for all subsequent actions at t_0, allowing us to observe both their choices at the onset of the free-choice stage and their plans for the subsequent actions. We compare the chosen actions at t_0 with the state-dependent optimal policy. Participants were informed of the horizon T and the success probability p before each round. After the blocks, we collected participants' subjective evaluation of the success probability, together with demographic information and feedback.

Design To test state-sensitivity, we designed three game types: lucky, unlucky, and free (table 1). In the lucky games, participants observed relatively more success outcomes in the forced-choice stage, and therefore started their free-choice stage with more valuable items. The unlucky games, on the contradictory, gave more null outcomes during forced-choice, and left the participants relatively low-value items when the free-choice stage began. Free games have no forced-choice stage, and participants made their own decisions for the entire round.

Horizon T	5	10	15	20
"Lucky" type game	(2, 1)	(4, 3)	(7, 5)	(9, 6)
"Unlucky" type game	(3, 0)	(5, 1)	(8, 2)	(10, 4)
"Free" type game	(0, 0)	(0, 0)	(0, 0)	(0, 0)

Table 1: Time-state pairs (t_0, s_{t_0}) selected for "lucky", "unlucky" and "free" type games of different horizons.

Each participant was randomly assigned to one of the four between-subject conditions: free-lucky, free-unlucky, lucky-free, and unlucky-free. Here, each game type represents a block of four rounds of games of that type, with varying $T = 5, 10, 15, 20$. The order of horizons was randomized within each game type block to control for potential order effects. To mitigate the potential effect of successive failures, two transition sequences are selected for the "unlucky" states in games with horizon $T = 10, 15, 20$. For all conditions, the fusion success probability is held constant $p = 0.6$, and explicitly communicated to the participants throughout.

Results

Participants' decisions were state-dependent, but in the opposite direction of optimality. Participants demonstrated sensitivity to the initial state at the start of the free-choice stage. Game type predicted the proportion of exploration (`fuse`) in the free-choice stage (game type: $F(2, 2019) = 333.81$, $p < .001$, Cohen's $f = 0.571$). However, contrary to the predictions of the optimal state-dependent policy, participants actually explored more in the "lucky" games than in the "unlucky" games (fig. 4a). Fo-

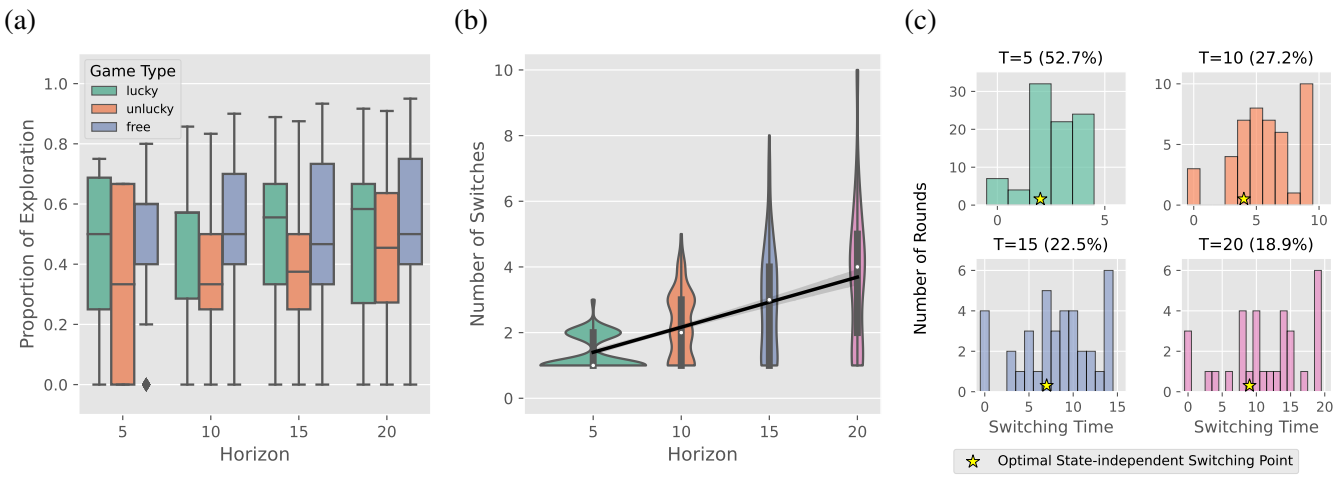

Figure 4: Participants' decisions for the discovery games. (a) Boxplot of the proportion of exploration (`fuse`) in participants' action sequences across different types of games with different horizons. (b) Number of switches from exploration to exploitation in "free" type games. The black regression line illustrates the increasing trend with the horizon. (c) The distribution of participants' switching points in "free" type games. The percentage numbers following the horizon indicate the proportion of participants that switched once for the "free" type game. The yellow star indicates the theoretical optimal switching point.

