Path Dependence under Adaptive AI Delegation

Lingxiao Huang
Nanjing University*

Nisheeth K. Vishnoi
Yale University

Abstract

Repeated AI assistance can improve immediate task performance while reducing the skill available for future independent work. We develop a mathematical framework for this long-run tradeoff. The model tracks two state variables: a latent human skill level governing expected independent performance, and a delegation level representing the learner's evolving tendency to rely on AI. Skill changes through error-driven learning under practice and decay under delegation; delegation responds to observed performance, increasing when AI-assisted work appears to outperform independent work. We analyze the resulting dynamics and contrast them with fixed delegation. With fixed delegation, skill follows a one-dimensional learning-decay process with a single stable equilibrium. With adaptive delegation, the coupled system has two attracting equilibria separated by the stable manifold of an interior saddle. The existence and geometry of this separatrix require a global phase-plane analysis of the coupled dynamics. The system is path-dependent: small differences in initial skill or reliance can lead to different long-run outcomes. We use this characterization to show that AI assistance can improve short-run performance while producing worse long-run performance than a no-AI baseline, even under fixed delegation. Increasing AI capability can enlarge the basin of attraction of the low-skill equilibrium, making delegation appear beneficial for longer while increasing the risk of eventual skill loss. The qualitative picture is observed to persist across alternative specifications. Together, these results show that the risk is not AI assistance itself, but the coupling between performance-driven reliance and use-dependent skill change.

1 Introduction

AI systems are increasingly used to support cognitive tasks such as writing, coding, tutoring, and problem solving [2, 17, 3, 18]. In many of these settings, AI assistance improves task performance, output quality, or throughput over a short horizon [3, 18, 17]. The short-run view treats AI use as a performance-enhancing intervention.

Repeated use changes the nature of the interaction. A learner who relies on AI across many related tasks is not only producing outputs; they are also changing the skill with which future outputs will be produced. Delegation today therefore changes both the immediate output and the human capacity that produces future outputs. The two effects can pull in opposite directions. Recent empirical work suggests that this distinction matters: studies of AI-assisted writing, programming, and mathematics report immediate gains in performance or productivity alongside reduced retention, weaker transfer, or worse post-test performance once assistance is removed [16, 3, 18, 19, 32, 8]. These effects vary across users [33, 18, 36]. Reliance also changes with feedback and beliefs formed through AI use [6].

These findings raise a structural question: *how does adaptive AI reliance shape the long-run dynamics of human skill, and when do differences in initial skill and reliance lead to different long-run outcomes?*

*State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing 210023, China

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Related work. The empirical studies above are often interpreted through cognitive offloading, automation bias, task substitution, or reduced cognitive engagement [24, 23, 30, 8]. These perspectives help explain why users may rely on AI in the moment, but they do not provide a dynamical account of how reliance and skill co-evolve over repeated use.

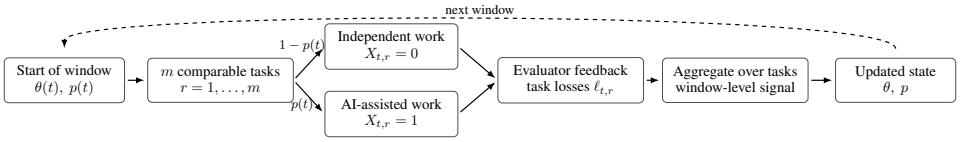
A complementary theoretical direction studies AI delegation as a one-shot decision problem. A recent model of delegation and verification under AI [15] considers an agent who optimizes how much to delegate and verify under task-level performance, effort, and verification tradeoffs. That framework explains workflow choice and institutional evaluation in a single decision environment. The present paper studies a different regime: delegation is an evolving reliance tendency formed across repeated tasks. The central object is therefore not within-task delegation, but the long-run feedback loop between performance-driven reliance and human skill formation.

Mathematical models of learning and forgetting capture error-driven improvement and use-dependent decay under fixed task execution [7, 22, 12, 11, 29, 26]. Classical stochastic approximation connects such rules to limiting ODEs whose global behavior governs convergence [5]. These models typically treat practice as exogenous or fixed. They do not capture settings in which reliance on AI responds to observed performance, thereby endogenously changing the practice that drives future skill. Related work studies how payoffs change strategy use [14, 31], how investment builds human capital [4], and how deployed models change data distributions [27, 21]. Thus, the novelty is not learning decay or adaptive reliance in isolation, but the basin structure created when reliance adapts to short-run performance while skill evolves through use and non-use.

