

DSSR

Discussion Paper No. 154

**Can Active Behavior Rewrite Fate?
Reinforcement-Learning Agents Learn to Lie Flat in
the Talent-versus-Luck Model**

Zhen Ren and Yinxing Li

September, 2026

**Data Science and Service Research
Discussion Paper**

Center for Data Science and Service Research
Graduate School of Economic and Management
Tohoku University
27-1 Kawauchi, Aobaku
Sendai 980-8576, JAPAN

Can Active Behavior Rewrite Fate? Reinforcement-Learning Agents Learn to Lie Flat in the Talent-versus-Luck Model

Zhen Ren¹ Yinxing Li¹

¹ Graduate School of Economics, Tohoku University

Abstract: Whether success comes from talent or from luck is an old debate. The Talent-versus-Luck (TvL) model drew wide attention with a blunt answer: luck decides who gets rich, and the richest are rarely the most talented. But its agents are passive. They stay put and take whatever luck reaches them, while real people move toward chances and away from trouble. To our knowledge we are the first to let TvL agents act and learn, using Double DQN, in a two-by-two design that crosses learning (off or on) with the kind of luck (asymmetric or symmetric).

Two results follow. The new one is behavioral. Under TvL’s original luck, where a chance is more likely to cost than to pay, the agents learn on their own to avoid opportunity altogether, a model version of “lying flat”. It is a rational choice rather than laziness, and the proof is that the same reward makes them seek events once luck is made fair. The learning also gives the shape of that retreat, which the arithmetic does not: the agents keep moving, meet almost nothing, and talent stops leaving any mark on wealth. The second result is that TvL’s core findings survive. The richest are still the luckiest rather than the most talented, and talent predicts wealth only once luck exposure is held fixed, because talent is simply the rate at which luck is turned into gains. Within the bounded agency modeled here, acting protects what an agent starts with but cannot make effort beat luck.

Keywords: Talent versus luck; reinforcement learning; agent-based model; luck-dominated success; talent as the conversion rate of luck; lying-flat; rational disengagement; incentive structure

Contents

1	Introduction	4
2	Literature Review	7
3	Model	8
3.1	Base Framework: Inheriting the TvL Model	8
3.2	Reinforcement-Learning Architecture: Double DQN	10
4	Experimental Design	18
4.1	Overview of Experiments	18
4.2	Baseline: Passive Random-Walk Benchmark (TvL Reproduction)	19
4.3	Exp.1: RL under Asymmetric Luck	19
4.4	Baseline2: Passive Benchmark under Symmetric Luck	20
4.5	Exp.2: RL under Symmetric Luck	20
5	Simulation Results	20
5.1	Baseline: A Reproduction of TvL’s Core Findings	21
5.2	Exp.1: RL under Asymmetric Luck, The Pure Effect of Agency	22
5.3	Baseline2: TvL’s Conclusions Survive under Symmetric Luck	24
5.4	Exp.2: RL under Symmetric Luck, and the Limits of Agency	25
5.5	Cross-Experiment Synthesis: The Four Experiments Jointly Confirm and Sharpen the Core TvL Conclusions	29
5.5.1	TvL confirmed: in every realization, the richest are always the luckiest	29
5.5.2	TvL sharpened: talent is the “conversion rate” of luck, and its zero cor- relation is only a masking effect	32
5.5.3	Cross-Experiment Comparison of Wealth Distribution and Talent Struc- ture	33
5.5.4	Cross-Experiment Comparison of the Emergent Strategy: Lying-Flat versus Seeking	40
5.6	Robustness Checks	41
6	Discussion	43
6.1	Two Laws of Individual Destiny	43
6.2	The New Finding: Lying-Flat as a Rational Choice	43
6.3	Limitations and Future Work	45
7	Conclusion	46
	References	48

Appendix A: Training Algorithm

50

1 Introduction

Whether individual success is primarily determined by innate talent or exogenous luck remains a central debate in the social and economic sciences. To quantitatively evaluate this dynamic, Pluchino et al. (2018) introduced an agent-based framework—the Talent-versus-Luck (TvL) model—wherein agents with normally distributed talent interact with randomly distributed positive and negative events throughout a simulated career. Despite assuming a Gaussian distribution of talent, their model generated a heavy-tailed, Pareto-like wealth distribution. Crucially, the most successful agents were characterized by exceptional luck rather than superior talent, directly challenging the meritocratic assumption that elite outcomes naturally reflect superior ability.

While the TvL model provided groundbreaking quantitative insights, it relies on a fundamental simplification: agents are entirely passive. In the original setting, agents remain stationary and absorb exogenous shocks without the capacity to actively pursue opportunities or mitigate risks. In reality, human individuals exhibit agency—dynamically adapting to their environments by relocating, changing occupations, and actively managing risk exposure. As Challet et al. (2020) observed, incorporating agent movement and strategic decision-making represents an essential yet unexplored extension of the baseline TvL model.

This study addresses this gap by examining whether active human agency—formalized through reinforcement learning—can mitigate or overturn the structural dominance of luck in wealth accumulation. We embed reinforcement learning agents capable of spatial navigation and adaptive strategy into the TvL environment.

The main contributions of this study are fourfold:

1. **Integration of Reinforcement Learning into the TvL Framework.** We present the first implementation of reinforcement learning within the TvL paradigm. Employing a 2×2 factorial experimental design (learning status $\times$ luck structure), we systematically test whether agency or structural luck symmetry can overturn the baseline conclusions, addressing the theoretical gap identified by Challet et al. (2020).
2. **Uncovering Structural Payoff Asymmetries.** Our learning framework exposes an implicit asymmetry in the original TvL design, wherein negative shocks strike unconditionally while positive opportunities require probabilistic conversion. This underlying imbalance directly motivates the implementation of a fairer, symmetric luck structure.
3. **Robustness of Luck Dominance under Bounded Agency.** We demonstrate that the core TvL conclusions remain remarkably resilient: across experimental conditions, the wealthiest individuals are consistently those who encounter the most favorable luck rather than those with the highest talent. Bounded agency alone is insufficient to dismantle the dominance of luck.

4. Emergent Rational Strategy of “Lying-Flat”. Under the original asymmetric luck payoff, agents spontaneously converge on a strategy of spatial withdrawal (“lying-flat”). Comparing this emergent policy across symmetric and asymmetric reward structures reveals that withdrawal is a fully rational adaptation to unfavorable risk-reward conditions rather than behavioral failure.

The paper builds up step by step around this question; Figure 1 shows the overall flow. We give each inherited TvL agent a Double DQN decision-making ability (§3) and run four experiments in a two-by-two design that crosses learning (off or on) with the luck structure (asymmetric or symmetric), read as two vertical contrasts, Baseline versus Exp.1 and Baseline2 versus Exp.2 (§5). The first contrast produced the key surprise, a learned lying-flat under asymmetric luck, and that discovery is what motivated the symmetric structure (supported by prior work on ability as risk avoidance, §3.1). A cross-experiment synthesis and discussion (§6) then draw the two conclusions, the survival and sharpening of TvL and the emergence of rational lying-flat, summed up as the idea that “agency has limits” (§7).

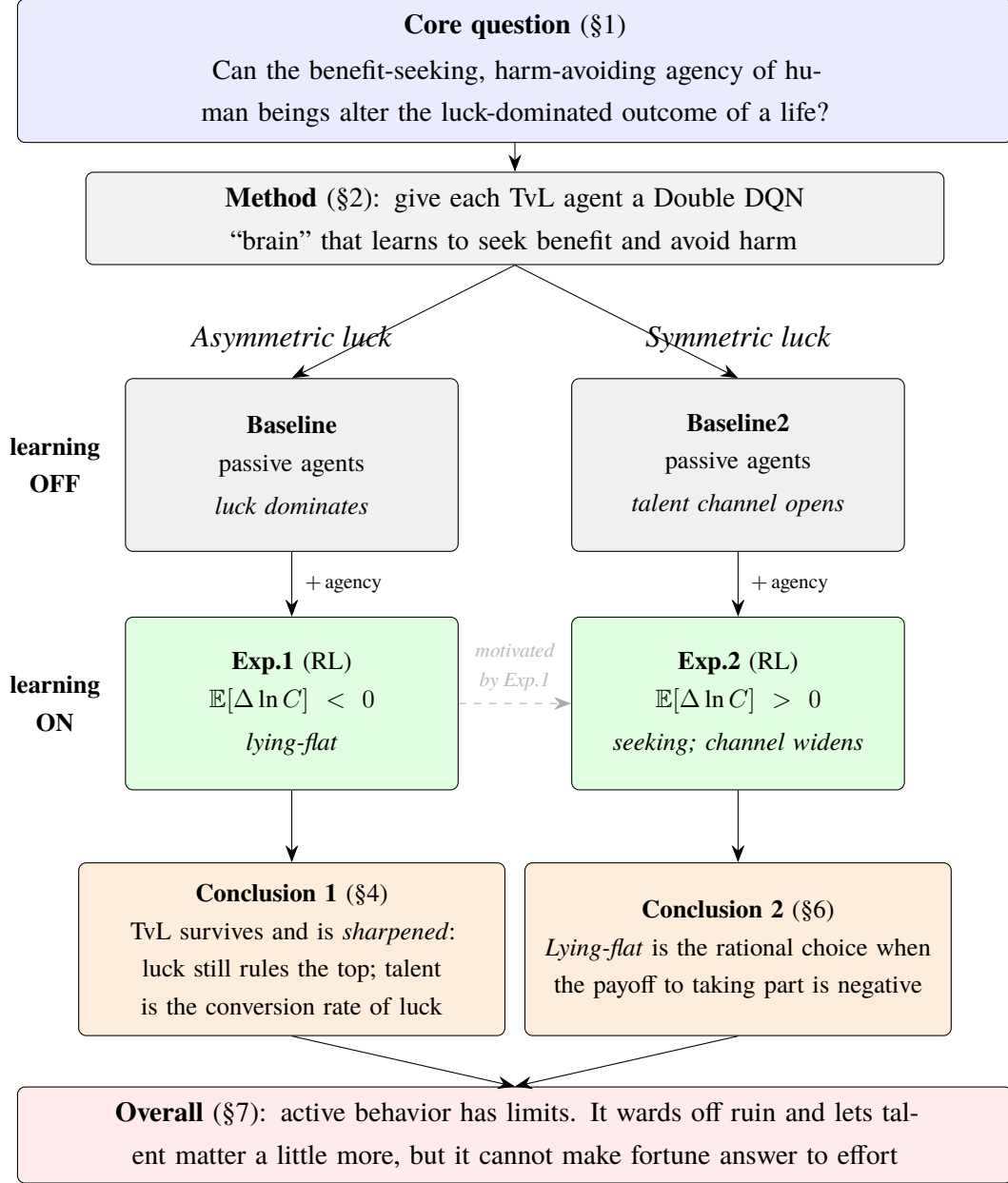


Figure 1. Overview of the study. From the core question (§1) and the RL-TvL model (§2), four experiments are run in a two-by-two design (§3) that crosses learning (off/on) with the luck structure (asymmetric/symmetric); reading down each column isolates the effect of agency. The rational retreat of Exp.1 under asymmetric luck motivated the fairer, symmetric structure. The results converge into two conclusions, the survival and sharpening of TvL (§4–5) and the emergence of lying-flat (§6), and an overall verdict that active behavior has limits (§7).

This paper is organized as follows. Section 2 introduce literature; Section 3 details the complete design of the RL-TvL model; Section 4 describes the experimental scheme; Section 5 reports the simulation results, including the cross-experiment confirmation of TvL and the socio-economic significance of the strategies that emerge; Section 6 discusses theoretical implications; Section 7 concludes.

2 Literature Review

The Talent-versus-Luck lineage. The idea that chance dominates success has deep roots: the “Matthew effect” of Merton (1968) and the cumulative-advantage mechanism of DiPrete & Eirich (2006) show how small early strokes of luck compound into large gaps, and Frank (2016) argues that luck’s role in winner-take-all careers is systematically underrated. Pluchino et al. (2018) made this computational in a deliberately simple agent-based model that has since become a widely used benchmark; Challet et al. (2020) then showed analytically why a Gaussian talent distribution becomes a heavy-tailed wealth distribution through multiplicative lucky events, echoing the CEO-pay mechanism of Gabaix & Landier (2008) in which talent acts as a conversion probability. Later extensions add variance-based decompositions of luck (Hamada, 2023), network interactions (Hur et al., 2024), and firm-level opportunity (Biondo et al., 2024), yet none lets its agents learn, and none takes up the alternative design Challet et al. (2020) pointed to but never tried, in which the events are fixed and the agents move. Ours is the first to do both. We keep these core settings unchanged: TvL’s Gaussian talent, its multiplicative lucky and unlucky events, and capital as the only measure of success. This is not because we think they are beyond question, but because keeping the same simple base fixed is what lets us cleanly separate the two things we add, agency and a fairer luck structure. Where those settings limit our conclusions we say so (§6).

Reinforcement learning as agency, and its use in social simulation. We model human agency with model-free reinforcement learning (Sutton & Barto, 2018), whose value-based branch runs from Q-learning (Watkins & Dayan, 1992) through deep Q-networks (Mnih et al., 2015) to the Double DQN (van Hasselt et al., 2016) we adopt. The most direct precedent is Sert et al. (2020), who placed Double-Q-learning agents on a spatial Schelling grid with an 11×11 observation window and five movement actions, exactly the setup our state and action design is built on. At the population scale, the AI Economist (Zheng et al., 2022) co-evolves an RL planner and RL households to discover tax policy, while RL agents have also been shown to learn prosocial norms (Peysakhovich & Lerer, 2018) and price competition (Calvano et al., 2020). In all of these, reinforcement learning is used to find the best behavior rather than to hand-code it, which is exactly why it can stand in for the gain-seeking, loss-avoiding agency that the original TvL model lacks.

Why only simulation can answer this. The question we ask cannot be settled in the real world, and reinforcement learning is what makes it workable. First, we live only once: a real person cannot replay a life hundreds of times to settle on a best strategy, and a career often ends before the best response to opportunity has even been found, whereas an RL agent carries the memory of each episode into the next and, in effect, lets “the same person” relive life until it approaches an optimum no individual could ever reach. Second, no real dataset records every step of every person’s wealth, decisions, and luck exposure, whereas the simulation preserves the full step-level trajectory (§4). Third, the key what-if tests here, switching agency on or off

and making luck symmetric or asymmetric while holding everything else fixed, cannot be run on a real society. For this study, then, agent-based reinforcement-learning simulation is the natural way to answer it.

3 Model

3.1 Base Framework: Inheriting the TvL Model

This study fully inherits the base settings of Pluchino et al. (2018): $N = 1000$ agents; a 201×201 toroidal spatial grid (with wrap-around boundaries); talent $T_i \sim \mathcal{N}(0.6, 0.1)$ fixed throughout after initialization; an initial capital $C(0) = 10$ common to all agents; and $N_E = 500$ event points (50% lucky / 50% unlucky). Fixing the initial capital identically for everyone deliberately strips out the “birth lottery” (Rawls, 1971; Roemer, 1998), so that the roles of talent, luck, and agency are separated cleanly from differences in origin. Following the original model, we also use wealth as the single measure of success. In reality success means many things; here, only to keep the model simple, we let the amount of wealth stand for all the things people pursue and value.

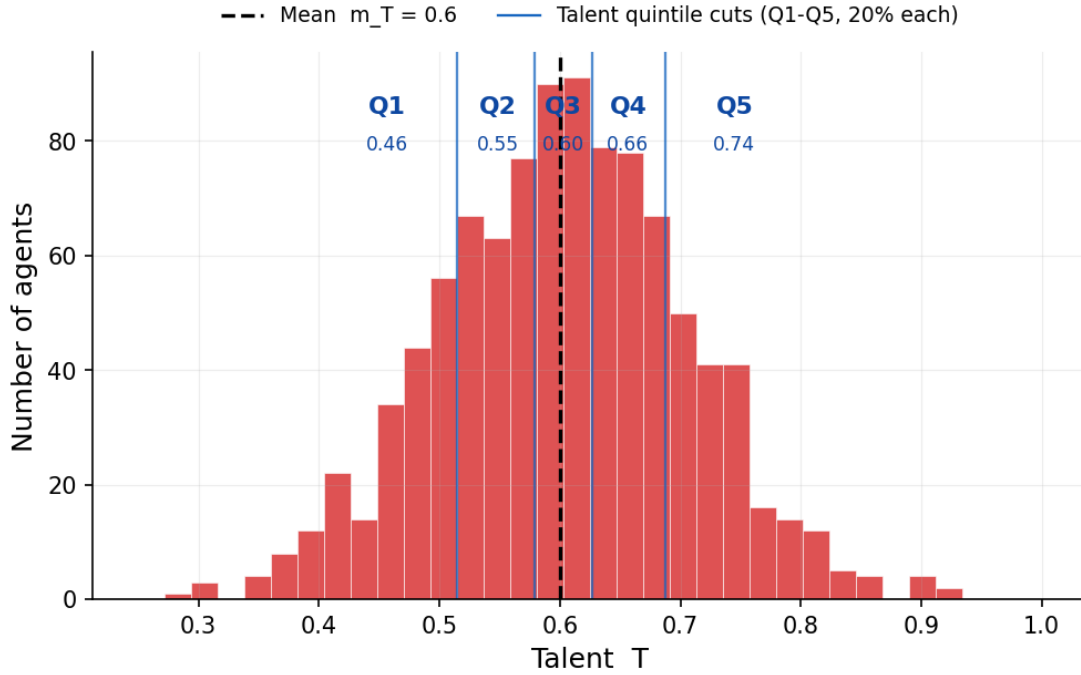


Figure 2. Talent distribution histogram ($N = 1000$, seed=2026). The horizontal axis is agent talent T_i , the vertical axis is the number of agents in each talent bin. Red bars are the actual talent distribution, a truncated normal $\mathcal{N}(0.6, 0.1)$ on $(0, 1)$. The black dashed line is the talent mean $m_T = 0.6$. The four blue vertical lines are the talent quintile cuts (the 20th, 40th, 60th, and 80th percentiles), which split the agents into five equal-sized groups Q1–Q5 of 20% each; the blue number above each group is its group-mean talent (0.46, 0.55, 0.60, 0.66, 0.74). Because talent is normal and tightly concentrated near 0.6, even the lowest quintile Q1 averages 0.46 rather than a value near zero, and 84% of agents have $T_i > 0.5$ (since 0.5 is one standard deviation below the mean). Most agents’ talent falls in 0.5–0.7, with few of very high (> 0.8) or very low (< 0.4) talent. The talent distribution stays fixed throughout the simulation; all talent differences are set by the initial sampling.