cusing on the first action after the forced-choice stage, game type also influenced the proportion of participants who chose to explore (game type: $F(1,672) = 3.782$, $p = .052$, Cohen's $f = 0.0748$). More participants chose `fuse` as the first action in "lucky" than in "unlucky" type games, especially when the horizon is short.

State-dependency might result from subjective probability Although participants were explicitly informed $p = 0.6$, they may internally update their beliefs based on the observations. We collected probability estimations after the "lucky" or "unlucky" type games in a debrief question for explorative purposes. The preceding game type significantly influenced the estimated success probability p (game type: $F(1,167) = 76.701$, $p < .001$, Cohen's $f = 0.677$). The average estimation was below 0.6 for the "unlucky" type games (mean: 0.455; $t(82) = 7.822$, $p < .001$, Cohen's $d = 0.859$) while above 0.6 for "lucky" type games (mean: 0.657; $t(85) = 4.101$, $p < .001$, Cohen's $d = 0.442$). The updated probability estimate may subsequently influence participants' choices. For participants in the "lucky-free" and "unlucky-free" conditions, the probability estimation predicted the proportion of exploration in the following "free" type games (estimation: $F(1,381) = 9.014$, $p = .003$, Cohen's $f = 0.153$). The proportion of `fuse` increases with the estimated $\hat{p}$ (estimation: $\beta_{\hat{p}} = 0.209, t(381) = 3.002$, $p = .003$).

Game order influences free-choice strategies Game order also significantly affects the choice patterns in "free" type games (game order: $F(3,668) = 7.407$, $p < .001$, Cohen's $f = 0.182$). Participants explored more in the "lucky-free" order than in the "free-lucky" order (Tukey's HSD: $p = .004$) and more in the "unlucky-free" order than in the "free-unlucky" order (Tukey's HSD: $p = .050$). This effect cannot be explained by subjective probability, because the "unlucky" rounds led to lower perceived success probability, and thus would lead to lower rate of exploration. In contrast, participants who first encountered these "unlucky" games actually explored more in subsequent "free" games, compared to those who started with "free" games right away.

Participants' plans deviate from optimality Participants exhibit excessive exploration in both "lucky" ($t(343) = 22.817$, $p < .001$, Cohen's $d = 1.230$) and "unlucky" ($t(331) = 14.580$, $p < .001$, Cohen's $d = 0.800$) games, where the optimal policies suggest exploitation since the second free choice. They also over-explored in "free" games ($t(899) = 14.256$, $p < .001$, Cohen's $d = 0.475$). The game horizon T influenced the proportion of exploration (horizon: $F(1,2019) = 8.303$, $p = .003$, Cohen's $f = 0.055$). Participants explored slightly more in games with longer horizons. (fig. 4a, $\beta_T = 0.0034, t(2019) = 1.937$, $p = .053$).

When focusing on the "free" games, participants can switch more than once, and the number of switches grows as the horizon increases (fig. 4b). For participants who switched once and exploited until the end, their switching points are distributed across the horizon (fig. 4c). The multi-modal pattern of the distribution implies notable heterogeneity in participants' strategies. Besides those who switched near the theoretical optimum, some chose to `fuse` until the second-to-last action and `extract` once, while others may `extract` throughout the entire game.

To evaluate participants' decisions, we sampled 1000 sequences of transitions and calculated the cumulative rewards.