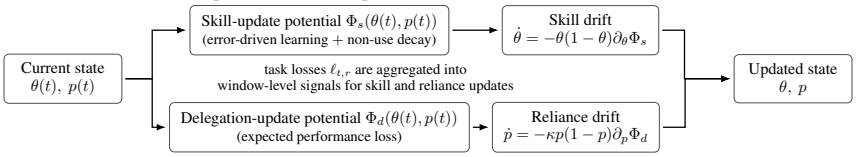
Our contributions. Answering the question above requires a model in which reliance and skill evolve together. We develop such a framework for repeated AI-assisted learning, with two coupled state variables: a latent skill level $\theta(t)$ governing expected independent performance, and a delegation level $p(t)$ representing the learner's tendency to rely on AI. Skill evolves through error-driven learning under practice and decay under delegation, while delegation responds to the short-run performance comparison between independent and AI-assisted work. This separation is central: reliance adapts to observed performance, while skill evolves through the consequences of use and non-use. Our main results are as follows.

- We characterize the equilibrium structure of the resulting dynamical system. With fixed delegation, the dynamics reduce to a one-dimensional learning-decay process with a single stable equilibrium. With adaptive delegation, the system has two attracting equilibria: a high-skill, low-delegation equilibrium and a low-skill, high-delegation equilibrium, together with an interior saddle (Theorem 3.1).
- We characterize which initial conditions lead to which long-run outcome. The stable manifold of the interior saddle is a global basin boundary in the (θ, p) state space, separating initial skill-delegation pairs that converge to sustained learning from those that converge to persistent delegation and low skill (Theorem 3.2). This yields path dependence: small differences in initial skill or reliance can lead to different long-run outcomes.
- We show that short-run performance gains can mask long-run losses. AI assistance can strictly improve performance initially while producing worse long-run performance than the no-AI baseline, even under fixed delegation (Theorem 3.5 and Section B.6). This effect arises because delegation reduces the independent practice that builds future skill, not from an explicit incentive misalignment.
- We analyze how the basin structure changes under AI capability and model variations. Increasing AI capability reshapes the basin boundary (Theorem 3.3), enlarging the set of initial conditions that converge to the low-skill equilibrium and extending the duration of apparent performance gains. The same qualitative structure is observed under jagged or misperceived AI performance, alternative objectives, multiple skills, learning from AI during delegation, and richer learner-AI dynamics (Sections 4 and D); the geometric view also suggests interventions that shift the basin boundary toward sustained-skill outcomes (Remark 3.4 and Section C).

Together, the framework and these results identify the possible long-run regimes, characterize which initial skill and reliance levels lead to each, and show how AI capability reshapes this partition. The source of long-run risk is not AI assistance itself, but the dynamic coupling between performance-driven reliance and use-dependent skill change.



(a) Repeated-task adaptation within a short time window.



(b) Coupled update structure.

Figure 1: Overview of the repeated-task adaptation model. In each short time window t, the learner faces many comparable tasks and delegates a fraction $p(t)$ of them. Evaluator feedback is aggregated across tasks and enters the skill-update and delegation-update mechanisms. The former combines error-driven learning under practice with non-use decay, while the latter updates reliance through relative performance feedback.

2 Model

We study a repeated task setting in which a learner faces many comparable tasks over time. In each short time window, the learner completes a fraction of tasks independently, delegates the remaining fraction to AI, and receives feedback on the resulting outputs. The window is chosen short enough that the learner's skill $\theta(t)$ and delegation level $p(t)$ can be treated as fixed within the window, but long enough to average over many task instances. The learner's state is therefore $(\theta(t), p(t))$. We focus on *error-driven* learners [29, 10], whose skill and reliance updates are driven by observed performance feedback.

Skill representation. We represent task-specific skill by a latent parameter $\theta \in [0, 1]$ that governs expected independent performance on a repeated class of comparable tasks. The variable θ is not the realized score on a particular task; realized performance may vary with task difficulty, attention, or execution noise. This abstraction is standard in latent-trait models of ability. In item response theory, ability is modeled as a continuous latent trait that combines with item difficulty to determine response probabilities [20, 9]. Our use of θ is analogous, normalized so that $\theta = 1$ denotes optimal expected independent performance. The scalar representation is an analytical simplification, not a claim that ability is intrinsically one-dimensional; Section D.3.1 analyzes a multiple-skill extension.

Let $\theta(t)$ denote the learner's skill during window t. Let $\theta_a \in [0, 1]$ denote the AI's *effective skill*, defined through its expected loss under the same task objective; under the squared-loss baseline, random AI performance S_a is represented by $(1 - \theta_a)^2 = \mathbb{E}[(1 - S_a)^2]$.

Delegation representation. Let $p(t) \in [0, 1]$ denote the delegation level, i.e., the fraction of tasks delegated to AI in window t. For a task sampled from this window, let $X \sim \mathrm{Bern}(p(t))$, where $X = 1$ indicates delegation to AI and $X = 0$ indicates independent work. Thus $p(t)$ is an aggregate reliance tendency shaped by feedback. The Bernoulli notation represents within-window task fractions, not a one-shot delegation decision. The same dynamics can also be used to describe continuous AI contribution p; see Section D.2.2.