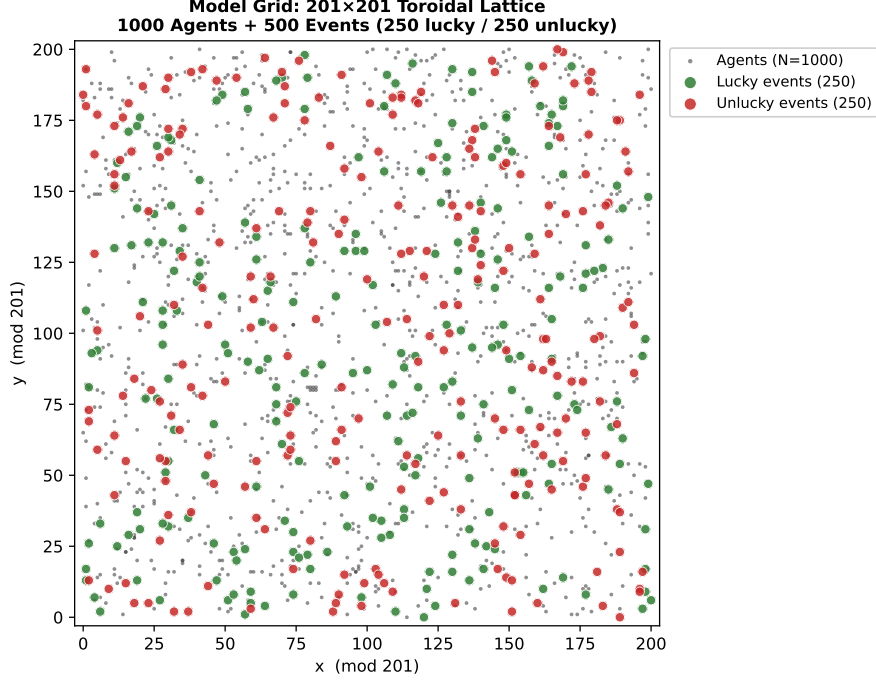


Figure 3. Schematic of the model grid (initial layout of one episode). On the 201×201 toroidal grid, $N = 1000$ agents (grey dots) and $N_E = 500$ event points (250 lucky, green; 250 unlucky, red) are randomly distributed. All positions are regenerated randomly at the start of each episode; thereafter events random-walk $v_E = 2$ cells per step, and agents move according to their policy (fixed in the Baseline). The grid is toroidal, with top-bottom and left-right boundaries connected respectively.

The event logic is identical to the original: when a lucky event strikes the position of agent k , capital doubles with probability T_k ; an unlucky event halves capital unconditionally:

$$C_k(t+1) = \begin{cases} 2C_k(t) & \text{with probability } T_k \text{ (lucky event strikes)} \\ C_k(t)/2 & \text{(unlucky event strikes, unconditional)} \\ C_k(t) & \text{(no event strikes)} \end{cases} \quad (1)$$

This is exactly as in Pluchino et al. (2018): a lucky event doubles capital with probability equal to the agent’s talent T_k , an unlucky event halves capital unconditionally, and capital is unchanged when no event strikes. On top of this multiplicative dynamic, this study introduces the active movement decisions of RL agents; the capital-update rule itself, however, is unchanged.

Time scale and event motion. Each agent makes a decision every $\delta t = 6$ months (as in Pluchino et al. 2018), and each episode spans 40 years over $T_{\max} = 80$ time steps. The $N_E = 500$ event points are not stationary: at each time step each performs a random walk of $v_E = 2$ cells on the 201×201 toroidal grid; both the positions of the events and their lucky/unlucky identities are reassigned randomly at the start of each episode. So agents keep facing a shifting, unpredictable mix of chances and risks, and must move to reach lucky events and avoid unlucky ones.

Encounter determination and trigger rules. “Striking” in Eq. (1) is determined by the following spatial rule. After both agent movement and event walking are complete at each step, define the *contact neighborhood* of agent k as the Von Neumann radius-1 neighborhood cen-

tered on its cell (y_k, x_k) (the center cell plus its up, down, left, and right cells, five in total):

$$\mathcal{N}_k = \{(y_k, x_k), (y_k \pm 1, x_k), (y_k, x_k \pm 1)\} \pmod{G}, \quad G = 201. \quad (2)$$

As soon as at least one lucky (or unlucky) event falls within $\mathcal{N}_k$, agent k is deemed to have encountered a lucky (or unlucky) event, recorded respectively as $\text{on}_k^L, \text{on}_k^U \in \{0, 1\}$ (multiple events of the same type within the neighborhood are recorded only as “present or not,” without multiplicity).

Using a radius-1 contact region rather than exact same-cell collision handles missed “pass-through” detections on the discrete grid: agents and events both move at speed 2, so under same-cell detection alone the two could swap positions within a single step and brush past undetected (e.g., at t the agent is in cell 10 and the event in cell 12; at $t+1$ they swap exactly, and same-cell detection fails entirely). Here speed 2 is the single-step *reachable range* and radius 1 is the *contact-determination range*; the two are conceptually distinct. Radius 1 strikes a balance between “not missing pass-throughs” and “not over-inflating the contact probability,” consistent with the small contact radius of Pluchino et al. (2018).

Whether an encounter actually *triggers* a capital change depends on talent T_k and varies with the luck structure:

- Lucky events (identical across all experiments): trigger with probability T_k , doubling capital ($\times 2$); with probability $1 - T_k$ they fail and capital is unchanged.
- Unlucky events: under *asymmetric* luck (Baseline, Exp.1) they trigger unconditionally, halving capital ($\times 0.5$), the case in Eq. (1); under *symmetric* luck (Baseline2, Exp.2) they trigger with probability $1 - T_k$, i.e., the agent actively dodges the loss with probability T_k .

If within the same step an agent both encounters and triggers a lucky and an unlucky event, the two cancel out, leaving no net change ($\times 2 \times 0.5 = \times 1$).

The symmetric rule makes a real claim about what talent is: not just the ability to grab opportunity, but also the ability to spot and avoid risk. People who are more financially literate make far fewer costly mistakes (Lusardi & Mitchell, 2014), lower cognitive ability raises the chance of costly errors in credit markets (Agarwal et al., 2009), and at the firm level dynamic capabilities let able firms cut losses and survive shocks (Teece et al., 1997). We therefore let the single parameter T_i govern both the opportunity-capture rate and the risk-dodge rate, treating talent as general ability rather than a narrow, industry-specific skill.

3.2 Reinforcement-Learning Architecture: Double DQN

State space. At each time step t , agent i observes a 245-dimensional state vector, formed by concatenating two channels of an 11×11 local field of view centered on its own position

with three scalar features:

$$s_i(t) = \left[\underbrace{\mathbf{p}_i(t)}_{\substack{121\text{-dim} \\ \text{event presence}}}, \underbrace{\mathbf{d}_i(t)}_{\substack{121\text{-dim} \\ \text{neighbor density}}}, \underbrace{\mathbf{x}_i(t)}_{\substack{3\text{-dim} \\ \text{scalars}}} \right] \in \mathbb{R}^{245}. \quad (3)$$

- Presence channel $\mathbf{p}_i(t) \in \{0, 1\}^{121}$: an 11×11 toroidal window of radius 5 centered on agent i 's current position, marking cell by cell whether at least one event is present (binary). This channel does not distinguish lucky from unlucky events, and records only “present or not” for multiple events in a cell; the agent can thus sense the spatial distribution and number of occupied cells of events within view, but cannot tell their good/bad type. Merging the two event types into one presence signal captures the fact that an opportunity cannot be told apart in advance, the “uncertainty” of Knight (1921) rather than measurable risk, along with an information-cascade setting in which an agent sees where others are rushing but not why (Bikhchandani et al., 1992; Banerjee, 1992).
- Density channel $\mathbf{d}_i(t) \in [0, 1]^{121}$: the number n of neighboring agents in each cell of the same window, normalized as $\min(n/10, 1)$ (saturating at 1 when a cell holds 10 agents), reflecting local crowding.
- Scalar features $\mathbf{x}_i(t) = [\log_{10} C_i(t)/10, g_i(t), t/T_{\max}] \in \mathbb{R}^3$, where: (i) $\log_{10} C_i(t)/10$ is the current log-capital (a scale-invariant wealth level); (ii) $g_i(t)$ is the most recent single-step log-return, defined as $g_i(t) = \text{clip}(\log_{10}[C_i(t)/C_i(t-1)], -1, 1)$, the only signal by which the agent indirectly senses its own recent fortunes (capital doubled/halved/unchanged last step); (iii) $t/T_{\max}$ is career progress ($T_{\max} = 80$ steps = 40 years).

Both spatial channels are *egocentric* local fields of view cropped around the agent's own position, the window center is always the agent itself, and each cell encodes an offset (δ_y, δ_x) relative to itself ($\delta_y, \delta_x \in \{-5, \dots, 5\}$). The agent can therefore distinguish, for each event and each neighboring agent in view, its bearing and distance relative to itself (e.g., “an event 3 cells directly above,” “3 companions in the first cell to the right”), but cannot learn any absolute coordinate, nor sense its own global position on the 201×201 grid (the toroidal boundary is transparent to it). This egocentric representation makes the observation *translation-invariant*: the agent's response to a given local configuration is the same wherever it appears on the grid, which is both what lets all agents share a single policy network (parameter sharing) and what allows the learned strategy to generalize across the whole grid.

The total state dimension is $11 \times 11 \times 2 + 3 = 245$. The state vector deliberately contains neither the agent's own talent T_i nor the lucky/unlucky type of events, the presence channel merges the two event types into a single “presence” signal. Talent is thus fully hidden, and the agent can sense its influence only through the indirect path “talent $\rightarrow$ capital trajectory $\rightarrow$ scalar g_i ”; this is precisely the basis on which this paper treats talent stratification as spontaneously

emergent rather than explicitly encoded, and it mirrors how poorly people know their own ability (Kruger & Dunning, 1999; Bénabou & Tirole, 2002).

The 11×11 observation window (radius 5 cells) jointly captures the bounded rationality of Simon (1955) and the rational-inattention limit of Sims (2003): the agent can neither observe the global distribution over the 201×201 grid nor process more than local information under a limited channel capacity, concentrating attention on its near neighborhood. Radius 5 corresponds to about 5.5% linear coverage of the grid, wide enough that nearby events (moving at 2 cells/step) can be sensed within a few steps, yet narrow enough that the agent must move actively to contact distant event clusters. The way the agent decides, from the signals visible within the window, whether to keep moving has the same logic as the optimal-stopping search of Stigler (1961) and the reservation-wage decision of McCall (1970): it decides, given the information, whether to keep “searching” (moving).

Action space. At each time step, the agent chooses one of 5 discrete actions:

$$\mathcal{A} = \{0, 1, 2, 3, 4\} \longleftrightarrow \{\text{stay}, \text{N}, \text{S}, \text{W}, \text{E}\}. \quad (4)$$

The four directional actions move the agent $v_A = 2$ cells along the corresponding direction on the toroidal grid (modulo $G = 201$); the stay action ($a = 0$) keeps the agent in its cell, where v_A is the agent’s movement speed. This “free wait” option lets the agent linger in a favorable location and act when the time is right, and it is central to the avoidance strategy that emerges in Exp.1.

Reward function. The immediate reward agent i receives at step t is the clipped single-step log-capital return:

$$R_i(t) = \text{clip}\left(\ln \frac{C_i(t+1)}{C_i(t)}, -\ln 2, +\ln 2\right), \quad (5)$$

where $C_i(t)$ is agent i ’s capital at time t and $\text{clip}(x, a, b)$ restricts x to the interval $[a, b]$. This reward is the single-step change in log-capital.

Why log-return, the Kelly criterion. The main reason for choosing log-return rather than an arithmetic return is that TvL’s wealth evolution is a *multiplicative process* (each event encounter multiplies capital by $\times 2$ or $\times 0.5$). Kelly (1956) showed that in a multiplicative process the growth rate of capital is governed by the expected log-return, and Breiman (1961) later proved that maximizing it almost surely leaves an agent richer in the long run than any other strategy (together, the Kelly criterion); conversely, maximizing arithmetic expectation (linear utility), though superior in single-step expectation, leads to over-betting and drives the long-run probability of ruin toward 1. Using log-return as reward is the same as having the agent pursue this long-run optimum on its own, and it fits the real-world fact that rational people avoid ruin. This multiplicative structure also directly determines the magnitude of the $\pm \ln 2$ clipping: a single event at most doubles capital ($\times 2$, log-return $+\ln 2$) or halves it ($\times 0.5$, log-return $-\ln 2$), so $\pm \ln 2 \approx \pm 0.69$ is the natural bound on the single-step reward; clipping to this magnitude

suppresses reward spikes from rare consecutive events and stabilizes training.

Beyond the theoretical rationale, log-return has two engineering advantages: (1) under multiplicative dynamics, log-returns are additive, so the reward signal is temporally stable; (2) when capital spans many orders of magnitude, the log transform keeps the gradient signal balanced across scales. It is also scale-invariant: doubling capital from 10 to 20 and from 1000 to 2000 produce the same reward $\ln 2 \approx +0.69$, which both reflects the diminishing marginal utility of wealth and prevents high-capital agents from dominating the gradient updates of the shared replay buffer. No idle penalty is introduced, ensuring that “stay” remains a legitimate optimal action.

The agent knows very little. It does not know the form of the reward; it does not know its own talent T_i ; within its observation window it cannot tell whether an event in view is lucky or unlucky (only the one-step change in its own capital gives it delayed, indirect feedback); and it does not know the probabilities with which either kind of event fires. Model-free reinforcement learning hands the agent exactly one signal per step: the realized reward, that is, whether its capital has just grown or shrunk. Every behavioral regularity reported in this paper, avoidance, opportunity-seeking, talent-stratified search, and disengagement alike, is therefore not a pre-programmed rule but the result of costly trial and error against the environment. Whether the expected log-return of an encounter is negative (Exp.1) or positive (Exp.2) is a truth the agent is never told; it can discover it only through its own accumulated gains and losses. The sharpest evidence that behavior is not encoded in the reward is the contrast between Exp.1 and Exp.2 themselves: the reward function is identical in both, yet the learned strategies are opposite, so what the agent converges to is determined by the structure of the environment as revealed through experience, not by the reward specification.

Double DQN architecture and training. The value network is a three-layer multi-layer perceptron (MLP), where each hidden layer consists of a fully connected (FC) linear transformation followed by a Rectified Linear Unit (ReLU) activation (using the Double DQN of van Hasselt et al. 2016, within the DQN framework of Mnih et al. 2015):

$$245 \xrightarrow{\text{FC+ReLU}} 128 \xrightarrow{\text{FC+ReLU}} 64 \xrightarrow{\text{FC}} 5 \text{ (} Q\text{-values)}, \quad (6)$$

totaling about 4.0×10^4 parameters. The architecture is deliberately lightweight, to avoid overfitting limited experience while retaining enough capacity to recognize local event patterns within the 11×11 window. The 1000 agents share a single set of network parameters: at each time step all agents’ observations are stacked into a $(1000, 245)$ matrix for one batched forward pass, which both keeps the policy globally consistent and reduces the computational cost from $O(N)$ forward passes to one. One consequence is worth stating early: because the parameters are shared, the agents all follow the same policy and differ only in talent and in where they stand, so the model has no room for strategic variety, imitation, or learning from a neighbor. We return to what this rules out in §6.

Double DQN suppresses the value-overestimation bias of standard DQN by separating *action selection* (online network Q_θ) from *value estimation* (target network Q_{θ^-}), with the temporal-difference target:

$$y_i(t) = R_i(t) + \gamma Q_{\theta^-}(s_i(t+1), a^*) (1 - d_i(t)), \quad a^* = \underset{a}{\operatorname{argmax}} Q_\theta(s_i(t+1), a), \quad (7)$$

where θ and θ^- are the parameters of the online and target networks, $\gamma = 0.99$ is the discount factor, and $d_i(t) \in \{0, 1\}$ marks whether the episode terminates at step t (masking the bootstrap term at termination). Training uses the Huber loss and the Adam optimizer (learning rate $\eta = 3 \times 10^{-4}$); the ε -greedy exploration rate decays linearly from 0.99 to 0.05 (reaching its minimum at episode 200); the target network is hard-copied from the online network every 10 episodes; and the transitions of all 1000 agents share a single replay buffer (capacity 8×10^5), providing diverse off-policy experience.

Convergence criterion. We stop RL training not by eyeballing a “flattening of the reward curve” but by a joint, talent-stratified test based on how the policy behaves. There are two reasons: (1) the mobility rate describes the policy itself and is less thrown off by a single random event than outcome measures such as capital; (2) splitting by talent keeps the strategy changes of high- and low-talent agents from canceling out in the population mean and faking convergence. Concretely, after each episode the $N = 1000$ agents are divided by the 20/40/60/80 percentiles of talent T_i into 5 talent quantile groups (about 200 each); within a sliding window of the most recent $W_c = 20$ episodes, we check for each group whether the standard deviation over the window of three mobility metrics falls below threshold:

1. Overall mobility rate (the fraction of steps in which a move action is executed): standard deviation < 0.05 ;
2. Mobility rate when events are present (counting only steps with an event in view): standard deviation < 0.08 ;
3. Mobility rate when no events are present (counting only steps with no event in view): standard deviation < 0.08 .

The latter two thresholds are slightly looser (0.08 vs 0.05), as they are conditional subsamples split by “encountered an event or not” and hence noisier; we additionally require at least 5 episodes with valid observations for that group within the window (a given quantile group may, in a given episode, encounter no events at all, leaving the conditional mobility rate undefined). All 5 quantile groups $\times$ 3 metrics = 15 conditions must be satisfied simultaneously, and the number of training episodes must have reached the lower bound $M_{\min} = 200$, before convergence is declared; if the upper bound $M_{\max} = 1000$ episodes is reached without satisfaction, training is forcibly terminated. This test requires that every mobility measure in every talent group has settled down, much stricter than watching aggregate measures alone, since it makes

sure the strategies of agents with different talent have each converged rather than hiding one another behind a false overall calm. After convergence, a further 100 post-convergence episodes are run (with exploration fixed at a near-greedy $\varepsilon = 0.05$ and no further parameter updates) as the final statistical basis, whose aggregation convention is detailed in the statistical-protocol note below.