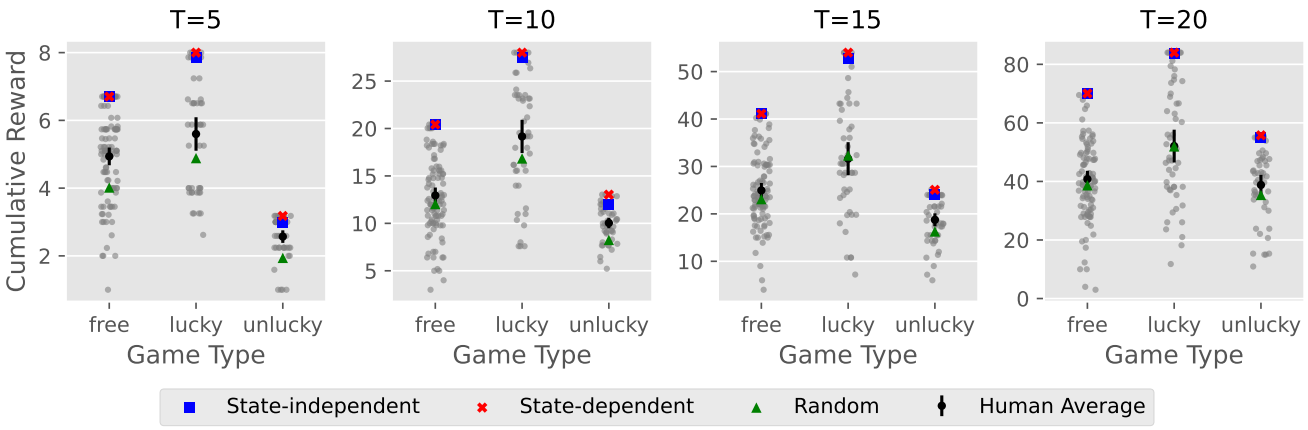

Figure 5: The expected cumulative reward over 1000 simulated outcome sequences. Each grey point shows the mean reward collected by a participant. The black point shows the human average performance with 95% confidence interval. The blue square and the red X indicate the anticipated reward collected if the participant had adhered to the optimal policies.

We compared the participants' plans against random choice baseline (50% `fuse`, 50% `extract`) and the optimal policies (fig. 5). Most participants outperform the random baseline ($t(1335) = 4.897$, $p < .001$, Cohen's $d = 0.134$) but underperform the optimal state-dependent policy ($t(1335) = 32.320$, $p < .001$, Cohen's $d = 0.884$) and the optimal state-independent policy ($t(1335) = 31.478$, $p < .001$, Cohen's $d = 0.861$).

Discussion

Successful innovation requires balancing exploring a vast combinatorial space of ideas with exploiting the established ones. The current study investigates innovation strategies theoretically and empirically using the additive discovery games framework. A rational innovator would exploit more in luckier rounds once an idea meets the time-dependent threshold. We find that participants indeed made state-sensitive decisions, but explored more after successful outcomes, which could be attributed to subjective probability updates. Participants' strategies also deviated from the theoretical optimum with excessive exploration, multiple switches between exploration and exploitation, and divergent switching points.

The current work has several limitations that future work could address. The experiment interface collects decisions for multiple steps in an one-off fashion. In future work, we aim to develop a more flexible and natural paradigm with iterative feedback to better assess decision rationality. Currently, the optimal policies use the informed p as ground truth. Unexpected outcomes are treated as incidental. In contrast, participants updated their subjective beliefs about p. This strategy can be more robust in real-world environments without perfect information. As a future direction, we may incorporate belief updates into the model, modeling how people learn about a latent p through the progress of the game (Lai & Robbins, 1985; Lattimore & Szepesvári, 2020).

Another future direction is to explore how other reward growth assumptions shape the optimal and human strategies. For example, the reward after successful fusions could be the sum of the rewards of the original items. Taking a step further, we may introduce item features to parameterize the success probability **P** and the reward **R**. This setup can capture the difficulty of navigating the combinatorial space of candidate ideas, which may ultimately limit the growth of knowledge (Weitzman, 1998).

In the current task, deriving the optimal policies requires integrating all possible outcome sequences, which is computationally intractable. Instead, people may resort to resource-rational thinking (Lieder & Griffiths, 2020). For example, participants might aim at satisficing instead of maximizing (Caplin et al., 2011). Notably, despite the efforts to derive them, both optimal policies are simple threshold-based decision rules. Instead of doing calculations, a decision-maker with limited cognitive resources might learn these decision rules through trial and error. Future work could investigate the resource-rational approaches for the game, as well as the role of learning throughout this process.