Performance loss. Task performance is evaluated against a fixed target by an external evaluator, which may represent a teacher, benchmark, or grader. Let $\ell(\theta)$ denote the expected performance loss of an output produced with latent skill $\theta \in [0, 1]$. We assume that $\ell(\theta)$ is monotonically decreasing, with $\ell(1) = 0$. A canonical example is the power loss $\ell(\theta) = (1 - \theta)^z$ for $z \geq 1$, with the squared loss $z = 2$ used as the baseline. The instantaneous performance loss for a task with delegation indicator X is $\ell(\theta(t), X) := (1 - X)\ell(\theta(t)) + X\ell(\theta_a)$. Averaging over $X \sim \mathrm{Bern}(p(t))$ gives

$$\ell(\theta(t), p(t)) = (1 - p(t))\ell(\theta(t)) + p(t)\ell(\theta_a). \tag{1}$$

When $\ell(\theta(t)) \geq \ell(\theta_a)$, i.e., $\theta_a \geq \theta(t)$, increasing delegation improves short-run performance.[2]

[2]The loss ℓ may implicitly reflect accuracy, time, or effort: any factor affecting observed task quality enters through the feedback signal. An additional delegation benefit, as in Section C.6, represents only benefits not already included in ℓ.

Skill update. After receiving feedback, the learner's skill changes depending on how the task was completed. If the learner works independently ($X = 0$), feedback generates an error-driven learning signal. If the task is delegated to AI ($X = 1$), the learner does not directly practice the underlying task. We model non-use as drift toward a default skill level $\theta_d \in [0, 1]$, consistent with learning and forgetting models in which performance improves through practice and weakens without continued practice or reinforcement [7, 22, 12, 25]. We encode these effects by the skill-update potential

$$\Phi_s(\theta(t); X) = (1 - X)\ell(\theta(t)) + \Delta X g(\theta(t), \theta_d),$$

where $\Delta > 0$ controls the strength of decay and $g(\theta, \theta_d) \geq 0$ measures distance from the default skill. This potential is distinct from the performance loss: its first term captures learning from practice, while its second term captures decay from non-use. The function g is left general at this stage; the squared distance used in the main analysis is chosen for tractability, and Section D.1.4 examines power-distance alternatives, showing that the qualitative phase structure is not specific to this choice.

Averaging over $X \sim \mathrm{Bern}(p(t))$ within window t gives

$$\Phi_s(\theta(t); p(t)) = (1 - p(t))\ell(\theta(t)) + \Delta p(t) g(\theta(t), \theta_d). \tag{2}$$

To keep the latent skill in $[0, 1]$, we use the one-dimensional information geometry $F(\theta) = 1/(\theta(1 - \theta))$, the Fisher metric of a Bernoulli parameter. The induced natural-gradient flow preconditions ordinary gradient descent by $F(\theta)^{-1} = \theta(1 - \theta)$, yielding multiplicative or replicator-style dynamics [1, 14]. This is a geometric modeling choice for bounded latent skill, not an empirical claim about the exact form of human learning. Applying this flow to Φ_s gives

$$\dot{\theta} = -F(\theta)^{-1}\partial_\theta \Phi_s = -\theta(t)(1 - \theta(t)) \left[(1 - p(t))\partial_\theta \ell(\theta(t)) + \Delta p(t)\partial_\theta g(\theta(t), \theta_d)\right]. \tag{3}$$

Since $\ell(\theta)$ decreases with skill, the first term improves skill under independent work; the second term pushes skill toward θ_d under delegation.

Delegation update. The learner adjusts reliance $p(t)$ based on observed performance across tasks. If AI-assisted outputs perform better than independent work, reliance increases; otherwise, it decreases. Assuming accurate perception of expected losses, we use the following delegation-update potential for performance-based reliance [6]:

$$\Phi_d(\theta(t), p(t)) := \ell(\theta(t), p(t)). \tag{4}$$

Unlike the skill-update potential, which separates learning from practice and decay from non-use, Φ_d measures which delegation level gives lower short-run loss. Since $p(t)$ is a bounded reliance variable, we use the multiplicative geometry $F(p) = 1/(p(1 - p))$. The induced natural-gradient flow is

$$\dot{p} = -\kappa F(p)^{-1}\partial_p \Phi_d(\theta(t), p(t)) = \kappa p(t)(1 - p(t))\big(\ell(\theta(t)) - \ell(\theta_a)\big), \tag{5}$$

where $\kappa > 0$ controls the relative speed of delegation adaptation. Thus, reliance increases when AI has lower expected loss than independent work, and decreases when independent work has lower expected loss.