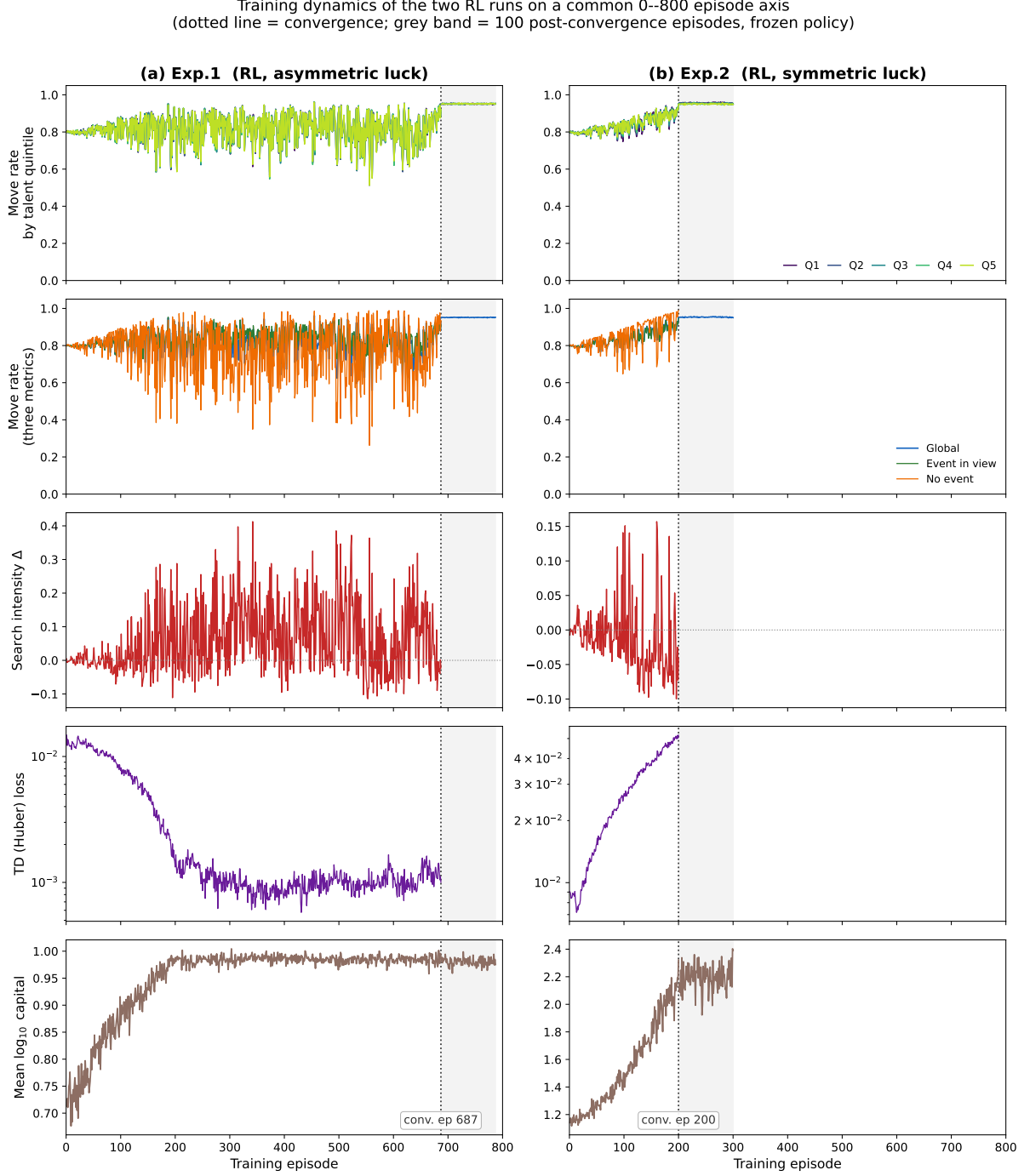


Figure 4. Training dynamics of the two RL runs on a common 0–800 episode axis, so the two runs are directly comparable despite converging at different episodes. Columns: (a) Exp.1 (RL, asymmetric luck), (b) Exp.2 (RL, symmetric luck). The five rows show, from top to bottom, every component the talent-stratified joint convergence criterion tracks: (1) the move rate by talent quintile (Q1–Q5), whose five lines nearly coincide, showing one undifferentiated strategy across talent; (2) the three aggregate move-rate metrics, global, with an event in view, and with no event in view; (3) the search intensity Δ ; (4) the TD (Huber) loss on a log scale; and (5) the population mean $\log_{10}$ capital. The dotted vertical line marks the convergence episode (687 for Exp.1, 200 for Exp.2), at which all 5 talent quintiles $\times$ 3 mobility metrics = 15 joint conditions hold at once; the grey band is the 100 post-convergence episodes (frozen policy, $\varepsilon = 0.05$). The passive Baseline and Baseline2 involve no learning and are omitted.

The full hyperparameters are given in Table 1; the complete training pseudocode is provided in Appendix A (Algorithm 1).

Table 1. RL-TvL model hyperparameters and rationale

Parameter	Symbol	Value	Rationale
<i>TvL environment parameters</i>			
Grid size	G	201×201 (toroidal)	follows Pluchino et al. (2018)
Number of agents	N	1,000	follows the original
Total events	N_E	500	250 lucky + 250 unlucky; follows the original
Steps/episode	$T_{\max}$	80	40 years $\times$ 6 months/step; follows the original
Initial capital	C_0	10	follows the original
Capital floor	$C_{\min}$	10^{-4}	prevents log 0
Event/agent speed	v_E, v_A	2 cells/step	equal speed; events random-walk each step
Talent distribution	T_i	$\mathcal{N}(0.6, 0.1)$, truncated (0, 1)	follows the original
<i>Observation and action</i>			
Observation window	W	11×11	bounded rationality; covers $\approx 5.5\%$ of grid width
State dimension	d	245	$11^2 \times 2 + 3$; excludes talent (hidden by design, sensed indirectly via capital)
Number of actions	$ \mathcal{A} $	5	stay + 4 directions
<i>Double DQN hyperparameters</i>			
Hidden layers	–	[128, 64]	enough capacity, avoids overfitting limited experience
Learning rate	η	3×10^{-4}	Adam default
Discount factor	γ	0.99	approximate time-consistency within an 80-step episode
Batch size	B	256	standard for DQN applications
Replay capacity	–	8×10^5	diverse off-policy experience across episodes
Target sync	–	hard copy every 10 episodes	prevents nonstationarity
ε start/end	–	0.99 / 0.05	early exploration $\rightarrow$ near-deterministic exploitation
ε decay length	–	200 episodes	decays to terminal value by episode 200
Min/max training episodes	–	200 / 1000	convergence fallback termination
Post-conv episodes	–	100	per-episode mean $\pm$ std statistical basis after convergence
Random seed	–	2026	reproducibility

Note: both RL experiments (Exp.1–2) share the parameters above, differing only in luck structure (asymmetric/symmetric) (see Table 2); the convergence criterion uses the talent-stratified multi-metric joint standard.

4 Experimental Design

4.1 Overview of Experiments

We design four experimental conditions in a clean two-by-two factorial layout, crossing two switches, learning off or on and luck asymmetric or symmetric, to systematically test the key properties of the RL-TvL model; see Table 2. The factorial structure lets us read off two effects cleanly: moving down a column isolates the effect of agency (switching learning on) at a fixed luck structure, and moving across a row isolates the effect of the luck structure (asymmetric to symmetric) at a fixed level of agency.

A word on how to read this design: the two-by-two layout is how we organize the analysis, not the order in which the experiments were conceived. The study proceeded sequentially. Symmetric luck was not a preplanned axis but a response to what Exp.1 revealed, that under the original asymmetric luck a learning agent rationally withdraws; and Baseline2 was added only afterward, once Exp.2 was in place, as its passive control to complete the factorial cell. We therefore introduce the four experiments below in the order we present them, Baseline $\rightarrow$ Exp.1 $\rightarrow$ Baseline2 $\rightarrow$ Exp.2, and keep the two-by-two table above purely for analysis.

Table 2. The two-by-two factorial design: learning $\times$ luck structure

	Asymmetric luck	Symmetric luck
Reinforcement learning: No	Baseline	Baseline2
Reinforcement learning: Yes	Exp.1	Exp.2

Note: rows switch reinforcement learning off/on; columns switch the luck structure between asymmetric (unlucky events always strike) and symmetric (unlucky events strike with probability $1 - T_i$, so talented agents dodge them with probability T_i). All four conditions share the identical environment.

Search intensity Δ . A diagnostic for whether an agent changes how much it moves when an event comes into view (it captures a change in move rate, not the direction the agent heads), defined as the difference between the mobility rate when events are present and when they are absent:

$$\Delta \equiv \bar{m}_{\text{event}} - \bar{m}_{\text{no-event}}, \quad (8)$$

where $\bar{m}_{\text{event}}$ is the average mobility rate over time steps with an event in view and $\bar{m}_{\text{no-event}}$ that over steps with no event in view (the two being metrics 2 and 3 of the convergence criterion). $\Delta > 0$ indicates more movement when events are in view; $\Delta \approx 0$ indicates a move rate independent of whether events are in view; $\Delta < 0$ indicates slowing down when events are in view. Note that Δ captures only whether an agent moves in response to a visible event, not which direction it takes; whether the population steers toward events or away from them is measured separately by the event-encounter rate (§5.4).

Data-logging protocol. To serve both convergence diagnosis and behavioral analysis, this paper uses different logging granularities for the two phases. Training phase (at most $M_{\max} =$

1000 episodes): each episode records only episode-level aggregates, overall mobility rate, conditional mobility rates with/without events in view, search intensity Δ , mean log-capital, TD loss, and each agent’s within-episode mobility rate (the last for the talent-stratified convergence test). These aggregates are used to plot learning curves and to run the convergence criterion; during training the strategy is not yet stable, and step-level trajectories have no analytical value while being costly to store (about $80 \times 1000 \times 1000$ records), so they are not logged. After convergence: a further 100 post-convergence episodes are run ($\varepsilon = 0.05$, network parameters frozen), recording the full step-level trajectory, each agent’s capital, action, coordinates, event encounter and trigger, and whether it can afford to move, at every step (each a $100 \times 1000 \times 80$ tensor). All analyses of the mature strategy’s behavior in this paper (mobility rate, search intensity, path-dependence, etc.) are based on these 100 episodes of step-level data.

Statistical protocol. All cross-sample scalar metrics in this paper (the talent–wealth correlation r , the fraction below initial capital, the top-20% wealth share, the mobility rate and search intensity Δ , etc.) follow the convention of “compute per realization, then take the mean $\pm$ standard deviation”, that is, for each independent run (Baseline, Baseline2) or each post-convergence episode (Exp.1–2) the metric is computed separately, and the mean and standard deviation are then taken over the resulting sequence of metrics, rather than first pooling or averaging the raw capital across realizations and then computing a single metric. The latter would artificially depress the variance and inflate the talent–wealth correlation, and is therefore never used here. Specifically, the Baseline and Baseline2 each use 100 independent runs and the RL experiments (Exp.1–2) use 100 post-convergence episodes. Figures that must show individual trajectories (such as capital-evolution paths) use a representative single run (the one with the highest final wealth), whereas wealth-distribution histograms and power-law fits are pooled over many realizations, both being visualizations rather than statistics entering the above means.

4.2 Baseline: Passive Random-Walk Benchmark (TvL Reproduction)

A full reproduction of the original model of Pluchino et al. (2018): $N = 1000$ agents stay fixed and immobile while events traverse the grid by random walk each step, with strikes checked every six months (80 steps = 40 years in total). We run 100 independent simulations and average them to remove the influence of single-run randomness. Main validation metrics: the power-law wealth distribution (top-20% wealth share, number of agents below initial capital), the weak talent–wealth correlation (Pearson $r \approx 0.06$), and a markedly higher count of lucky-event strikes among the top-wealth tier than the bottom.

4.3 Exp.1: RL under Asymmetric Luck

We replace the passive agents of the Baseline with Double DQN agents while keeping the asymmetric luck structure unchanged. This experiment isolates the effect of “RL agency” itself on the wealth distribution. We run at most 1000 episodes and, after convergence, take 100 post-

convergence episodes, computing metrics per realization and then taking the mean $\pm$ standard deviation as the statistical basis.

4.4 Baseline2: Passive Benchmark under Symmetric Luck

Baseline2 is the passive control for the symmetric luck structure, and its motivation comes from what Exp.1 will reveal (§5.2): once learning shows that the original asymmetric luck makes participation a losing game, the natural next question is what a fairer luck structure would change, first without agency (here) and then with it (Exp.2). Baseline2 was thus added after Exp.2, as its passive counterpart that completes the two-by-two cell. It is identical to the Baseline in every respect, the same $N = 1000$ passive, immobile agents and the same wandering events over 100 independent runs, with a single change: the luck structure is made *symmetric*. Unlucky events no longer strike unconditionally; instead they strike with probability $(1 - T_i)$, so that a talented agent dodges misfortune with probability T_i , exactly as it seizes opportunity with probability T_i . Because the agents still cannot move, this condition isolates the pure effect of the luck structure with agency held off: any change in the talent–wealth relationship, relative to the Baseline, must come from the shape of luck alone, not from behavior. Baseline2 thus separates the structural contribution of symmetric luck from the behavioral contribution that RL adds in Exp.2.

4.5 Exp.2: RL under Symmetric Luck

Exp.1 reveals that asymmetric luck (unlucky events striking 100% of the time) makes the expected log-return of any event encounter negative, leading RL to learn a pure avoidance strategy in which talent’s predictive power for wealth drops to zero. Exp.2 dismantles this mechanism by changing the luck structure to symmetric and talent-dependent: lucky events strike with probability T_i (higher talent $\rightarrow$ easier to seize opportunity), unlucky events strike with probability $(1 - T_i)$ (higher talent $\rightarrow$ better at avoiding bad luck).

Under this setting the expected log-return changes from zero into a linear function of talent:

$$\mathbb{E}\left[\ln \frac{C'}{C} \mid \text{encounter}\right] = \frac{1}{2} \cdot T_i \ln 2 + \frac{1}{2} \cdot (1 - T_i) \cdot (-\ln 2) = \ln 2 (T_i - 0.5).$$

When $T_i > 0.5$ the expected log-return is positive; when $T_i < 0.5$ it is negative; at $T_i = 0.5$ it is exactly zero. Taking the mean talent $T_i = 0.6$ as an example, $\ln 2 (0.6 - 0.5) \approx +0.069$, each encounter yields a small but positive expected log-growth. Whether the talent–wealth channel can thereby reopen is the core question of this experiment.

5 Simulation Results

This section reports the four experiments in turn, reading each the same way. For each condition we first ask whether the core TvL propositions (luck dominance, power-law wealth) still

hold, and then record the socio-economic behavior that emerges spontaneously as the RL agents adapt without supervision. A short summary closes each experiment, with a global synthesis in §5.5. All experiments start from $C_0 = 10$; statistical metrics are based on 100 post-convergence episodes, computed per realization and then averaged as mean $\pm$ standard deviation (for the convention see the statistical protocol in §4).

5.1 Baseline: A Reproduction of TvL’s Core Findings

The original (Pluchino et al., 2018) was in NetLogo; we reproduce it in Python as the baseline for the RL experiments. The setup matches the original: $N = 1000$ fixed agents, $N_E = 500$ random-walking events (50% lucky / 50% unlucky), $T_i \sim \mathcal{N}(0.6, 0.1)$ (truncated to $(0, 1)$), initial capital $C_0 = 10$, over 80 steps (40 years, 6 months per step); 100 independent runs.

The Baseline results are read from panel (a) of the cross-experiment figures collected in §5.5: talent versus final wealth (Figure 9), the power-law wealth distribution (Figure 10) and its histogram form (Figure 11), the capital paths by talent quintile (Figure 12), and the event-encounter distributions (Figure 7). The model grid and the talent distribution appear in the model section (Figures 3 and 2).

We reproduced every statistic in the original and compared it with our Python version; the qualitative summary follows.

Seven statistics come out consistent with the original, covering all its qualitative core propositions: (1) the most successful agent’s talent is only slightly above the mean ($T^* \approx m_T$), not the highest; (2) the capital distribution is extremely unequal and power-law in form; (3) lucky/unlucky event frequencies follow exponential distributions (the decay is slower in proportion to the higher encounter rate of the 5-cell neighborhood); (4) about 16% neutral lives; (5) high-talent agents have slightly higher mean capital than mean-talent agents; (6) high-talent agents ($T > 0.8$) very rarely become the most successful (3–8%); (7) roughly a third of agents with $T > 0.7$ (34% vs. the original’s 32%) end with capital above the initial value. The quantitative differences are all explained by the single implementation difference (collision area) and do not affect the comparability of the model’s conclusions. The largest such difference is the top-20% wealth share, which reaches 97% here rather than the 80% of a strict Pareto rule: our 5-cell collision neighborhood raises the encounter rate, so multiplicative compounding runs stronger and concentrates wealth further. The direction, a heavy luck-driven tail, is TvL’s; the exact concentration is sharper.

The Baseline gives a reliable benchmark. We check each later experiment against five core indicators: ① power-law wealth (α, R^2); ② luck’s dominance over wealth (talent–wealth r); ③ the talent of the most successful (T^*); ④ the wealth of the most talented ($C(T_{\max})$); and ⑤ the fraction ending below C_0 . On all five the Baseline matches the original TvL: $\alpha = 0.32$ ($R^2 = 0.70$), $r = +0.063 \approx 0$ (luck dominates), $T^* = 0.69 \approx m_T = 0.60$ (far from the highest

talent in the same draw, ≈ 0.92), $C(T_{\max}) = 10 = C_0$, and 45.7% below C_0 . The values for all four conditions are collected in the master table (Table 6).

5.2 Exp.1: RL under Asymmetric Luck, The Pure Effect of Agency

In the Baseline the agents are passive, fixed in place, able only to wait for luck to reach them. Real people act, moving toward chances and away from danger. Exp.1 restores this: it holds the original asymmetric luck fixed and switches learning on, so any change from the Baseline comes from the agents’ new ability to act, and nothing else.

We replace the agents with Double DQN (asymmetric luck, seed=2026); training converges at episode 687 (passing the talent-stratified convergence test; the training dynamics of the two RL conditions on a common episode axis are shown in Figure 4). The core results over 100 post-convergence episodes:

- 78.8% of agents end with capital equal to $C_0 = 10$ ($\pm 0.1\%$), corresponding to a mean of 10.78 and a median of 10.00;
- only 12.9% end below C_0 , far better than the Baseline’s 45.7%;
- the top 20% hold only 32.5% of wealth (Baseline: 97.3%), inequality nearly eliminated;
- the talent–wealth correlation $r \approx +0.03$ (essentially zero), with the top performers’ average talent at 0.634.

After convergence the agents’ mobility rate reaches 95.1%, moving almost every step, and the mobility rates of the different talent quantiles (Q1 to Q5) are exactly the same (all about 95.1%), showing one uniform “keep moving” strategy rather than talent-dependent exploration. Oddly, though, all this movement barely yields more event encounters: the per-capita encounter with lucky events is only 0.21 (Baseline: 2.43) and with unlucky events only 0.19 (Baseline: 2.45), and 74.2% of agents encounter no event at all over their lifetime (Baseline: 15.9%).