Acknowledgments

This work was supported by a grant from the National Natural Science Foundation of China (NSFC 72192824).

References

Baumann, C., Schlegelmilch, R., & von Helversen, B. (2025). Beyond risk preferences in sequential decision-making: How probability representation, sequential structure and choice perseverance bias optimal search. *Cognition*, *254*, 106001.

Baumann, C., Singmann, H., Gershman, S. J., & von Helversen, B. (2020). A linear threshold model for optimal stopping behavior. *Proceedings of the National Academy of Sciences*, *117*(23), 12750–12755.

Bhatia, S., He, L., Zhao, W. J., & Analytis, P. P. (2021). Cognitive models of optimal sequential search with recall. *Cognition*, *210*, 104595.

Bramley, N. R., Zhao, B., Quillien, T., & Lucas, C. G. (2023). Local search and the evolution of world models. *Topics in Cognitive Science*.

Caplin, A., Dean, M., & Martin, D. (2011). Search and satisficing. *American Economic Review*, *101*(7), 2899–2922.

Chu, J. S., & Evans, J. A. (2021). Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences*, *118*(41), e2021636118.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 933–942.

Ferguson, T. S. (2006). *Optimal stopping and applications.* UCLA, Los Angeles, CA, USA.

Huang, S., Lu, W., Bu, Y., & Huang, Y. (2022). Revisiting the exploration-exploitation behavior of scholars' research topic selection: Evidence from a large-scale bibliographic database. *Information Processing & Management*, *59*(6), 103110.

Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, *6*(1), 4–22.

Lattimore, T., & Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, *43*, e1.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., ... Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*(3), 191.

Reverdy, P., Srivastava, V., & Leonard, N. E. (2013). Modeling human decision-making in multi-armed bandits. In *Multidisciplinary conf. on reinforcement learning and decision making, princeton, nj, usa.*

Sang, K., Todd, P. M., Goldstone, R. L., & Hills, T. T. (2020). Simple threshold rules solve explore/exploit trade-offs in a resource accumulation search task. *Cognitive science*, *44*(2), e12817.

Song, M., Bnaya, Z., & Ma, W. J. (2019). Sources of suboptimality in a minimalistic explore–exploit task. *Nature human behaviour*, *3*(4), 361–368.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to Reinforcement Learning*. MIT Press.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*(6157), 468–472.

Weitzman, M. L. (1998). Recombinant growth. *The Quarterly Journal of Economics*, *113*(2), 331–360.

Xiao, T., Makhija, M., & Karim, S. (2022). A knowledge recombination perspective of innovation: review and new research directions. *Journal of Management*, *48*(6), 1724–1777.

Zhao, B., Vélez, N., & Griffiths, T. (2024). A rational model of innovation by recombination. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Appendix

Proof for theorem 1

Proof. We first prove the "switch-once" property by construction. Now consider a given but unknown binary sequence of fusion outcomes $\omega = \{o_j\}_{j=1}^{T} \in \Omega$, where o_j stands for the outcome of the j-th fusion. For a policy $\pi : S \to A$, $a_t = \pi(s_t)$. If $a_t = \texttt{fuse}$, and it is the j_t-th `fuse` until time t, $\sum_{\tau=1}^{t} \mathbb{I}[a_\tau = \texttt{fuse}] = j$, the transition is determined by

$$s_{t+1} = s_t + o_{j_t}.$$

For any $\omega \in \Omega$, π uniquely determines a sequence of actions $\{a_t^{\pi,\omega}\}_{t=1}^{T}$. If there exists $t_1 < t_2$, such that $a_{t_1}^{\pi,\omega} = \texttt{extract}$, $a_{t_2}^{\pi,\omega} = \texttt{fuse}$, we can construct another action sequence with

$$a_t^{\pi',\omega} = \begin{cases} \texttt{fuse}, & \text{if } t = t_1, \\ \texttt{extract} & \text{if } t = t_2, \\ a_t^{\pi,\omega}, & \text{otherwise.} \end{cases} \tag{9}$$

Comparing the two action sequences, only actions at t_1 and t_2 are swapped. For any outcome sequence ω, the differences between the rewards collected by the two action sequences can only appear from t_1 to t_2. Since $s_{t_1} + o_{j_{t_1}} \geqslant s_{t_1}$, and other parts of the sequences are identical, we can conclude that the latter one collects equal or more reward. And since this holds for any $\omega \in \Omega$, taking an expectation over ω renders the conclusion: **for any action sequence, there exists an alternative sequence that switches at most once and achieves equal or greater reward**.