Dynamics. We now specialize the general skill and delegation updates to the main specification analyzed in the paper:

$$\ell(\theta) = (1 - \theta)^2, \qquad g(\theta, \theta_d) = (\theta - \theta_d)^2.$$

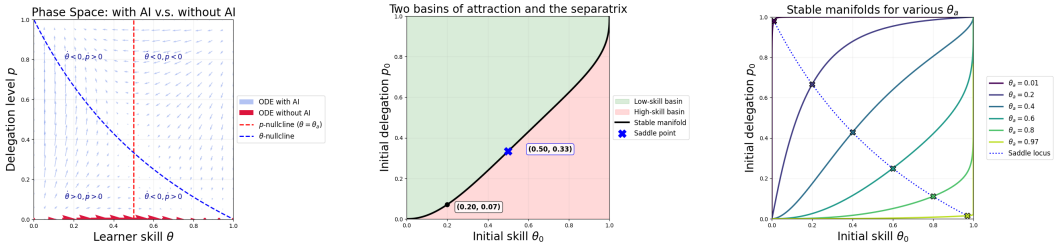
This yields the coupled dynamics

$$\dot{\theta} = 2\theta(1 - \theta)\big((1 - p)(1 - \theta) + \Delta p(\theta_d - \theta)\big), \;\; \dot{p} = \kappa p(1 - p)\big((1 - \theta)^2 - (1 - \theta_a)^2\big). \tag{6}$$

Throughout the analysis, we compare the long-run behavior of (6) across different initial conditions (θ_0, p_0) and against the no-AI case. The initial delegation level p_0 reflects the learner's prior belief about the relative effectiveness of AI-assisted and independent work, formed through prior experience or external information.

Unless otherwise stated, we consider the regime $\theta_a, \theta_0 \in [\theta_d, 1]$. For the main analysis, we set $\theta_d = 0$, such that non-use decay is toward the lowest skill level on the chosen scale. We also rescale time to remove the factor 2 in the skill equation and absorb the resulting constant into the free relative-speed parameter κ. This gives

$$\dot{\theta} = \theta(1 - \theta)\left((1 - p)(1 - \theta) - \Delta p\theta\right), \;\; \dot{p} = \kappa p(1 - p)\left((1 - \theta)^2 - (1 - \theta_a)^2\right). \tag{7}$$



(a) Fixed versus adaptive delegation

(b) Stable manifold and basin partition

(c) AI skill reshapes the basin boundary

Figure 2: Phase portraits for the adaptive-delegation dynamics. (a) Fixed delegation has a single attracting skill level, while adaptive delegation creates competing attracting regimes. (b) The stable manifold of the interior saddle separates the high-skill and low-skill basins. (c) Increasing AI skill shifts the basin boundary downward, expanding the low-skill basin. Default parameters are $(\theta_a, \kappa, \Delta) = (0.5, 3, 2)$ unless varied.

The general ODE (6) is derived and analyzed in Section A; biased perceived AI skill and stochastic feedback are discussed in Sections D.1.2 and D.1.3, respectively. The model is intended to characterize qualitative dynamics rather than to serve as a calibrated structural model of human behavior. Table 4 lists these assumptions, their roles, and the extensions studied here.

The functional forms above make the phase-plane analysis tractable; the same phase structure is observed to persist under the alternative specifications studied in Section 4, including alternative losses and decay functions, jagged AI performance, biased delegation updates, multiple skills, and learning from AI during delegation. In simulations, noisy feedback shows a sharp finite-horizon transition near the deterministic separatrix.

Fixed delegation and no-AI case. When the delegation level is held fixed at $p(t) \equiv p_0$, the system reduces to the one-dimensional skill dynamics

$$\dot{\theta} = \theta(1-\theta)\big((1-p_0)(1-\theta) - \Delta p_0 \theta\big). \tag{8}$$

The no-AI case corresponds to $p_0 = 0$. For any fixed $p_0 \in [0,1]$, (8) has the unique stable equilibrium $\theta^\infty(p_0) = \frac{1-p_0}{1-p_0+\Delta p_0}$. Thus every trajectory with $\theta_0 \in (0,1)$ converges to $\theta^\infty(p_0)$, and $\theta^\infty(0) = 1$. With fixed full delegation, $p_0 = 1$, skill decays to $\theta^\infty(1) = 0$.

This fixed-delegation case separates ordinary non-use decay from the adaptive effect studied below. With fixed delegation, the model is one-dimensional and has a single stable equilibrium. Path dependence and multiple basins arise only when delegation adapts jointly with skill. These feedback directions do not, however, guarantee multiple basins. Section B.1 gives a system with the same feedback directions as (7), in which every interior trajectory converges to the same equilibrium. Figure 2a contrasts the no-AI and adaptive-delegation phase portraits.

3 Theoretical results

We analyze the long-run behavior of the adaptive-delegation dynamics (7) on the interior $(0,1)^2$; basic well-posedness facts are proved in Section A.3. We begin with the fixed-point structure. Fixed delegation already captures the direct effect of reduced practice on long-run skill. The results below show what changes when practice is endogenous: the two-way coupling between skill and reliance creates an interior saddle whose stable manifold separates two long-run regimes.