Theoretical explanation (see §6.2). This is the inevitable result of the log-utility (Kelly-criterion) optimal strategy. The agent’s observation window shows only “whether an event is present,” not distinguishing lucky from unlucky (each 50%). Under the asymmetric luck structure, unlucky events strike with probability 100% while lucky events strike only with the talent probability $T_i < 1$, so the expected log-return of encountering any unknown event is always negative:

$$\mathbb{E} \left[\ln \frac{C'}{C} \mid \text{encounter} \right] = \frac{1}{2} \cdot T_i \ln 2 + \frac{1}{2} \cdot (-\ln 2) = \frac{\ln 2}{2} (T_i - 1) < 0 \quad \forall T_i < 1.$$

RL spontaneously learns the optimal strategy: move actively to avoid all events, because any encounter lowers the expected log-return. Note that although T_i appears in the formula, its effect only changes the *magnitude* of the negative value ($T_i = 0.7 \rightarrow -0.104$; $T_i = 0.3 \rightarrow -0.243$),

not the *sign*, for every agent with $T_i < 1$ the expected log-return is always negative. RL therefore gets no talent-based strategy signal: whatever the talent, “avoid every encounter” is the one log-optimal choice. This directly predicts the experimental observation: the mobility rates of Q1 to Q5 are exactly the same (all about 95.1%), and the talent–wealth correlation is close to zero ($r \approx 0$). This “keep-moving avoidance” strategy cuts the poverty rate from the Baseline’s 45.7% to 12.9%, but also gives up the upside of lucky events (the median maximum capital is only 160; Baseline: 8.2×10^4).

Table 3 compares the full statistics of Exp.1 and the Baseline.

Table 3. Exp.1 vs. Baseline core statistics (passive Baseline: 100 independent runs; Exp.1: 100 post-convergence episodes)

Category	Statistic	Baseline	Exp.1 (RL)	Change
Wealth dist.	% below C_0 ($\pm 0.1\%$ band)	45.7%	12.9%	↓ 72% (downside protection)
	% equal to C_0 ($\pm 0.1\%$)	24.4%	78.8%	RL-specific mass at C_0
	% above C_0 ($\pm 0.1\%$ band)	29.9%	8.3%	↓ less upside
	Top-20% wealth share	97.3%	32.5%	↓ sharply lower inequality
Power-law	α (R^2), rank–wealth slope	0.32 (0.70)	1.47 (0.49)	tail destroyed; fit fails ($R^2 < 0.5$), mass at C_0
Event encounters	lucky encounters (mean)	2.43	0.21	↓ 91% (active avoidance)
	unlucky encounters (mean)	2.45	0.19	↓ 92%
	fraction of neutral lives (met no event)	15.9%	74.2%	↑ no events is itself a strategic outcome
	mobility rate	0% (fixed)	95.1%	RL active exploration
Talent–wealth	talent–wealth correlation r	+0.063	+0.029	talent’s effect nearly vanishes
	$\bar{T}_{\text{top}10}$ (mean talent, richest 10%)	0.66	0.63	luck still weakly dominant
	C_{mt} (median capital, $T > 0.7$)	10	10	typical talented agent stays at C_0
	$C_{\text{mt}}/C_{\text{at}}$ ratio (medians; $C_{\text{at}}: T - 0.6 \leq 0.05$)	1.0×	1.0×	no median talent advantage

Three emergent findings: (1) *Kelly-optimal avoidance strategy*: without supervision, RL spontaneously converges to the log-utility-optimal “continuous-movement avoidance” strategy, a computational instance of the classic Kelly criterion; (2) *equality-stagnation paradox*: the flat-test wealth distribution (the top 20% hold just 32.5%) and the lowest upward channel ($P(C > C_0) = 8.3\%$) appear together, revealing that “equality” is not “opportunity”; (3) *talent-undifferentiated strategy*: the mobility rates of Q1–Q5 are nearly identical (about 95%), and the incentive distortion almost entirely erases talent’s effect on strategy at the system level.

Social mapping: this strategy of “everyone moving all the time, avoiding every event” matches lying-flat forced out by bad incentives: when the expected return to taking part is negative ($\frac{\ln 2}{2}(T_i - 1) < 0$ for all $T_i < 1$), rational people drop out entirely. This lying-flat is not physical stillness. The agents keep moving almost every step (mobility $\approx 95\%$), but they steer away from events, holding their encounter rate down to 0.5%, one-twelfth of the 6% passive-agent baseline, the rate an agent meets events when it does not steer toward or away from them (the same 6% whether it stays put like the passive Baseline or moves at random). Lying flat here is the active avoidance of exposure, not sitting still; sitting still is instead the passive Baseline, where agents cannot act and simply take whatever luck drifts onto them, closer to resignation than to the deliberate withdrawal RL discovers. This has a social cost on two sides, both visible

above. With almost everyone stuck at C_0 , wealth stops growing and the whole group produces less, so the equality here is the equality of a stalled economy, not of shared wealth. And because the avoidance ignores talent, talent leaves no mark on wealth at all ($r \approx 0$; in Figure 5 and Figure 13 the richest are no more talented than anyone else), so ability has nowhere to go and the most able people are wasted along with everyone else. What this suggests (that the structure of incentives matters more than urging people to try harder, and the parallel with markets that prey on those who take part) is developed in detail in §6.2.

Why introduce symmetric luck: the bridge from Exp.1 to Exp.2

Exp.1 delivered an unexpected result: once given the ability to act, the agents did not chase opportunity but learned to avoid it, because under the original asymmetric luck every encounter carries a negative expected log-return, so withdrawal, not effort, is what pays. The natural question this raises is whether the avoidance is a feature of agency itself, or only of that particular harsh luck structure. The original model makes bad luck always strike, so talent can boost a gain but can never block a loss. Yet a broad literature treats ability as exactly the capacity to spot and sidestep risk, not just to seize opportunity (§3.1): financially literate people make far fewer costly mistakes (Lusardi & Mitchell, 2014), lower cognitive ability raises the rate of credit-market errors (Agarwal et al., 2009), and firms with strong dynamic capabilities cut losses and survive shocks (Teece et al., 1997). We therefore introduce a *symmetric* luck structure, in which a talented agent dodges an unlucky event with probability T_i just as it seizes a lucky one, turning the expected log-return of an encounter positive for the majority. The next two conditions, Baseline2 (passive) and Exp.2 (with agency), test what this fairer structure changes and, decisively, whether it finally lets agency overturn the TvL conclusions.

5.3 Baseline2: TvL’s Conclusions Survive under Symmetric Luck

Baseline2 is the passive control for the symmetric structure: agency is held off and only the luck rule changes. The first question is the same one we ask of every condition, whether TvL’s core conclusions survive, and here the answer is unambiguous, they do. Over 100 independent runs the wealth distribution stays heavy-tailed and highly unequal (the top 20% hold 97.0% of wealth, essentially the Baseline’s 97.3%), and the richest agent is again only near-average in talent ($T^* = 0.69$, against a maximum of about 0.92): making luck fairer does not make the top of wealth the top of talent, and luck still decides who ends up on top.

The one thing that does move, and it will matter for reading Exp.2, is that talent leaves a slightly clearer mark. The talent–wealth correlation rises from $r = +0.063$ to $r = +0.182$ (per-run mean, s.d. 0.031), and the fraction of agents ending below C_0 falls from 45.7% to 28.0%, because the expected log-return of an encounter is now $\ln 2 (T_i - 0.5)$, positive for the majority whose talent exceeds 0.5. This shift is modest, and, decisively, it is achieved with no agency at all: it is a property of the luck structure, not of behavior, and it supplies the passive baseline

against which Exp.2 measures what reinforcement learning adds on top.

5.4 Exp.2: RL under Symmetric Luck, and the Limits of Agency

Exp.1 showed that under asymmetric luck avoiding all events is best, because unlucky events strike 100% of the time and make the expected log-return of any encounter negative. Exp.2 makes luck symmetric and talent-dependent (lucky events trigger with T_i , unlucky with $(1 - T_i)$), which removes that negative bias, and tests whether the talent–wealth channel reopens.

The expected log-return of an encounter:

$$\mathbb{E}\left[\ln\frac{C'}{C} \mid \text{encounter}\right] = \frac{1}{2} \cdot T_i \cdot \ln 2 + \frac{1}{2} \cdot (1 - T_i) \cdot (-\ln 2) = \ln 2 (T_i - 0.5).$$

It is positive when $T_i > 0.5$, negative when $T_i < 0.5$, and exactly zero at $T_i=0.5$. At the mean talent ($T_i = 0.6$) the expected return is about +0.069, positive but tiny. Most agents lie in the positive-return region, so RL receives a signal pointing toward events, and it acts on that signal: the mature policy is not a random walk. Agents land on an event in about 24% of steps, roughly four times the 6.0% passive-agent baseline, which shows the population actively steers toward opportunity rather than drifting. This is a whole-population shift, not a talent-sorted one: mobility is a near-uniform 95% and the seeking is about equally strong across talent. The talent–wealth correlation $r = +0.317$ then comes from the conversion channel: because active seeking lifts everyone’s exposure to multiplicative events, the per-encounter advantage of higher talent (a higher rate of doubling and of dodging, $0.75(1 + T_i)$ in arithmetic terms) compounds into a wider wealth gap than it could in the passive Baseline2, where agents met events only by chance.

The results of Exp.2 are shown in Table 4, with three key findings.

Table 4. Core wealth-outcome statistics across the four experiments (passive conditions: 100 runs; RL: 100 post-convergence episodes; seed=2026)

Experiment	Below C_0	At C_0	Above C_0	$r(\text{tal.} - \text{wea.})$	Mobility
Baseline (passive, asymmetric)	45.7%	24.4%	29.9%	+0.063	0.0%
Baseline2 (passive, symmetric)	28.0%	32.5%	39.6%	+0.182	0.0%
Exp.1 (RL, asymmetric)	12.9%	78.8%	8.3%	+0.029	95.1%
Exp.2 (RL, symmetric)	35.0%	5.3%	59.7%	+0.317	95.3%

(1) The talent–wealth correlation re-emerges from zero. In Exp.1, RL’s capital-preserving strategy left almost all agents at C_0 , and talent no longer predicted wealth ($r \approx 0$). In Exp.2 the negative bias in the expected log-return is removed, so agents lose the rational basis for avoiding events and instead actively seek them (encounter rate about 24% of steps, four times the 6% passive-agent baseline), while mobility stays around 95%. Since both lucky and unlucky

events trigger with the talent probability T_i , high-talent agents trigger more often in both directions, enjoying higher variance under the symmetric expectation, which drives the talent–wealth correlation from ≈ 0 in Exp.1 up to $r = +0.317$.

The move-preference metric $\Delta \approx -0.005$ stays near zero, but this does not mean the population ignores events: Δ tracks only whether an agent changes its move rate when an event comes into view, not the direction it heads, and the fourfold rise in encounter rate over the passive-agent baseline shows the population steers toward events rather than waiting for them. The correlation comes from talent’s higher conversion rate compounding over this denser stream of encounters. The key comparison is with Baseline2, which shares the identical symmetric-luck structure but holds agency off: there the correlation is already $+0.182$, so of the full $+0.317$ under Exp.2, a large part comes from the luck structure itself and the rest from agency, which lifts the whole population’s exposure to events and so boosts talent’s conversion advantage.

(2) Extreme inequality coexists with explosive wealth. Inequality is near-complete, with the top 20% holding 100% of wealth; meanwhile the median $C_{\max} = 1.34 \times 10^9$ (16,400 times the Baseline’s 8.2×10^4). This is a characteristic feature of a log-multiplicative process: although each encounter carries only a small positive expected log-return for most agents ($\ln 2 (T_i - 0.5)$, about $+0.07$ at the mean), its arithmetic expectation grows much faster with talent (a per-encounter factor of $0.75(1 + T_i)$), so a few high-talent agents who meet many events run away to extreme wealth, producing a power-law tail of “most are mediocre, a few winners are extremely rich.” The mean talent of the richest 10% is $\bar{T}_{\text{top}10} = 0.73$ (vs. Baseline 0.66), and those with talent > 0.8 make up 35% of the top performers.

(3) The protective effect of talent is restored. Among the high-talent group ($T > 0.7$), 76.8% end with capital above C_0 (Baseline only 34.2%; per-episode means). The typical talented agent finally pulls ahead too: the group’s *median* capital rises to about 1280 (roughly $128 \times C_0$), whereas in the Baseline the median high-talent agent ends right at C_0 , no better off than average. A few talented agents do far better and pull the group’s *mean* capital up to about 1.2×10^8 , but that mean reflects a handful of explosive winners, not the typical talented life. This contrasts with Exp.1: there talent’s “help” showed only in more precise capital preservation, whereas in Exp.2 talent, through the variance advantage of the multiplicative process, translates directly into exponential wealth differences. Talent here does two things. It lifts the top, as the exponential wealth above shows; and it protects the bottom, since dodging an unlucky event with probability T makes the more talented much less likely to fall below their starting capital, so talent also helps a person hold onto what they began with.

Under symmetric luck the expected log-return of an encounter turns positive for the talented majority, so the rational strategy flips: the same learning rule that taught withdrawal in Exp.1 now teaches engagement, and the population abandons lying-flat for active seeking. Three emergent findings capture this: (1) *Reactivation of the talent channel*: without external guidance, RL spontaneously produces a significant effect of talent on wealth, an emergent outcome beyond the model’s design; (2) *Population-wide opportunity-seeking*: the population actively steers toward

events (encounter rate about four times the 6% passive-agent baseline), a whole-population shift with no talent sorting (mobility a near-uniform 95%); (3) *Talent-variance boosting mechanism*: because agency lifts everyone’s exposure to multiplicative events, the higher-talent agent, by its arithmetic-expectation advantage $0.75(1 + T_i)$, pulls ahead exponentially, so a shared strategy still produces “unequal outcomes.”

Social mapping: even with agents actively seeking opportunity, the top of wealth still goes to the luckiest, not the most talented. Once incentives are repaired (symmetric luck makes $\ln 2 (T_i - 0.5)$ positive for most), talent predicts wealth a little more clearly, because active seeking gives everyone more events and so boosts talent’s *conversion-rate* advantage (r rises from ≈ 0 to $+0.32$), yet the richest agents remain those who met the most good luck. Even under agency and fairer luck combined, in other words, TvL’s verdict holds: luck decides who ends up on top.

Table 5 compares the full statistics of Exp.2 and Baseline2, using the same protocol, data source, and layout as the Exp.1-vs-Baseline table (Table 3) so the two pairs are directly comparable. One asymmetry in that table is worth a note: in the passive conditions lucky and unlucky encounters are equal by construction (both about 2.44), but Exp.2 agents meet more lucky than unlucky events (11.85 vs 8.36). They still cannot see an event’s type, and the ratio (about 1.42) is stable across all four seeds. The one-step capital change is the only channel that could produce it, and the data fit that reading: an agent next to a lucky event is still next to it one step later 32.1% of the time, against 29.1% for an unlucky one (Exp.1: 25.6% against 18.8%). That is a win-stay, lose-shift response to its own gains and losses, not any sight of what the event is. Note that sitting still is not how an agent holds on to an event here, since events move two cells per step as well, so following one means moving. This per-step gap is small and does not by itself account for the full lifetime ratio, so we report the asymmetry as a stable and only partly explained side effect of ex-post feedback rather than a settled mechanism. It changes none of the conclusions.

Table 5. Exp.2 vs. Baseline2 core statistics (passive Baseline2: 100 independent runs; Exp.2: 100 post-convergence episodes); same protocol and data source as Table 3

Category	Statistic	Baseline2	Exp.2 (RL)	Change
Wealth dist.	% below C_0 ($\pm 0.1\%$ band)	28.0%	35.0%	$\uparrow$ RL accepts more downside
	% equal to C_0 ($\pm 0.1\%$)	32.5%	5.3%	$\downarrow$ almost none stay at C_0
	% above C_0 ($\pm 0.1\%$ band)	39.6%	59.7%	$\uparrow$ majority end above C_0
	Top-20% wealth share	97.0%	100%	concentration stays near-complete
Power-law	α (R^2), rank-wealth slope	0.51 (0.87)	0.21 (0.91)	$\downarrow$ fatter power-law tail
Event encounters	lucky encounters (mean)	2.44	11.85	$\uparrow 4.9\times$ (active seeking)
	unlucky encounters (mean)	2.44	8.36	$\uparrow 3.4\times$ (seeking raises exposure)
	fraction of neutral lives (met no event)	16.0%	0.0%	$\downarrow$ no event-free lives
	mobility rate	0% (fixed)	95.3%	RL active exploration
Talent-wealth	talent-wealth correlation r	+0.182	+0.317	$\uparrow$ agency strengthens talent channel
	$\bar{T}_{\text{top}10}$ (mean talent, richest 10%)	0.66	0.73	$\uparrow$ the richest become more talented
	C_{mt} (median capital, $T > 0.7$)	10	1280	talent lifts the typical agent only here
	$C_{\text{mt}}/C_{\text{at}}$ ratio (medians; $C_{\text{at}}: T - 0.6 \leq 0.05$)	$1.0\times$	$16\times$	median advantage appears only here

Having run all four conditions, we collect them into a single master table (Table 6) that reads the whole two-by-two design along both of its axes; it is also the comprehensive reference for the cross-experiment synthesis that follows, so all four-condition comparisons below draw on it.

Table 6. Core performance metrics across the 2×2 factorial conditions.

Condition	Top-20%	% $<C_0$	$P(C>C_0)$	Enc.%	$r(T, \log C)$	$\bar{T}_{\text{top}10}$	α (R^2)	$C(T_{\text{max}})$	Core mechanism
Original TvL reference	80%	$\approx 50\%$	n/a	n/a	≈ 0	n/a	power-law	n/a	Luck dominates (Pluchino et al., 2018)
Baseline (passive, asym.)	97.3%	45.7%	29.9%	6.0	+0.063	0.66	0.32 (0.70)	10	Luck dominates; fixed agents
Baseline2 (passive, sym.)	97.0%	28.0%	39.6%	6.0	+0.182	0.66	0.51 (0.87)	10	Symmetric luck opens talent channel
Exp.1 (RL, asym.)	32.5%	12.9%	8.3%	0.5	+0.029	0.63	1.47 (0.49*)	10	Active avoidance (lying-flat)
Exp.2 (RL, sym.)	100%	35.0%	59.7%	24	+0.317	0.73	0.21 (0.91)	2.05×10^4	Active seeking opens talent channel

Note: Metrics report per-realization means over 100 runs (passive) or 100 post-convergence episodes (RL), except $C(T_{\text{max}})$ (100-realization median) and α (R^2 , pooled fit). Initial capital $C_0 = 10$. Enc.%: share of steps near/on an event (passive baseline $\approx 6.0\%$). $\bar{T}_{\text{top}10}$: mean talent of top 10% by wealth (population mean = 0.60, max ≈ 0.92). α : OLS rank-wealth log-log slope (* $R^2 < 0.5$ indicates fit failure due to wealth clustering at C_0). $C(T_{\text{max}})$: final capital of the most talented agent.

Comparing the two rows within a luck block isolates what agency (RL) buys. Under asymmetric luck it buys lying-flat: agents drive their encounter rate from the 6% passive-agent baseline down to 0.5%, upward mobility falls from 29.9% to 8.3%, and the talent channel stays shut (r : 0.06 $\rightarrow$ 0.03). Under symmetric luck the identical learning rule buys the opposite, active

seeking: the encounter rate rises to 24%, upward mobility climbs from 39.6% to 59.7%, and the talent channel opens further (r : $0.18 \rightarrow 0.32$). The same RL algorithm produces opposite population behavior, fixed only by the sign of the expected return.