Now, the remaining issue is the existence of such π', mapping each $\omega \in \Omega$ to an action sequence that exhibits the "switch-once" property. This is resolved by construction. Assuming the optimal π^* satisfies the "switch-once" policy, we can derive it by greedily comparing `fuse` and `extract`. The resulting π^* is presented in eq. (5)

□

Simulation results

The simulation results for policy comparison are presented here in table 2.

Experiment design details

We include the selected initial time-state pairs and the corresponding transition sequences used in our experiment in fig. 6.

Table 2: Cumulative reward (mean $\pm$ std. dev.) collected by the optimal state-dependent and the state-independent policy.

	$T = 5$		$T = 10$		$T = 15$	
	state-dependent	state-independent	state-dependent	state-independent	state-dependent	state-independent
$p = 0.2$	5.00$\pm$0.00	5.00$\pm$0.00	11.47$\pm$5.14	10.98$\pm$4.34	20.73$\pm$9.01	19.94$\pm$8.80
$p = 0.5$	6.32$\pm$2.06	6.05$\pm$2.00	18.11$\pm$5.91	17.71$\pm$5.92	37.35$\pm$10.70	36.22$\pm$11.06
$p = 0.8$	7.87$\pm$1.59	7.84$\pm$1.66	25.53$\pm$4.52	25.24$\pm$4.74	53.24$\pm$8.12	53.02$\pm$8.30

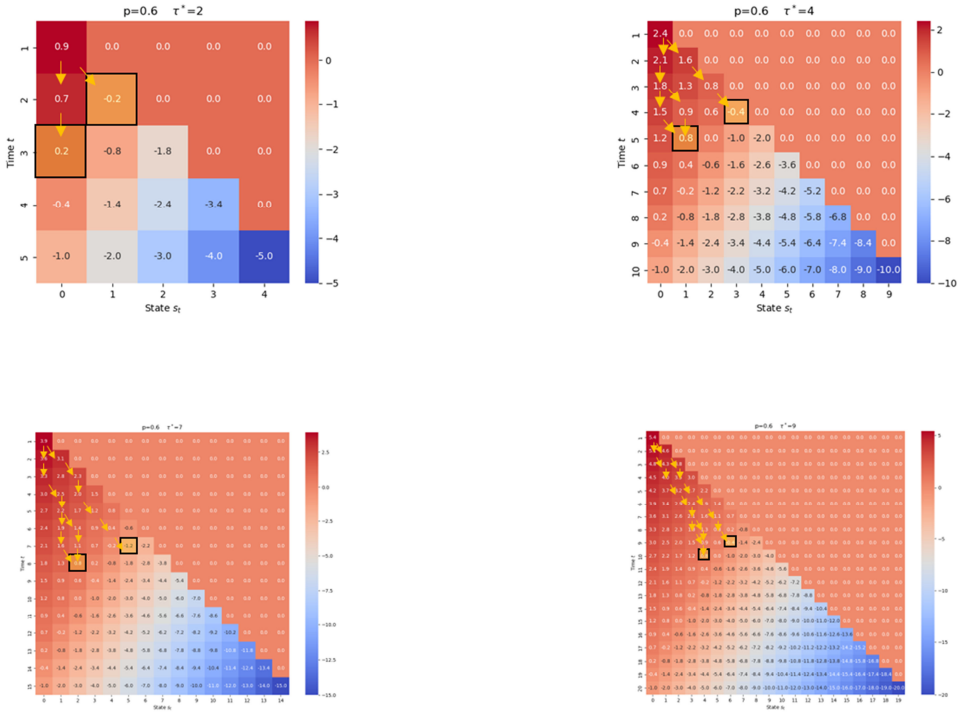

Figure 6: The transition sequences selected for the lucky and unlucky states. Black boxes indicate the selected time-state pairs (t_0, s_{t_0}). Yellow arrows mark the transition sequences leading to the selected states. The color map indicates the differences in q-values of `fuse` and `extract` predicted by the optimal state-dependent policy.