Theorem 3.1 (Equilibria of ODE (7)). *Let θ_a, κ, Δ be the parameters of ODE (7). When $\theta_a \in (0,1)$, the system has two corner equilibria, $(1,0)$ and $(0,1)$, that are locally asymptotically stable relative to $[0,1]^2$; two repelling corner equilibria, $(0,0)$ and $(1,1)$; and one hyperbolic interior saddle $(\theta^\dagger, p^\dagger) = \left(\theta_a, \frac{1-\theta_a}{1-(1-\Delta)\theta_a}\right)$. The stable corner equilibria are non-hyperbolic because of the boundary factors in the vector field.*

The key qualitative change is the coexistence of high-skill and low-skill attracting regimes. In contrast, neither the no-AI baseline nor fixed delegation with $p_0 < 1$ admits a low-skill equilibrium $(0,1)$ (Section 2). The remaining question is which initial conditions converge to which attractor. Theorem 3.2 identifies the basin boundary as the stable manifold of the interior saddle.

Theorem 3.2 (Basin partition by the stable manifold). *Let θ_a, κ, Δ be the parameters of ODE* (7)*. Let $(\theta^\dagger, p^\dagger)$ be the interior saddle point guaranteed by Theorem 3.1. There exists a nondecreasing differentiable function $\psi : (0,1) \to (0,1)$ that extends continuously to $[0,1]$ with $\psi(0) = 0$ and $\psi(1) = 1$, whose graph coincides with the one-dimensional stable manifold of $(\theta^\dagger, p^\dagger)$ within $(0,1)^2$. The curve $p = \psi(\theta)$ partitions the interior state space $(0,1)^2$ into two distinct basins of attraction: Initial states with $p_0 > \psi(\theta_0)$ converge to $(0,1)$, while those with $p_0 < \psi(\theta_0)$ converge to $(1,0)$.*

Theorem 3.2 gives the path-dependence statement: for the same parameters, small changes in the initial state can place trajectories on opposite sides of the separatrix and lead to different long-run outcomes. Figure 2b visualizes the stable manifold and the two basins.

The monotonicity of ψ also shows why early-stage learners are more fragile: when θ_0 is small, even modest delegation can place the trajectory in the low-skill basin. Since $\psi(\theta) < 1$ for every $\theta < 1$, sufficiently heavy reliance can place even high-skill learners in the low-skill basin. A piecewise-polynomial approximation $\widetilde{\psi}(\cdot)$ of the one-skill basin boundary appears in Section B.8, Eq. (27).

In the two-skill extension, simulations suggest that a separating surface replaces this curve; Section 4 summarizes the supporting local analysis and numerical evidence.

Effect of AI quality on the basin boundary. We next characterize how the basin boundary changes with AI skill.

Theorem 3.3 (AI skill monotonically sweeps the basin boundary). *For every $\theta \in (0,1)$, the boundary value $\psi_{\theta_a}(\theta)$ is nonincreasing and continuously differentiable in θ_a. Moreover, the family $\{\psi_{\theta_a}\}_{\theta_a \in (0,1)}$ sweeps the state space in the following sense: for every $(\theta_0, p_0) \in (0,1)^2$, there exists $\theta_a \in (0,1)$ such that $p_0 = \psi_{\theta_a}(\theta_0)$.*

Thus, increasing AI skill shifts the basin boundary downward and enlarges the set of initial conditions that converge to the low-skill equilibrium. More capable AI can therefore make delegation attractive for more learners while raising the threshold for sustained independent skill. Figure 2c illustrates this deformation; variation with respect to κ and Δ is studied in Section B.7. The same comparative statics imply that timing matters: early independent practice can move a learner toward the high-skill side of the boundary, while early delegation can move the learner toward the low-skill side.

Remark 3.4 (Interventions as basin-boundary shifts). The basin characterization gives a geometric view of interventions: a temporary reduction in reliance or increase in skill moves the learner relative to the separatrix $p = \psi(\theta)$, whereas sustained changes to delegation incentives can shift the boundary itself. Representative mechanisms are studied in Section C.

3.1 Analyzing performance loss across time

Adaptive delegation may reduce instantaneous loss in the short run while lowering long-run skill, a pattern observed in recent experiments [16, 3]. We formalize this reversal by comparing the same learner with and without AI access, holding initial skill fixed.

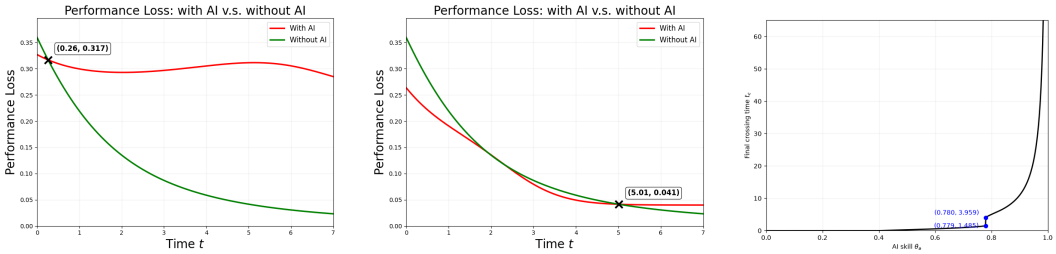
For an initial state $(\theta_0, p_0) \in (0,1)^2$, let $\theta(t; \theta_0, p_0)$ and $p(t; \theta_0, p_0)$ denote the learner's skill and delegation level at time t, respectively, under adaptive delegation. In the no-AI baseline, $\theta(t; \theta_0, 0)$ denotes the learner's skill trajectory starting from the same initial skill, yielding performance loss $(1 - \theta(t; \theta_0, 0))^2$. We define the *performance gap* at time t as