Comparing the asymmetric block with the symmetric block isolates what a fairer luck structure buys, and the decisive step is Baseline2 to Exp.2. Symmetric luck alone, with agents still frozen, already opens half the talent channel (r : $0 \rightarrow 0.18$) simply by letting talent dodge misfortune, a gift of the structure that costs no effort. Switching agency on then opens the rest ($0.18 \rightarrow 0.32$) by pushing the whole population from passively waiting for luck to actively seeking it, which lifts upward mobility by about half ($39.6\% \rightarrow 59.7\%$). Yet across all four cells the rich’s mean talent never exceeds 0.73, far short of the most talented (≈ 0.92): whatever agency and a fairer structure add, the summit still belongs to luck.

5.5 Cross-Experiment Synthesis: The Four Experiments Jointly Confirm and Sharpen the Core TvL Conclusions

Across all four conditions together, the improved model does not overturn TvL’s core conclusion that luck dominates success. Instead, the four conditions make it stronger in two ways: they confirm it and they sharpen it. It holds under every condition, so it is not a fluke of any single experiment.

5.5.1 TvL confirmed: in every realization, the richest are always the luckiest

For each experimental condition, within each independent realization (the 100 runs of the Baseline, the 100 post-convergence episodes of the RL experiments), we rank by final wealth to find the 10 richest and 10 poorest agents, count the lucky and unlucky events each *encounters*, and then aggregate across 100 realizations. The core conclusion holds under all four experimental conditions: the net good luck encountered by the 10 richest is always positive and that of the 10 poorest always negative. In Exp.1 the separation is real but small, since lying-flat leaves almost everyone near C_0 ; the richest are still the luckiest, only by a slim margin. Ranking within each realization (rather than “averaging first, then ranking”) is the key method choice: it directly tests whether “in each life realization, the one who reaches the top is the luckiest in that realization,” and holds up with small cross-realization error bars. Figure 5 shows this as a continuous map: across the whole talent–wealth plane, net good luck rises with wealth at every talent level.

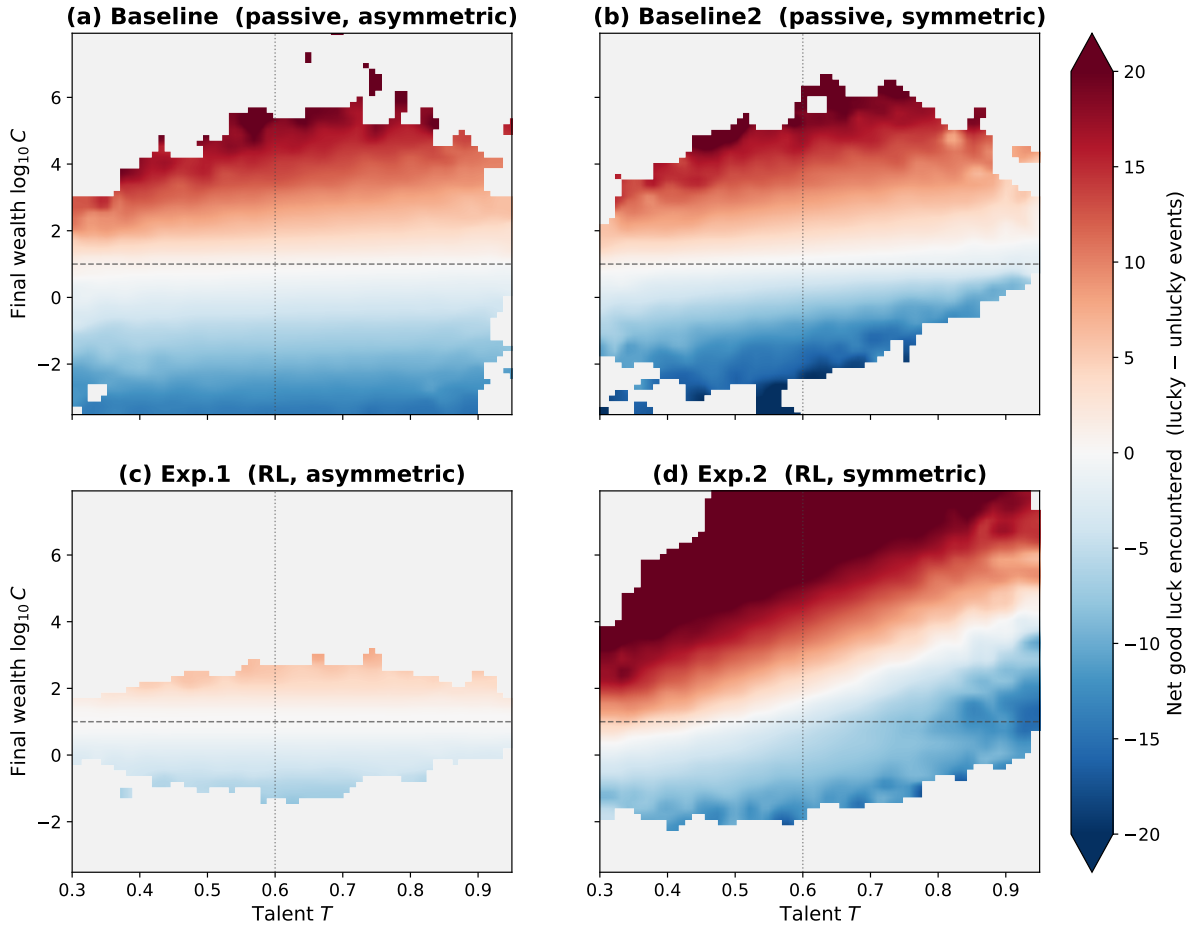
Net luck across talent and final wealth: the wealthy are the lucky, at every talent level

Figure 5. Net good luck across talent and final wealth, under all four conditions. For each condition the plane is binned by talent (x) and final $\log_{10}$ capital (y), and each cell is colored by the mean net good luck of the agents in it (lucky minus unlucky events met; red positive, blue negative), pooled over all realizations and smoothed for a continuous view. The dashed horizontal line is the initial capital $C_0 = 10$ ($\log_{10} = 1$); the dotted vertical line is the mean talent 0.6. In every condition the color tracks the vertical (wealth) axis far more than the horizontal (talent) axis: at any given talent, richer agents met more good luck and poorer agents met less. Exp.1 (c) collapses near C_0 with pale colors, the mark of avoidance; Exp.2 (d) shows a diagonal in which talent tilts the boundary, yet wealth still tracks luck.

We count the encounters of events, not their triggers. Because capital evolution satisfies the identity $\log_2 C = \log_2 C_0 + (\text{lucky triggers}) - (\text{unlucky triggers})$, plotting “trigger counts” would be circular reasoning; “encounter counts,” by contrast, are pure luck exposure, since meeting an event does not always turn into wealth (it still has to be converted by talent T), so “the richest meet more good luck” is a genuine, testable conclusion. In the Baseline panel of Figure 6, the 10 richest and 10 poorest have almost identical talent, yet differ by about 28 in net good luck, precisely TvL’s most counterintuitive finding, which we fully reproduce and reinforce with stricter per-realization statistics.

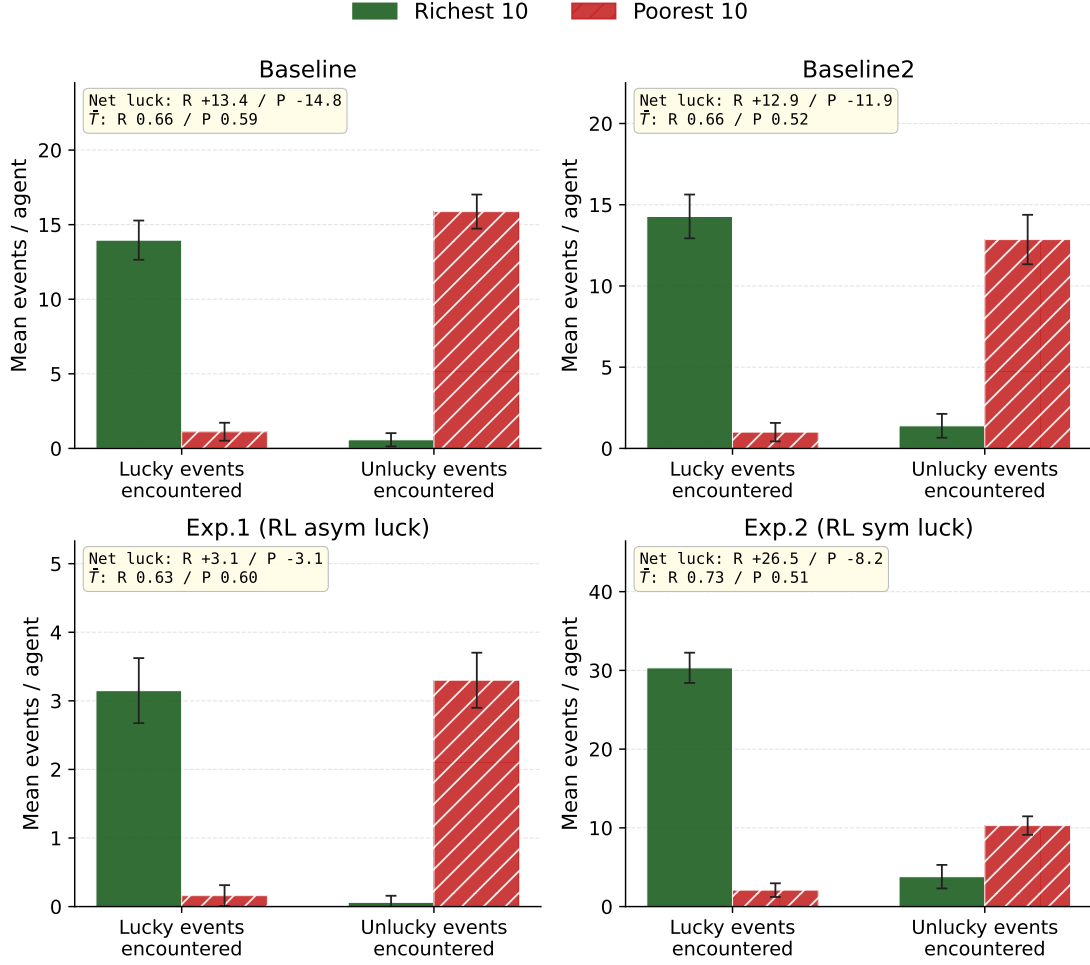


Figure 6. Lucky/unlucky event encounters of the 10 richest vs. 10 poorest in each experiment (mean $\pm$ standard deviation, 100 realizations). In each cell the green bars (richest) dominate on lucky events and the red bars (poorest) on unlucky events. The annotation box gives the net good luck and the two groups' mean talent $\bar{T}$. The Baseline column is especially key: the richest and poorest have almost identical talent ($\bar{T} = 0.66$ vs 0.59), yet differ by about 28 in net good luck, pure luck dominance, consistent with the original TvL.

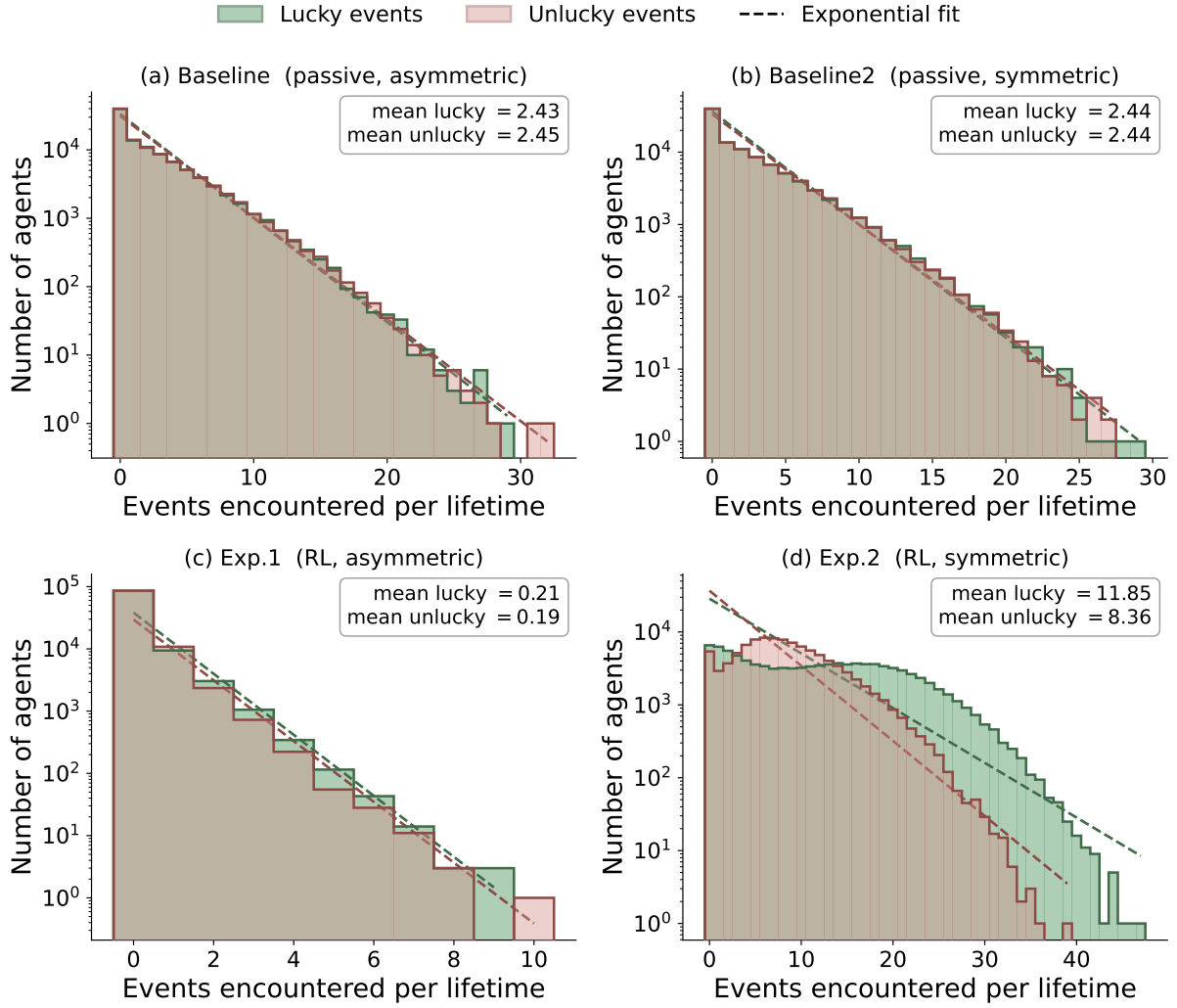


Figure 7. Distributions of lifetime event encounters per agent (lucky: green; unlucky: red; log-scale counts) under all four conditions. Panels, arranged as the two-by-two factorial grid: (a) Baseline (passive, asymmetric luck); (b) Baseline2 (passive, symmetric luck); (c) Exp.1 (RL, asymmetric luck); (d) Exp.2 (RL, symmetric luck), so that reading down each column isolates the effect of agency at a fixed luck structure. Passive conditions pool 100 runs; RL panels pool 100 post-convergence episodes. Insets report the mean number of lucky and unlucky encounters per lifetime.

5.5.2 TvL sharpened: talent is the “conversion rate” of luck, and its zero correlation is only a masking effect

TvL’s most cited conclusion is that talent and wealth are almost uncorrelated ($r \approx 0$) and the richest are not the most talented. Our model reproduces this (Baseline $r = +0.06$, Table 6). But logging encounters and triggers separately lets us take the appearance apart and reveal the talent effect that the spread of luck hides.

Figure 8(A) reproduces TvL’s zero correlation; but this is only because the variance of luck exposure drowns out the talent effect. Figure 8(B) shows that, once luck is controlled for, talent immediately predicts wealth clearly, with a regression slope equal to the number of lucky events encountered. This is not an independent empirical result but a direct consequence of the multiplicative mechanism: each lucky event doubles capital with probability T , so under k

encounters the expected slope of $\log_2$ wealth on talent is exactly k . The measured slope only confirms the bookkeeping is right.

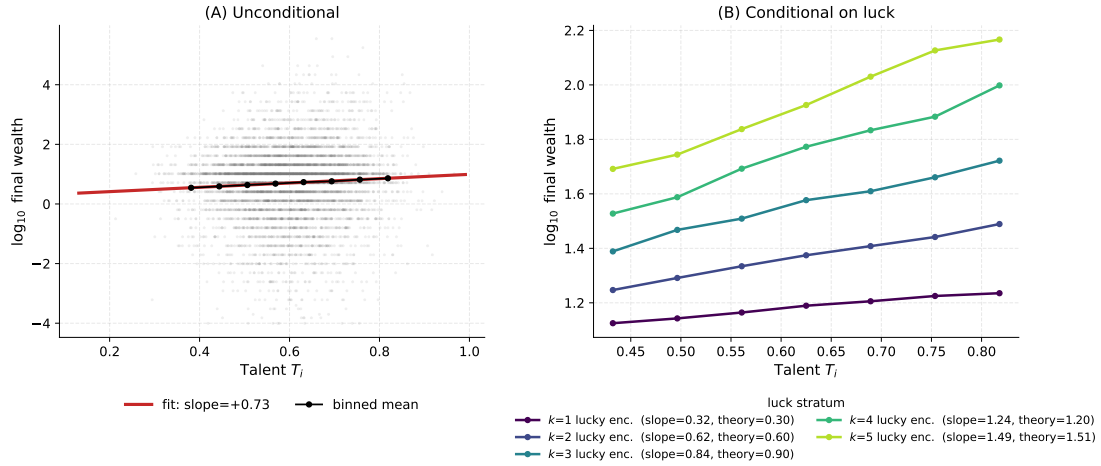


Figure 8. Unmasking talent: TvL’s zero correlation is only how things look before luck is accounted for. (A) Pooling all agents in the Baseline, talent and wealth are almost uncorrelated ($r = +0.06$), i.e., TvL’s classic conclusion. (B) Once luck is controlled (comparing only agents who encountered the same number k of lucky events and zero unlucky), talent immediately shows clear positive predictive power, with a regression slope exactly equal to $k \cdot \log_{10} 2$, each lucky event doubles capital with probability T , so under k encounters the slope of $\log_2$ wealth on talent is exactly k . As a consistency check, the measured slope matches this identity ($k=5$: measured 1.49, predicted 1.51).

The micro-mechanism is direct: in the symmetric-luck Exp.2, the rate at which an agent converts “encountered good luck” into “actual doubling,” and the rate at which it “dodges” “encountered bad luck,” are both exactly equal to its talent value T .

Across the two symmetric-luck conditions, for the majority of “ordinary people” who encounter typical (modal) luck, talent still significantly and positively predicts wealth. This is the paper’s first real reinforcement of TvL, and it fits in one sentence: luck decides how many opportunities arrive; talent decides how many of them are seized. TvL’s “talent-irrelevance” is only how it looks before luck is accounted for; once luck exposure is controlled for, talent is always the real force behind who ends up where, and the mechanism is exactly luck’s conversion rate. Reading talent as the rate at which luck is converted is not new in itself. It is how talent enters the CEO-pay model of Gabaix & Landier (2008), and Challet et al. (2020) showed analytically how multiplicative lucky events turn a Gaussian talent distribution into a heavy-tailed one. What we add is the measurement: logging encounters and triggers separately lets us hold luck exposure fixed and show what the conversion rate is worth inside the TvL world itself. Our model, rather than shaking TvL, offers a way to read it: the zero correlation is a masking effect, and talent’s role is exactly luck’s conversion rate.

5.5.3 Cross-Experiment Comparison of Wealth Distribution and Talent Structure

This subsection returns to TvL’s two most central observations, the extreme inequality of wealth and that the richest are not the most talented, and checks whether they still hold across the

four conditions, while showing how RL agency reshapes the talent structure of wealth without overturning either.

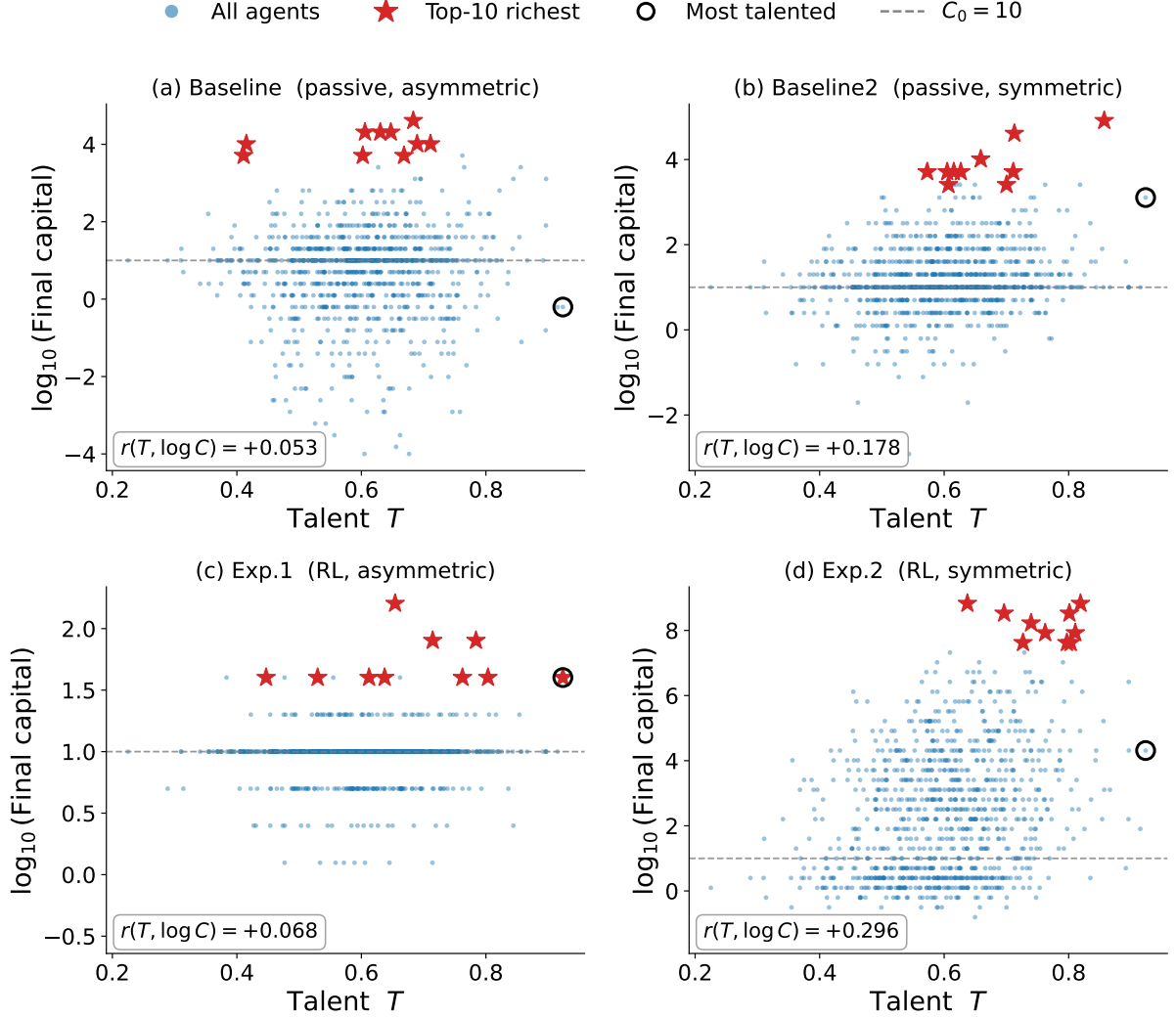


Figure 9. Talent versus final log-capital under all four experimental conditions. Panels, arranged as the two-by-two factorial grid: (a) Baseline (passive, asymmetric luck); (b) Baseline2 (passive, symmetric luck); (c) Exp.1 (RL, asymmetric luck); (d) Exp.2 (RL, symmetric luck), so that reading down each column isolates the effect of agency at a fixed luck structure. Blue dots: all 1,000 agents (Baseline: run 0; RL: final post-convergence episode). Red stars: the ten richest agents; open circle: the most talented agent; dashed line: initial capital $C_0 = 10$. Each panel reports the Pearson correlation $r(T, \log C)$ for the plotted realization. In every condition the richest agents are usually not the most talented, with the distribution over 100 realizations given in Figure 13, while the talent–wealth correlation strengthens under symmetric luck: from +0.05 (Baseline) to +0.18 once luck alone is made symmetric (Baseline2), and to +0.30 once agency is added (Exp.2). The corresponding population means over all realizations are +0.06, +0.18, and +0.32 (Table 6). Panel (c) plots one Exp.1 realization, where about 79% of agents are tied at C_0 and the wealth ranking is therefore close to arbitrary.

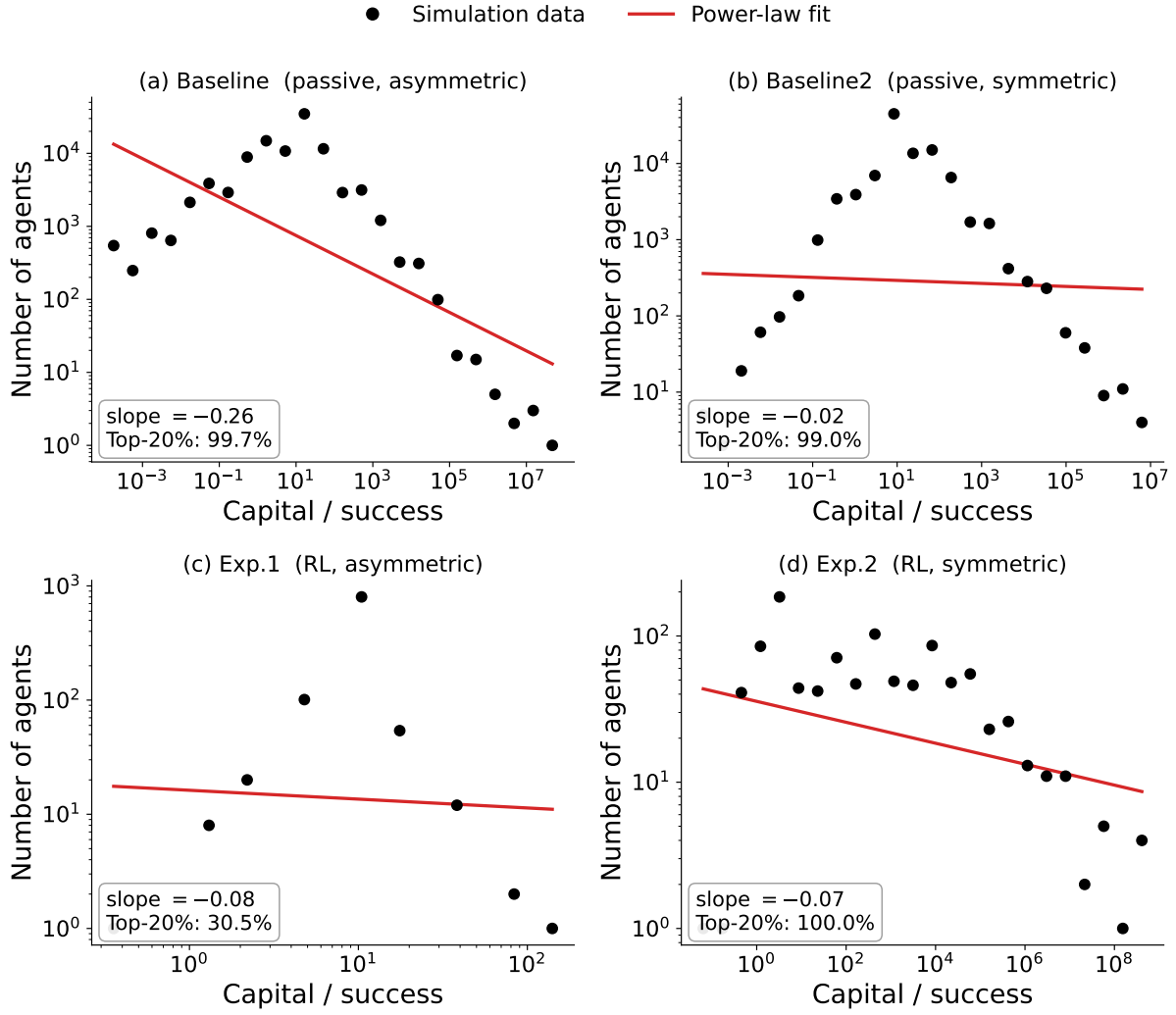


Figure 10. Final-capital distributions on log-log axes with power-law fits, under all four conditions. Panels, arranged as the two-by-two factorial grid: (a) Baseline (passive, asymmetric luck); (b) Baseline2 (passive, symmetric luck); (c) Exp.1 (RL, asymmetric luck); (d) Exp.2 (RL, symmetric luck), so that reading down each column isolates the effect of agency at a fixed luck structure. The passive conditions (Baseline, Baseline2) pool 100 independent runs; the RL panels use the final post-convergence episode. Insets report the fitted slope and the top-20% wealth share. The heavy-tailed character of the wealth distribution is reproduced in every condition except Exp.1, where the distribution collapses onto C_0 .

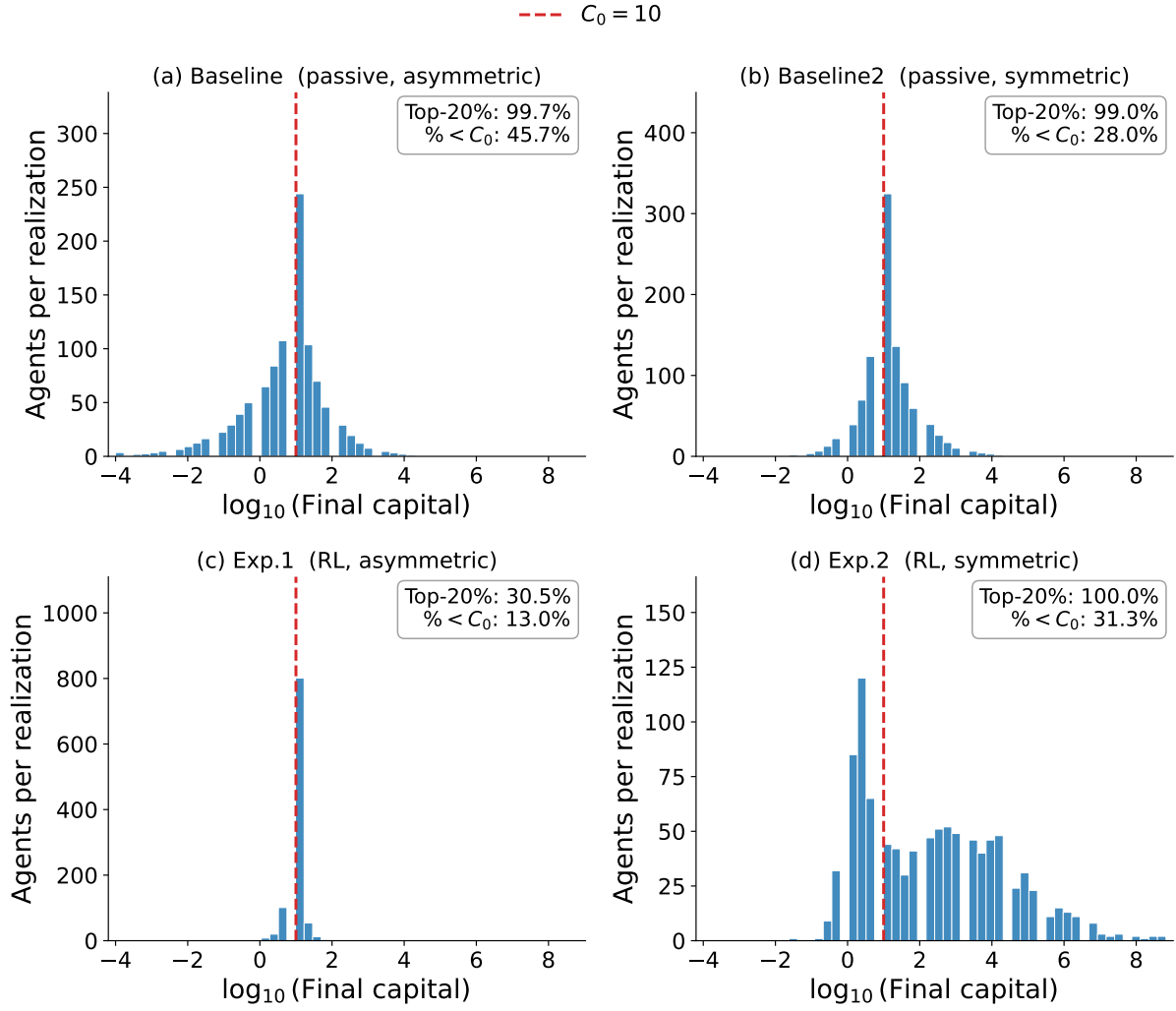


Figure 11. Histograms of final log-capital under all four conditions. Panels, arranged as the two-by-two factorial grid: (a) Baseline (passive, asymmetric luck); (b) Baseline2 (passive, symmetric luck); (c) Exp.1 (RL, asymmetric luck); (d) Exp.2 (RL, symmetric luck), so that reading down each column isolates the effect of agency at a fixed luck structure. All panels share identical axes: the horizontal axis spans the same $\log_{10}$ -capital range, and the vertical axis reports the number of agents per realization (the pooled passive-condition counts are divided by their 100 runs), so that the four distributions are directly comparable. Dashed red line: initial capital $C_0 = 10$. Insets report the top-20% wealth share and the fraction of agents ending below C_0 .

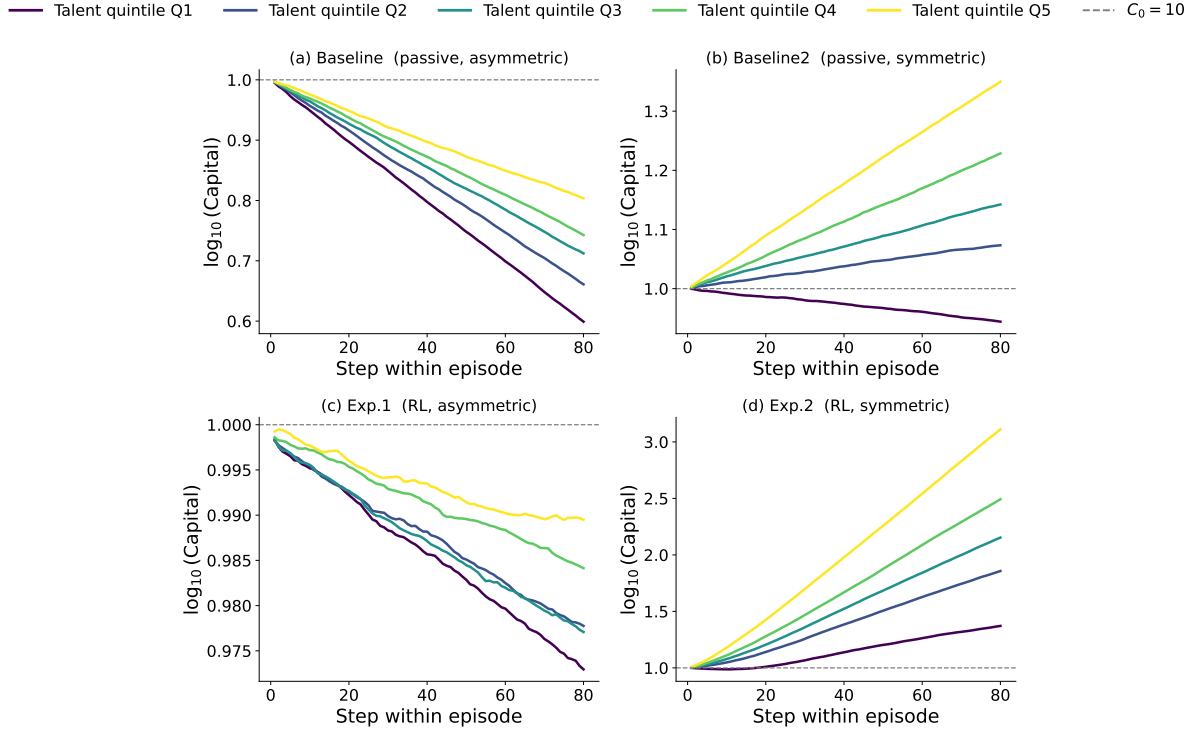


Figure 12. Within-episode capital trajectories by talent quintile (mean $\log_{10}$ capital) under all four conditions. Panels, arranged as the two-by-two factorial grid: (a) Baseline (passive, asymmetric luck); (b) Baseline2 (passive, symmetric luck); (c) Exp.1 (RL, asymmetric luck); (d) Exp.2 (RL, symmetric luck), so that reading down each column isolates the effect of agency at a fixed luck structure. Passive conditions (Baseline, Baseline2): per-run quintile means averaged over 100 independent runs; RL panels: means over 100 post-convergence episodes. Dashed line: $C_0 = 10$.

Figure 13 provides a direct cross-experiment test of TvL’s claim that “the richest are not the most talented.” The conclusion has two layers: (i) what holds in every condition, the mean talent of the top 10% of wealth (0.63–0.73) is only slightly above the population mean 0.6 and systematically far from the highest talent (≈ 0.92 –0.99), so the top of wealth is rarely identical to the top of talent; (ii) what changes with condition, the size of this “rich talent premium”: in the asymmetric-luck Exp.1 the premium is essentially zero (the rich and the population share a mean talent of about 0.60, because lying-flat leaves talent no room to act), while once luck is symmetric the rich’s mean talent rises, most strongly when agency is also switched on (Exp.2, a top-10 mean talent of 0.73). In other words, agency cannot make “the most talented necessarily the richest,” but it does give talent a measurable, nonzero advantage in the competition for wealth.