$$G_\ell(t) := \ell\left(\theta(t; \theta_0, p_0), p(t; \theta_0, p_0)\right) - (1 - \theta(t; \theta_0, 0))^2,$$

so $G_\ell(t) < 0$ means that delegation yields lower instantaneous loss than the no-AI baseline. If $\theta_0 \geq \theta_a$, delegation provides no initial performance advantage, and Lemma B.2 implies that the no-AI trajectory has weakly higher skill and lower loss for all $t > 0$. We therefore focus on $\theta_0 < \theta_a$.

Theorem 3.5 (Short-term gains, long-term losses). *Let θ_a, κ, Δ be the parameters of ODE* (7)*. Fix an initial state $(\theta_0, p_0) \in (0,1)^2$ with $\theta_0 < \theta_a < 1$. Let $t_c := \inf\{t \geq 0 : \forall t' > t,\ G_\ell(t') > 0\}$ denote the final crossing time, after which adaptive delegation incurs higher loss than the no-AI baseline. Then $t_c < \infty$ and $\theta(t_c; \theta_0, 0) \leq \theta_a$.*

After the final crossing time t_c, delegation no longer offers an advantage in instantaneous task loss. The same reversal also occurs under fixed positive delegation (Appendix B.6). The post-crossing behavior depends on the basin of the adaptive-delegation trajectory. In the low-skill basin, the



(a) Short-run gains and long-run losses

(b) Multiple crossings at high AI skill

(c) Crossing time versus AI skill

Figure 3: Plots illustrating the instantaneous performance loss over time with and without AI, and how the crossing point varies with respect to AI skill θ_a, with default settings $(\theta_a, \kappa, \Delta, \theta_0, p_0) = (0.5, 3, 2, 0.4, 0.3)$.

trajectory converges to persistent reliance and low skill, so the performance gap remains positive in the limit. In the high-skill basin, both trajectories converge to high skill, but the no-AI trajectory does so more directly; $G_\ell(t)$ remains positive after t_c and vanishes asymptotically. This formalizes a mechanism consistent with cognitive debt [16]: short-term gains can be outweighed by structurally foregone practice.

Effect of AI quality on the crossing time. Figure 3c shows that t_c can grow sharply with θ_a: the no-AI learner may need longer to surpass a highly capable AI. Thus higher AI capability can extend the window of apparent gains while also expanding the low-skill basin (Theorem 3.3).

Remark 3.6 (Empirical predictions). Although the analysis is theoretical, the state variables and parameters have direct counterparts in repeated-task data. Given task-level delegation indicators $X_{t,r}$, realized losses $\ell_{t,r}$, and AI-only benchmark losses $\ell^a_{t,r}$, the delegation level $p(t)$ can be estimated by the fraction of delegated tasks, while $\theta(t)$ and θ_a can be estimated from randomized, matched no-AI probe-task losses and AI-only benchmark losses. The decay rate Δ and adaptation rate κ can then be estimated from finite differences on the resulting time series. The fitted system yields an estimated separatrix $p = \widehat{\psi}(\theta)$. The model predicts that learners initialized on opposite sides of this boundary diverge in long-run reliance and independent skill, while learners initialized on the same side converge to the same long-run regime. Section E gives the full calibration procedure, a synthetic illustration, and falsifiable predictions for a proposed longitudinal test.

3.2 Overview of the proofs

We present the main mathematical ingredients; full proofs appear in Section B.

Proof overview of Theorem 3.1. Write the vector field in (7) as

$$f(\theta, p) = \theta(1-\theta)\big((1-p)(1-\theta) - \Delta p\theta\big), \qquad h(\theta, p) = \kappa p(1-p)\big((1-\theta)^2 - (1-\theta_a)^2\big).$$

The equations $f = h = 0$ give the four corner equilibria through the boundary factors $\theta(1-\theta)$ and $p(1-p)$. In the interior, the nullclines are $(1-p)(1-\theta) - \Delta p\theta = 0$, and $(1-\theta)^2 - (1-\theta_a)^2 = 0$. Since $\theta, \theta_a \in [0,1]$, the second nullcline gives $\theta = \theta_a$, and the first then gives $(\theta^\dagger, p^\dagger) = \left(\theta_a, \frac{1-\theta_a}{1-(1-\Delta)\theta_a}\right)$.