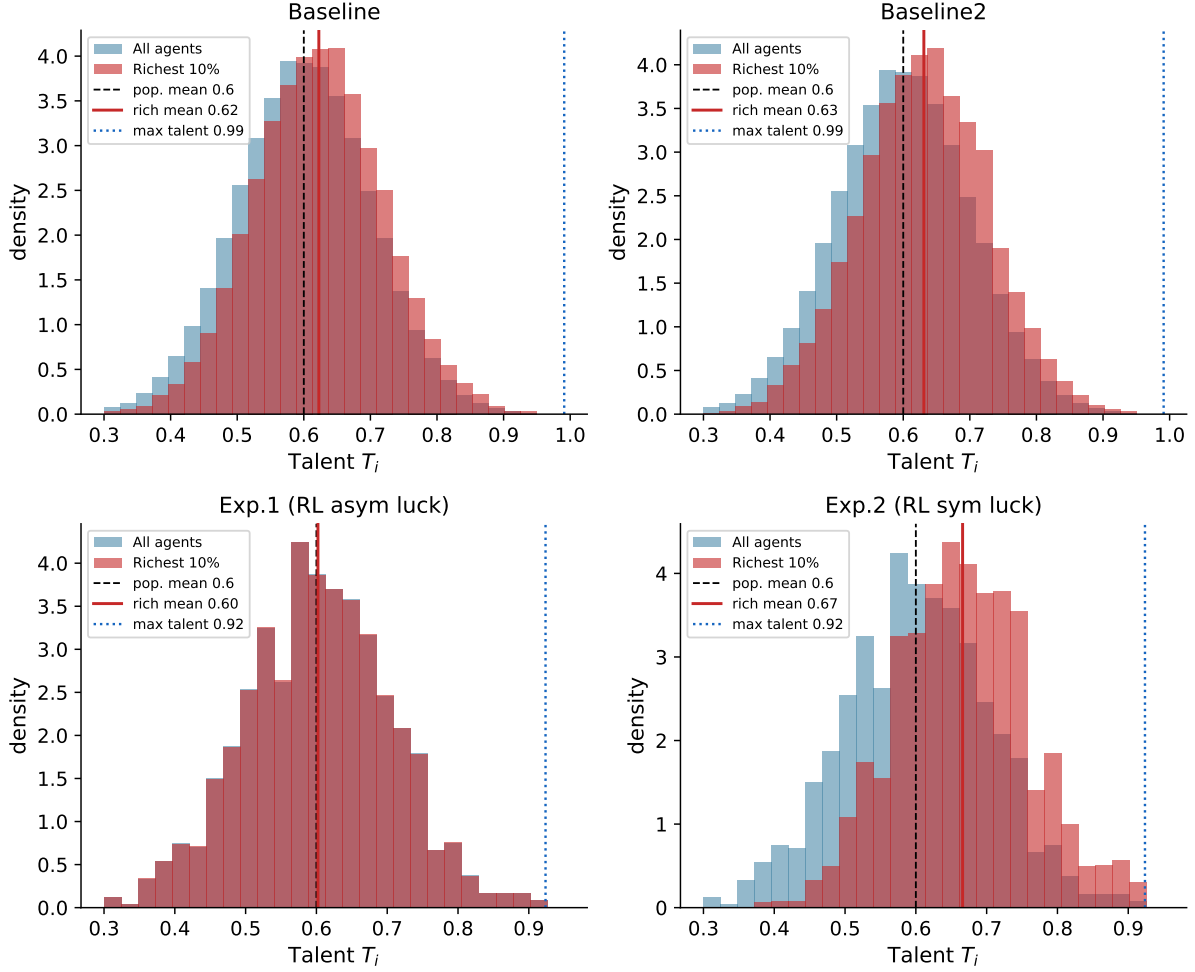


Figure 13. The richest 10% are not the most talented, the talent distribution of the top-10%-wealth agents under each condition. Grey is the overall talent distribution, red the talent distribution of the top 10% by wealth; the black dashed line is the population mean $\bar{T} = 0.6$, the red solid line the rich’s mean talent, the blue dotted line the overall highest talent. Under all four conditions, the rich’s mean talent stays only slightly above the population mean (roughly 0.63–0.73) and far below the overall highest talent (0.92–0.99), the top of wealth is only rarely occupied by the top of talent, and TvL’s signature conclusion holds robustly in every experiment. The luck structure leaves a quantifiable imprint: relative to the Baseline, making luck symmetric lifts the rich’s mean talent already in the passive Baseline2 and further in Exp.2, whereas in the asymmetric-luck Exp.1 it falls back to the population mean, “lying-flat” rendering talent wholly ineffective.

Figure 14 further backs up the above gradient through distribution shape: here we do not control luck but directly show how the marginal wealth distribution shifts with talent quantile: the five Baseline violins almost coincide, backing up TvL’s unconditional zero correlation; the Baseline2 violins already show a mild lift from Q1 to Q5, symmetric luck alone tilting the distribution toward talent; the Exp.1 violins all collapse to C_0 ; while Exp.2 shows a pronounced monotonic lift and widening from Q1 to Q5, the high-talent quantiles having not only higher medians but also longer upper tails, talent both raising the central tendency of wealth and boosting its upside potential.

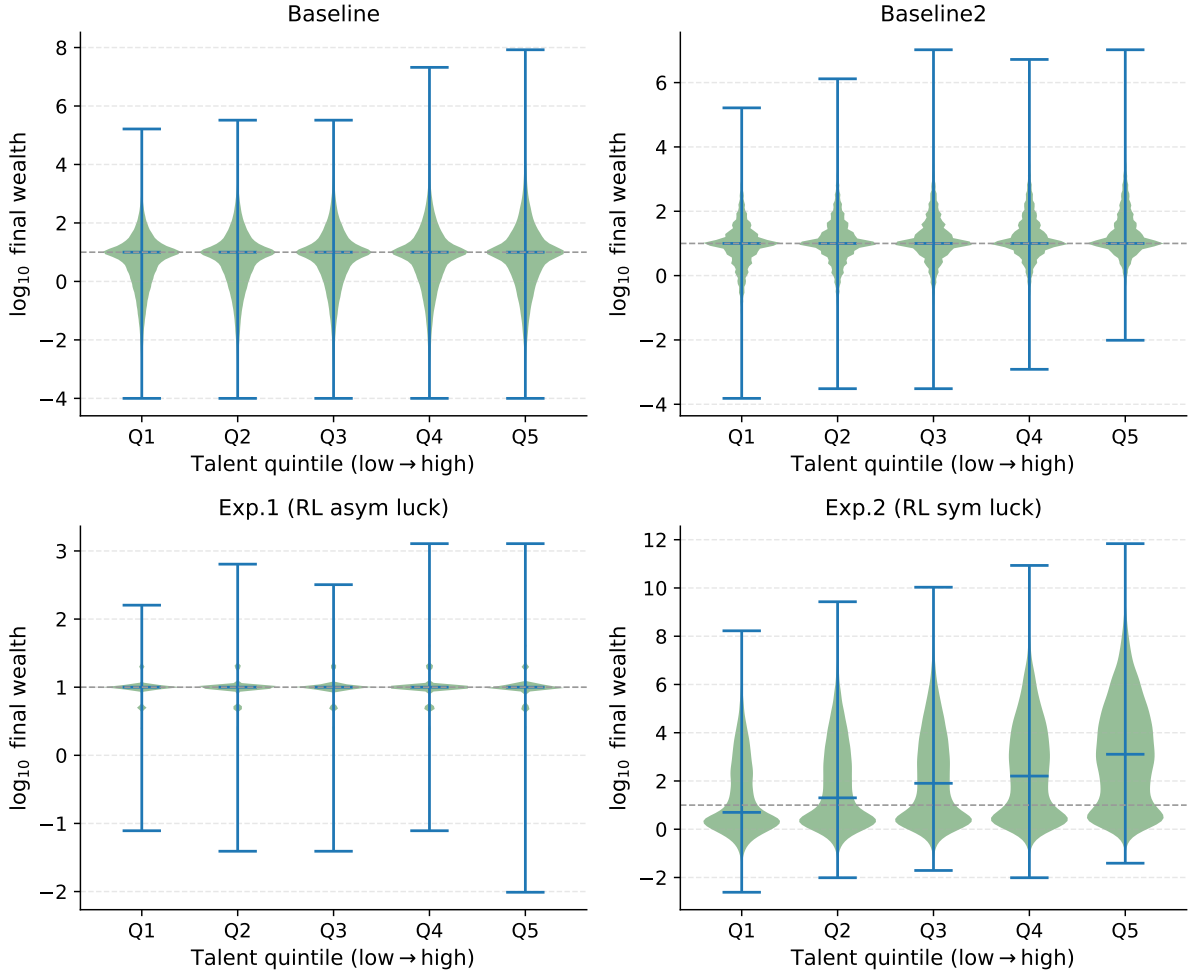


Figure 14. Distribution of final wealth ($\log_{10}$) by talent quantile (Q1 lowest $\rightarrow$ Q5 highest) under each condition. Grey dashed line is $C_0 = 10$. Baseline: the five quantiles’ distributions almost completely overlap, unconditionally, talent is unrelated to wealth (the TvL appearance). Exp.1 (asymmetric): all quantiles collapse to the C_0 line, lying-flat erases talent differences. Baseline2 (passive, symmetric): a mild upward gradient already appears, symmetric luck alone tilting wealth toward talent even without agency. Exp.2 (RL, symmetric): a clear monotonic gradient, the median rising from $\log_{10} C \approx 0.7$ at Q1 to ≈ 3.1 at Q5 (about a 250-fold wealth gap), the talent gradient fully open once agency is added. This is the quintile-violin view complementary to the main-text talent–wealth scatter and rich’s-talent figures.

A cross-experiment measure of wealth inequality: the P99/P10 ratio. To put these differences on one inequality measure, we compute the ratio of the top 1% to the bottom 10% of wealth (P99/P10) under each condition: 1.6×10^4 for the Baseline, falling to 2.1×10^3 for Baseline2 (symmetric luck lifts the poorest decile off the floor, so the same top wealth now spans a narrower ratio); in asymmetric-luck Exp.1 it plunges to just 8 (the avoidance strategy locks almost everyone near C_0 , nearly eliminating inequality); and in symmetric-luck Exp.2 it jumps to 1.7×10^7 (multiplicative growth produces extreme spread under positive-expectation incentives). This single measure captures the two faces of agency: it can both remove inequality (Exp.1, at the cost of universal stagnation) and magnify it (Exp.2, where the majority in fact ends better off, 59.7% above C_0 against 39.6% in Baseline2, but the share falling below C_0 also rises from 28.0% to 35.0%).

5.5.4 Cross-Experiment Comparison of the Emergent Strategy: Lying-Flat versus Seeking

This subsection turns to the social meaning of the group strategy RL finds under each luck structure. The contrast is stark and covers the whole population: under asymmetric luck the agents drive their event-encounter rate down to 0.5%, one-twelfth of the 6% passive-agent baseline, steering away from events; under symmetric luck they lift it to about 24%, four times the baseline, seeking them out. In neither case does the strategy sort by talent, high- and low-talent agents move about equally; the detailed mechanism and social meaning of each are given in the corresponding experiment sections (§5.2, §5.4). Figure 15 makes the contrast concrete: under an identical reward function and an identical $\approx 95\%$ mobility, the two luck structures push the population to opposite ends on every behavioral axis, avoidance versus seeking, stagnation versus mobility, a closed versus an open talent channel, which is the sharpest single piece of evidence that lying flat is a rational response to the reward-risk balance, not laziness or a fixed trait.

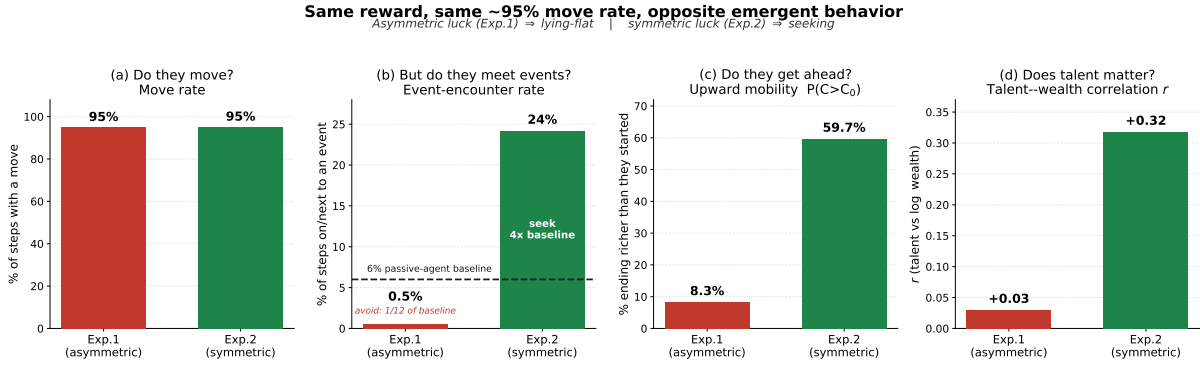


Figure 15. Same reward, opposite behavior, the mechanism behind lying-flat. Both RL conditions optimize the identical reward (the clipped single-step log-return), yet the sign of the expected return alone flips the emergent population strategy. (a) Move rate: both conditions stay near 95%, so agents move almost every step in either world. (b) Event-encounter rate: under asymmetric luck (Exp.1) agents drive it down to 0.5%, one-twelfth of the 6% passive-agent baseline (steering away from events), while under symmetric luck (Exp.2) they lift it to 24%, four times the baseline (seeking them out); so both move constantly, but one steers away from events and the other toward them. (c) Upward mobility $P(C > C_0)$: 8.3% versus 59.7%. (d) Talent-wealth correlation: +0.03 versus +0.32. Because the reward never changes, the reversal is driven by the payoff structure, not by any change in the agents’ preferences, the core evidence that lying-flat is taught by a negative-expectation payoff rather than encoded as laziness.

Table 6 shows that the luck structure determines both whether the talent channel opens (asymmetric $\rightarrow$ closed, $r \approx 0$; symmetric $\rightarrow$ open, r rising from +0.18 without agency to +0.32 with it) and the type of strategy RL emerges (asymmetric $\rightarrow$ avoidance and lying-flat; symmetric $\rightarrow$ active opportunity-seeking by the whole population with a reopened talent channel), and that social mobility $P(C > C_0)$ is highly sensitive to the incentive structure (Exp.1 only 8.3%, jumping to 59.7% in the incentive-repaired Exp.2). These strategy differences and their social meaning, above all the “lying-flat” under asymmetric luck, are developed in detail in the discussion (§6) and not repeated here.

5.6 Robustness Checks

Four checks confirm that the conclusions do not depend on a single random draw or on one arbitrary setting; all are shown below.

Baseline across initial conditions. Repeating the Baseline over 100 independent seeds (each drawing a fresh talent vector, agent positions and event trajectories) leaves the core TvL indicators stable (Table 7): the fraction of high-talent agents ending above C_0 holds near 34%, the 95% interval of the talent–wealth correlation straddles zero, the most successful agent’s talent is only slightly above the mean rather than the highest, and about half the agents end below C_0 . The qualitative TvL picture is reproduced under every one of the 100 initial conditions.

Post-convergence statistical protocol. After convergence is certified by the 15-condition talent-stratified criterion, all scalar metrics are averaged over 100 further post-convergence episodes. Across these episodes the structural metrics are highly stable ($\%<C_0$ standard deviation only 1.2–1.8 points, and the top-20% share barely moves), and the talent–wealth correlation, though noisier (s.d. 0.030–0.035), keeps a consistent sign ($r > 0$ in 100% of Exp.2 episodes and 80% of Exp.1’s; Table 8). The 100-episode mean $\pm$ s.d. basis used throughout is therefore reliable.

Table 7. Baseline 100-repetition check: stability of the core TvL indicators across initial conditions

Indicator	Original TvL	Mean	Std.	Range
P_T (fraction with $C > C_0$ among high-talent)	32%	34.2%	3.8%	[23.9%, 45.9%]
T^* (talent of the most successful)	0.61	0.687	0.089	[0.437, 0.951]
$\%<C_0$ (fraction below initial capital)	$\approx 50\%$	45.7%	1.7%	[41.6%, 49.2%]
$r(T, \log C)$ (talent–wealth correlation)	≈ 0	+0.063	0.030	[−0.012, +0.153]

Note: $P_T = P(C_{\text{end}} > C_0 \mid T > 0.7)$, consistent with the original TvL definition. T^* is the talent of the highest-final-wealth agent in each run.

Table 8. Post-convergence 100-episode statistical stability of the RL experiments (mean $\pm$ standard deviation)

Experiment	$\%<C_0$	$r(T, \log C)$	$r > 0$ occurrence
Exp.1 (asymmetric)	12.9% $\pm$ 1.2%	+0.029 $\pm$ 0.035	80%
Exp.2 (symmetric)	35.0% $\pm$ 1.8%	+0.317 $\pm$ 0.030	100%

Note: each metric is the mean $\pm$ standard deviation *computed per episode* over the 100 post-convergence episodes; “ $r > 0$ occurrence” is the fraction of the 100 episodes in which the talent–wealth correlation is positive, reflecting directional consistency.

Multi-seed replication of the RL experiments. To confirm the two emergent strategies do not depend on the single training seed, we re-ran Exp.1 and Exp.2 from three further independent seeds each (Table 9), re-seeding the environment, the network initialization, and the replay sampling together. The qualitative conclusions are unchanged across all four seeds. Under asymmetric luck the agents always collapse to lying-flat: the encounter rate stays near 0.5% (against

the 6% passive-agent baseline), 75–82% of agents end at C_0 , and the talent–wealth correlation stays near zero ($r = +0.03$). Under symmetric luck they always switch to active seeking: the encounter rate rises to 24–25%, a majority end above C_0 , and r stays near +0.3. Mobility is high in every run. The strict 15-condition convergence test was met within 1000 episodes for the original seed and for one of the three new Exp.1 seeds; in the other two the same lying-flat behavior was reached and stable, but the conservative test did not formally certify it within the episode cap, so their statistics are read from the post-cap episodes. Either way, the direction and size of both emergent strategies are robust to the random seed.

Table 9. Multi-seed robustness of the RL experiments: the two emergent strategies hold across four independent seeds

Experiment	seed	Enc.%	% at C_0	% above C_0	$r(T, \log C)$	Mobility%
Exp.1 (lying-flat)	2026	0.5	78.8	8.3	+0.029	95.1
	101	0.43	82.4	6.9	+0.027	96.1
	202	0.61	75.4	9.9	+0.036	97.0
	303	0.42	82.3	6.8	+0.027	79.0
Exp.2 (seeking)	2026	24.1	5.3	59.7	+0.317	95.3
	101	23.6	4.8	57.6	+0.319	93.6
	202	24.1	5.4	61.3	+0.311	92.5
	303	22.9	5.8	59.5	+0.287	92.9

Note: seed 2026 is the run used throughout the paper; seeds 101/202/303 are new. “Enc.%” is the share of steps an agent is on or next to an event, the same definition used in Table 6; all metrics are per-episode means over the 100 post-convergence (or post-cap) episodes. Exp.1 seeds 101 and 202 did not trigger the strict 15-condition convergence test within the 1000-episode cap but reached the same stable lying-flat behavior. Mobility varies more across seeds than the other metrics (Exp.1: 79.0–97.0%), yet seed 303 reaches the same 0.42% encounter rate and the same wealth outcome with the lowest mobility of the four, so it is the encounter rate, not how much the agents move, that carries the result.