Stability is determined from the Jacobian $J(\theta, p) = \begin{pmatrix} \partial_\theta f & \partial_p f \\ \partial_\theta h & \partial_p h \end{pmatrix}$. At the corners, the off-diagonal entries vanish, and the stability within the square follows from the diagonal terms together with the one-dimensional boundary drifts. This gives two locally asymptotically stable corners relative to the square, $(1,0)$ and $(0,1)$, and two repelling corners, $(0,0)$ and $(1,1)$. The stable corners are non-hyperbolic because one direction is tangent to the boundary. At the interior equilibrium, the nullcline conditions simplify the Jacobian. In particular, $\partial_p h(\theta^\dagger, p^\dagger) = 0, \partial_p f(\theta^\dagger, p^\dagger) < 0, \partial_\theta h(\theta^\dagger, p^\dagger) < 0$. Therefore, $\det J(\theta^\dagger, p^\dagger) = -\partial_p f(\theta^\dagger, p^\dagger)\partial_\theta h(\theta^\dagger, p^\dagger) < 0$. Thus, the interior equilibrium has one positive and one negative eigenvalue and is a saddle. This negative determinant is the local signature of the coupling: higher delegation lowers the skill drift, while lower skill raises the delegation drift.

Proof overview of Theorem 3.2. The main difficulty is global geometry: the stable manifold of the saddle point must form a single separatrix rather than fold, cycle or create disconnected basin boundaries.

By the stable manifold theorem [28], the interior saddle has a one-dimensional local stable manifold $W^s(\theta^\dagger, p^\dagger)$. The stable eigenvector gives this local curve a positive slope in (θ, p)-coordinates. Let Γ be one branch and continue it backward under the flow:

$$\Gamma^- = \{\varphi_{-t}(x) : x \in \Gamma,\ t \geq 0\}.$$

The boundary of $[0,1]^2$ is invariant and solutions are unique, so an interior branch cannot hit the boundary in finite time. Suppose, for contradiction, that its backward continuation stays a positive distance from the boundary. Then its alpha-limit set is a compact invariant set in $(0,1)^2$. With the Dulac function

$$D(\theta, p) = \frac{1}{\theta(1-\theta)p(1-p)},$$

we can compute

$$\partial_\theta(Df) + \partial_p(Dh) = -\tfrac{(1-p)+\Delta p}{p(1-p)} < 0 \qquad \text{on } (0,1)^2.$$

Hence, the Bendixson-Dulac criterion [28] rules out periodic or homoclinic orbits in the interior. Since the saddle $(\theta^\dagger, p^\dagger)$ is the only interior equilibrium, the planar limit-set classification would then force the backward branch to converge to that saddle, producing a homoclinic orbit. This contradiction shows that the closure of the backward branch intersects $\partial[0,1]^2$. Applying this to both branches extends W^s globally.

The global stable manifold is invariant, so trajectories cannot cross it. Below, we show that its two backward branches approach $(0,0)$ and $(1,1)$, dividing the interior into two connected components. Since every interior trajectory converges to an equilibrium (Theorem A.3) and the only attracting equilibria are $(1,0)$ and $(0,1)$, the two open attracting basins cover $(0,1)^2 \setminus W^s$. Connectedness therefore places each component in one basin. Conversely, any point on the interior boundary of either attracting basin cannot converge to an attracting or repelling corner, and therefore must converge to the saddle. Thus, the resulting invariant stable manifold $W^s(\theta^\dagger, p^\dagger)$ acts as a separatrix in the planar phase portrait [28].

It remains to represent this boundary as $p = \psi(\theta)$. Set $x = 1-\theta$. In (x,p)-coordinates,

$$\dot{x} = x(1-x)\big[\Delta p(1-x) - (1-p)x\big], \qquad \dot{p} = \kappa p(1-p)\big[x^2 - (1-\theta_a)^2\big].$$

The off-diagonal derivatives satisfy

$$\tfrac{\partial \dot{x}}{\partial p} = x(1-x)\big[\Delta(1-x) + x\big] \geq 0, \qquad \tfrac{\partial \dot{p}}{\partial x} = 2\kappa p(1-p)x \geq 0.$$

Thus, the system is cooperative in (x,p). Kamke's comparison theorem for monotone dynamical systems [34] gives the comparison principle: larger initial skill gap x_0, larger initial delegation p_0, or larger AI skill θ_a weakly increase future skill gap and delegation (Lemma B.2). It follows that the low-skill basin is upward closed in skill gap and delegation.

Strict positivity of the off-diagonal derivatives in the interior makes the flow strictly order preserving. Suppose two distinct points $z_i = (x_i, p_i) \in W^s$ are ordered by $x_1 \leq x_2, p_1 \leq p_2$. Their forward images remain ordered and eventually lie on the local stable curve near the saddle. But this curve is strictly decreasing in (x,p), so it contains no two distinct ordered points, a contradiction. Hence, distinct points of W^s satisfy $x_1 \neq x_2$ and $x_1 < x_2 \to p_1 > p_2$. Thus, each x corresponds to at most one p. Since $\theta = 1-x$, W^s is a strictly increasing graph $p = \psi(\theta)$ over its projection onto the θ-axis.