What sets the 6% passive baseline. Every behavioral comparison in this paper is read against the 6% rate at which a non-steering agent meets events, so it matters that this number is not a free parameter. With N_E events spread over an $L \times L$ grid and a collision neighborhood of k cells, an agent that does not steer meets an event on a given step with probability $1 - (1 - k/L^2)^{N_E}$. For our setting ($L = 201$, $N_E = 500$, $k = 5$) this gives 6.0%, which is what we measure. The rate moves with those choices as expected: holding $N_E = 500$, a single-cell collision rule gives 1.2%, a 3×3 Moore neighborhood 10.5%, and a radius-2 von Neumann neighborhood 14.9%; holding $k = 5$, halving the events to $N_E = 250$ gives 3.1% and doubling them to $N_E = 1000$ gives 11.6%. The learned strategies should therefore be read against this baseline rather than as absolute rates. What carries the argument is that Exp.1 drives encounters far below the passive rate while Exp.2 drives them far above it, and that ordering does not depend on which of these settings is chosen.

6 Discussion

6.1 Two Laws of Individual Destiny

At the level of a single life, the results give two laws. First, the extremes of wealth are set by luck, not talent, yet talent is not irrelevant: it is the *realization rate* of luck, so once luck exposure is held fixed, talent predicts wealth cleanly (§5.5). Luck decides how many opportunities arrive, and talent decides how much of each is realized, so TvL’s “talent-irrelevance” is only how this looks before luck is accounted for. Second, whether talent matters at all is decided by the structure of luck, not by effort: under asymmetric luck the channel stays shut whether agents are passive or learning ($r \approx 0.03$ – 0.06); under symmetric luck it opens in two stages, first structurally, when a talented agent can dodge misfortune (Baseline2, $r = +0.18$), then behaviorally, once agency is switched on (Exp.2, $r = +0.32$). The structural stage is the larger surprise, a talented person gains ground not by striving but by living in a world where talent is allowed to matter, so reshaping that structure does more for the talented than their own effort does.

6.2 The New Finding: Lying-Flat as a Rational Choice

Beyond backing up TvL, the model gives one finding of its own: it shows how lying-flat emerges by itself from the learning. Many societies now have a name for this withdrawal of the young from the labor market, tangping in China (Matthyssen, 2024), satori and hikikomori in Japan (Furlong, 2008; Uchida & Norasakkunkit, 2015), the “N-po generation” in Korea, “quiet quitting” in the West; the model reads it as a rational response to an unfair payoff, not as a moral failing. Labor economics has long had a narrower version of the same idea in the discouraged worker, someone who stops looking for work not because they stopped wanting it but because looking has stopped paying.

It is fair to ask what the learning adds, since the arithmetic alone says that a bet with a negative expected return should be declined. The answer is the shape of the retreat, which the arithmetic does not give. The agents do not simply stop. They move almost every step, and still hold their encounter rate to 0.5%, a twelfth of the passive rate; 74.2% of them live a life in which nothing happens at all; and talent stops leaving any mark on wealth, with r falling to $+0.03$ and the richest no more talented than anyone else. A closed-form expected value gives the sign of the answer. It does not tell us that a whole population ends up busy, safe, equal and stalled.

Three layers of not-knowing. A real person deciding whether to strive faces three kinds of ignorance, each built into our agents. They do not know their own talent (the agent never observes T_i); they cannot tell an opportunity from a trap in advance (the observation window does not separate lucky from unlucky events); and they cannot control when good luck arrives (events wander at random). With so little to go on, the only thing that can rationally guide action is the return participation actually brings, discovered through trial and error, not known in advance.

When repeated trials keep showing that taking part shrinks wealth more often than it grows it, holding back is simply the rational response.

The asymmetric world is an unfair payoff. In Exp.1 the two kinds of event carry deeply unequal odds: an unlucky event strikes with probability 1, a lucky one only with the talent probability $T_i < 1$. The expected log-return of meeting an unknown event is therefore

$$\mathbb{E}\left[\ln \frac{C'}{C} \mid \text{encounter}\right] = \frac{1}{2}T_i \ln 2 + \frac{1}{2}(-\ln 2) = \frac{\ln 2}{2}(T_i - 1) < 0 \quad \forall T_i < 1,$$

negative for everyone. Many readers will find real examples of this, the familiar reality where effort and reward do not match: overtime that never lifts pay, a venture far likelier to burn savings than to grow them, credit for one's work taken by a boss or an advisor, and entry blocked by barriers, corruption, and missing information. Whatever the local cause, in the model these all reduce to one thing, a negative expected return to effort; seen this way, the obstacle is a structural one, not a personal failing.

Lying flat as a rational strategy, not a failure of effort. In the model, lying flat is not laziness; it is the best strategy the agent reaches after trying, failing, and learning from the feedback. Facing this payoff, RL converges with no prior on the one rational strategy: hold on to the starting capital and avoid every event. 78.8% of agents end with exactly their initial C_0 , and the upward channel $P(C_{\text{end}} > C_0)$ is only 8.3%, the most blocked of all four conditions. Read onto real life, this is a person keeping the capital their family gave them rather than spending it on a game they are bound to lose: showing up without exerting themselves, doing the minimum, staying home and waiting. It looks like idleness, but in the model it is an active, chosen defense, and an optimal one, in line with de Courson et al. (2025) on how the level of one's resources sets the best risk strategy.

Restlessness, not lying flat, is what does the damage. We need no separate experiment to see that hustling ends worse; the learning result already contains the answer. Reinforcement learning is trial and error in search of wealth, and in the rigged world it converged on avoidance. That convergence is itself the proof: had chasing events paid much better, the exploration phase would most likely have found it and the policy would have settled there instead, although as our second limitation notes we cannot rule out a local optimum. The arithmetic says the same thing. Because every encounter carries a negative expected log-return, $\frac{\ln 2}{2}(T_i - 1) < 0$ for all $T_i < 1$, expected log-wealth falls as the number of encounters grows, so the more an agent steers into events, the slower its wealth grows over a life. It is worth being exact about which average this is. In money terms a single encounter is a small gain, since $\mathbb{E}[C_i(t+1)/C_i(t)] = 0.75 + 0.5T_i = 1.05$ at the mean talent, but an unlucky event fires every time while a lucky one pays only with probability T_i , so a halving (50% of encounters) is more likely than a doubling ($0.5T_i$, about

30%). An agent that maximized expected wealth would therefore take these chances; one that maximizes the growth rate of its own capital, as ours does, declines them. Which criterion is right for a single life that cannot be replayed is a real question, and our reward function answers it by choosing growth. The passive Baseline puts a number on this: its agents meet events at the 6% chance rate rather than the 0.5% of learned avoidance, and the share ending below their starting capital rises accordingly, from 12.9% under Exp.1 to 45.7% in the Baseline. Steering deliberately toward events would only push exposure higher, and losses with it. The reading the model suggests runs opposite to the usual scolding: under a rigged payoff, sitting tight protects the principal, while rushing in is what spends it down.

The exit is a fair payoff, not more effort. In the model, none of this is a fixed trait. The moment the payoff is made fair, in Exp.2, where gains and losses both track talent and effort, the identical learning rule flips from avoidance to engagement: agents stop dodging events and start seeking them, and the share of high-talent agents ($T > 0.7$) ending above their starting capital, P_T , jumps from 9.9% under Exp.1 to 76.8% under Exp.2 (Baseline 34.2%, Baseline2 47.6%). The same agents who lay flat now strive. Within the model, then, lying flat is a response to the payoff rather than a trait, and what brings engagement back is a fairer payoff, not urging the agent to try harder. Read onto the world, this is consistent with the view that disengagement is better met by reshaping the returns to effort than by moral appeals to try harder.

The limit of the claim. The model offers no policy fix, since these imbalances are built into the system and no single reform removes them; what it offers is a mechanism, a way to understand why, where the returns to effort are held down, disengagement can be the rational response rather than a failure of character.

Lying flat is itself a privilege. One caveat keeps the mapping honest: sitting out needs a principal to defend. Our agents all start from the same C_0 , but in reality the initial endowment is unequal and is itself a draw in the birth lottery (§3), what Korea’s “spoon class” idiom calls being born with a gold spoon rather than a dirt one. Those born with little cannot afford to sit out the negative-expectation game and must play it regardless. We hold the endowment equal here to isolate the mechanism, and leave its unequal distribution to future work.

6.3 Limitations and Future Work

While this framework extends the TvL model by introducing deep reinforcement learning, several simplified assumptions highlight directions for future research:

Simplified Spatial Dynamics and Frictionless Movement. The environment assumes a homogeneous grid with frictionless agent movement and non-rival, exogenous events. Incorporating spatial heterogeneity, relocation/switching costs, and endogenous event congestion would

transform the environment into a genuine multi-agent spatial game requiring multi-agent reinforcement learning (MARL).

Bounded Agency and Policy Homogeneity. Agents operate with local perception (11×11 window), memoryless policies, and shared network parameters, learning over multi-lifetime episodes rather than during a single life. Extending this to heterogeneous policies, social learning, memory, or talent-aware agents could provide deeper insights into individual behavioral variation.

Algorithmic and Convergence Limits. Strategies are learned via Double DQN, which does not guarantee global optimality. Emergent behaviors reflect learned policies under specific exploration constraints rather than theoretical global optima. Incorporating advanced exploration mechanisms or analytical benchmarks remains a priority for future quantification.

Stylized Talent and Luck Mechanisms. Talent is treated as a fixed, unobserved parameter affecting event conversion rather than direct event exposure, while luck events rely on fixed multiplicative jumps ($\times 2, \times 0.5$). Endogenizing talent evolution through experience, modeling talent-driven exposure, and applying empirical, heavy-tailed luck distributions represent natural next steps.

Unidimensional Success Metric. Success is measured solely by terminal capital, ignoring the time-value of wealth and non-monetary dimensions of life (e.g., well-being, health, relationships). Evaluating agents under a multidimensional utility function may reframe behaviors like “lying flat” as rational revaluations rather than passive withdrawals.

7 Conclusion

This study set out to learn whether active behavior can rewrite fate. Within the bounded agency we model, it cannot. We gave TvL’s agents a Double DQN brain and let them seek gains and avoid losses at every step, and across a two-by-two design crossing agency with the structure of luck, luck still decided where a life ended up.

Two conclusions follow. The first is the one the passive model cannot reach, and it is where this paper adds something new: lying-flat as the rational choice. Under the original asymmetric luck, where a chance is more likely to cost than to pay, the agents converge on their own to avoid all opportunity, not out of laziness but because holding back is what pays. The same reward makes them seek events once luck is made symmetric, showing that lying flat is a rational response to the reward-risk balance rather than laziness, and what revives engagement is a change in the payoff rather than urging people to try harder. In the model that change is a switch we flip. Outside it, the returns to effort are set by institutions, bargaining power and

unequal starting points, so the model points to where the problem sits rather than prescribing a fix. This pulling back costs the whole group, not just the person: when everyone holds back, the group produces less and talent has nowhere to go, so the most able people are wasted along with everyone else. What the learning adds beyond the arithmetic is the shape of this retreat: a population that stays busy, keeps itself safe, ends up equal, and goes nowhere.

The second conclusion is that the core TvL findings survive this test, and the learning framework sharpens them. In every condition the richest agents are the luckiest, and are usually not the most talented, and wealth stays heavy-tailed in all four except Exp.1, where lying-flat flattens it onto C_0 ; either way, agency does not overturn luck. And TvL's near-zero talent–wealth correlation is only how things look before luck is accounted for: once luck exposure is held fixed, talent predicts wealth cleanly, because talent is simply the rate at which luck is turned into gains. Luck decides how many opportunities arrive; talent decides how many are seized.

Active behavior, then, has limits, but it is not powerless. Under punishing luck it keeps most agents from ending worse than they began, with the share below their starting capital falling from 45.7% to 12.9%; under fairer luck it lets talent leave a real mark on wealth, with the talent–wealth correlation rising from near zero to +0.32, and the more talented are also less likely to fall below the money they started with, though only where luck is symmetric: among agents with $T > 0.7$ the share ending below C_0 is 18.7% in Baseline2 and 18.6% in Exp.2, against 36.3% and 52.4% for those with $T < 0.5$. Under the original asymmetric luck that protection is thin (Baseline 43.1% against 48.6%) or gone (Exp.1 12.8% against 13.0%, where holding back shields everyone alike). We cannot change our fate, yet within its limits we can still make ourselves a little better off.

References

- Calvano, E., et al. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297.
- DiPrete, T. A., & Eirich, G. M. (2006). Cumulative advantage as a mechanism for inequality. *Annual Review of Sociology*, 32, 271–297.
- Frank, R. H. (2016). *Success and Luck*. Princeton University Press.
- van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double Q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).
- Merton, R. K. (1968). The Matthew effect in science. *Science*, 159(3810), 56–63.
- Mnih, V., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529–533.
- Peysakhovich, A., & Lerer, A. (2018). Prosocial learning agents solve generalized Stag Hunts better than selfish ones. In *Proceedings of AAMAS 2018*.
- Pluchino, A., Biondo, A. E., & Rapisarda, A. (2018). Talent-vs-Luck: The role of randomness in success and failure. *Advances in Complex Systems*, 21(03n04), 1850014.
- Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, 69(1), 99–118.
- Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3–4), 279–292.
- Bénabou, R., & Tirole, J. (2002). Self-confidence and personal motivation. *Quarterly Journal of Economics*, 117(3), 871–915.
- Challet, D., Pluchino, A., Biondo, A. E., & Rapisarda, A. (2020). The origins of extreme wealth inequality in the talent-versus-luck model. *Advances in Complex Systems*, 23(02), 2050004.
- Gabaix, X., & Landier, A. (2008). Why has CEO pay increased so much? *Quarterly Journal of Economics*, 123(1), 49–100.
- McCall, J. J. (1970). Economics of information and job search. *Quarterly Journal of Economics*, 84(1), 113–126.
- Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50(3), 665–690.
- Stigler, G. J. (1961). The economics of information. *Journal of Political Economy*, 69(3), 213–225.
- Zheng, S., Trott, A., Srinivasa, S., Parkes, D. C., & Socher, R. (2022). The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science Advances*, 8(18), eabk2607.
- Rudd-Jones, J., Musolesi, M., & Pérez-Ortiz, M. (2025). Multi-agent reinforcement learning simulation for environmental policy synthesis. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '25)*, 2890–2895.

- Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.
- Biondo, A. E., Pluchino, A., & Zanola, R. (2024). Simulating the emergence of superstar firms: The role of luck vs talent. *Physica A: Statistical Mechanics and Its Applications*, 644, 129848.
- de Courson, B., Frankenhuis, W. E., & Nettle, D. (2025). Poverty is associated with both risk avoidance and risk taking. *Proceedings of the Royal Society B*, 292, 20242071.
- Hamada, H. (2023). Luck of outcome in the Talent-vs-Luck Model. *Advances in Complex Systems*, 26(04n05), 2350008.
- Hur, J., Ha, M., & Jeong, H. (2024). Interplay of network structure and talent configuration on wealth dynamics. *Physical Review E*, 110, 024312.
- Breiman, L. (1961). Optimal gambling systems for favorable games. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 65–78. University of California Press.
- Kelly, J. L. (1956). A new interpretation of information rate. *Bell System Technical Journal*, 35(4), 917–926.
- Sert, E., Bar-Yam, Y., & Morales, A. J. (2020). Segregation dynamics with reinforcement learning and agent-based modeling. *Scientific Reports*, 10, 11771.
- Agarwal, S., Driscoll, J. C., Gabaix, X., & Laibson, D. (2009). The age of reason: Financial decisions over the life cycle and implications for regulation. *Brookings Papers on Economic Activity*, 2009(2), 51–117.
- Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3), 797–817.
- Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5), 992–1026.
- Knight, F. H. (1921). *Risk, Uncertainty and Profit*. Boston: Houghton Mifflin.
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134.
- Lusardi, A., & Mitchell, O. S. (2014). The economic importance of financial literacy: Theory and evidence. *Journal of Economic Literature*, 52(1), 5–44.
- Roemer, J. E. (1998). *Equality of Opportunity*. Cambridge, MA: Harvard University Press.
- Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. *Strategic Management Journal*, 18(7), 509–533.
- Furlong, A. (2008). The Japanese hikikomori phenomenon: Acute social withdrawal among young people. *The Sociological Review*, 56(2), 309–325.
- Uchida, Y., & Norasakkunkit, V. (2015). The NEET and Hikikomori spectrum: Assessing the risks and consequences of becoming culturally marginalized. *Frontiers in Psychology*, 6, 1117.

Matthyssen, M. (2024). Tangping (lying flat) among young adults: Shameful, courageous or just fleeting resistance? In *Self-Development, Ethics and Politics in China Today: A Keyword Approach*. Amsterdam University Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). Cambridge, MA: MIT Press.

Appendix A: Training Algorithm

Algorithm 1 RL-TvL model training procedure

```

1: Initialize: grid  $G$  ( $201 \times 201$ , toroidal),  $N = 1000$  agents,  $N_E = 500$  event points
2: Initialize Q-network  $Q(\cdot; \theta)$  and target network  $\hat{Q}(\cdot; \hat{\theta})$ , experience replay buffer  $\mathcal{D}$ 
3: for episode = 1 to  $E_{\max}$  (at most 1000) do
4:   Randomly initialize agent positions, reset capital  $C_i \leftarrow 10$ , initialize event positions
5:   for  $t = 1$  to  $T_{\max}$  (80 six-month steps) do
6:     Each event random-walks  $v_E = 2$  cells (toroidal grid)
7:     for each agent  $i = 1$  to  $N$  (in parallel) do
8:       Build local observation  $s_i(t)$  ( $11 \times 11$  window + personal state)
9:       Select action  $a_i \in \mathcal{A}$  by  $\varepsilon$ -greedy policy
10:      Execute action: if  $a_i \neq \text{stay}$ , move  $v_A = 2$  cells in that direction
11:      Check whether struck by an event; update  $C_i$  by Eq. (1)
12:      Compute immediate reward  $R_i(t)$ , observe new state  $s_i(t+1)$ 
13:      Store transition  $(s_i, a_i, R_i, s'_i)$  in  $\mathcal{D}$ 
14:    end for
15:    if  $|\mathcal{D}| > B$  then
16:      Sample a mini-batch from  $\mathcal{D}$ , compute the Double DQN loss, update  $\theta$ 
17:    end if
18:  end for
19:  Decay  $\varepsilon$ ; sync target network  $\hat{\theta} \leftarrow \theta$  every 10 episodes
20:  Record this episode's wealth distribution and per-talent-group average mobility rate
21: end for

```