The decreasing skill nullcline $p = q(\theta) := (1-\theta)/(1-\theta+\Delta\theta)$ intersects the increasing stable graph W^s only at the saddle. Thus, $f \neq 0$ elsewhere on W^s, while at the saddle the stable eigenvector has a nonzero θ-component. Hence, the smooth curve W^s has no vertical tangencies, so ψ is differentiable. The same nullcline comparison shows that, backward in time, both coordinates decrease along the branch below the saddle and increase along the branch above it. Their bounded monotone limits must be equilibria, respectively $(0,0)$ and $(1,1)$. Hence, the graph spans $(0,1)$, with endpoint limits 0 and 1. Order preservation and convergence then identify the two basins.

Proof overview of Theorem 3.3. Fix $\theta_{a,1} < \theta_{a,2}$, and let ψ_1, ψ_2 be the corresponding separatrices. Suppose for contradiction that $\psi_2(\theta^*) > \psi_1(\theta^*)$ for some θ^*. Choose p_0 with $\psi_1(\theta^*) < p_0 < \psi_2(\theta^*)$. Then (θ^*, p_0) lies in the low-skill basin for $\theta_{a,1}$, but in the high-skill basin for $\theta_{a,2}$. Hence, $\theta^\infty(\theta_{a,1}) = 0$ while $\theta^\infty(\theta_{a,2}) = 1$. But Lemma B.2 implies that increasing θ_a weakly decreases $\theta(t)$ for all $t \geq 0$, so $\theta^\infty(\theta_{a,2}) \leq \theta^\infty(\theta_{a,1})$, contradicting $0 < 1$. Thus, $\psi_{\theta_a}(\theta)$ is nonincreasing in θ_a.

Differentiability with respect to θ_a follows from the parameter-dependent stable manifold theorem [28], since the saddle remains hyperbolic for every $\theta_a \in (0,1)$.

The sweeping property follows from continuity in θ_a and the endpoint limits $\lim_{\theta_a \downarrow 0} \psi_{\theta_a}(\theta) = 1, \lim_{\theta_a \uparrow 1} \psi_{\theta_a}(\theta) = 0$. These limits are obtained by comparing with the limiting systems $\theta_a = 0$ and $\theta_a = 1$, using finite-time continuous dependence and invariant attracting rectangles. At $\theta_a = 0$, delegation is nonincreasing and the high-skill basin fills the interior; at $\theta_a = 1$, delegation is nondecreasing and the low-skill basin fills the interior. The intermediate value theorem then gives, for any $(\theta_0, p_0) \in (0,1)^2$, a value of θ_a such that $p_0 = \psi_{\theta_a}(\theta_0)$.

Proof overview of Theorem 3.5. Expanding $\ell(\theta, p) = (1-p)(1-\theta)^2 + p(1-\theta_a)^2$ gives $G_\ell(t) =$

$$(1-p(t;\theta_0,p_0))\Big[(1-\theta(t;\theta_0,p_0))^2-(1-\theta(t;\theta_0,0))^2\Big]+p(t;\theta_0,p_0)\Big[(1-\theta_a)^2-(1-\theta(t;\theta_0,0))^2\Big].$$

Lemma B.2 gives that for any $t > 0$, $\theta(t;\theta_0,p_0) \le \theta(t;\theta_0,0)$, so the first bracket is nonnegative. The no-AI trajectory satisfies $\dot\theta = \theta(1-\theta)^2$, hence increases monotonically from θ_0 to 1. Since $\theta_0 < \theta_a$, there exists a finite t^* with $\theta(t^*;\theta_0,0) = \theta_a$. For $t > t^*$, the second bracket is strictly positive, and $p(t;\theta_0,p_0) > 0$. Hence, $G_\ell(t) > 0$ for all $t > t^*$, so $t_c \le t^* < \infty$. Moreover, once the no-AI trajectory exceeds θ_a, the performance gap is strictly positive; therefore $\theta(t_c;\theta_0,0) \le \theta_a$.

4 Alternative specifications and robustness

The main specification in (7) uses scalar skill, squared performance loss, and exact perception of fixed AI skill. We summarize variants that relax these choices. The main two-dimensional model is analyzed theoretically; the variants below are robustness checks, supported by exact reductions, global or local analysis, or simulations as indicated in Section D. Most variants preserve the two-basin structure, while noisy feedback and evolving AI are discussed separately below. Thus, the basin mechanism is not tied to the particular squared-loss, scalar-skill specification used for the main phase-plane analysis. Full derivations, simulations, and intervention variants appear in Sections D and C.

Performance and perception.

Jagged AI performance (Section D.1.1). If AI performance varies across tasks, the AI-side loss in the delegation update becomes $\mathbb{E}_{s\sim\mu_a}[(1-s)^2]$. Defining $(1-\theta_a^{\text{eff}})^2 = \mathbb{E}_{s\sim\mu_a}[(1-s)^2]$ reduces the expected dynamics to the main specification with θ_a replaced by θ_a^{eff}. Thus, jaggedness enters through expected task loss while preserving the basin structure when $\theta_a^{\text{eff}} \in (0,1)$.

Biased perceived AI skill (Section D.1.2). Learners may update reliance using a perceived AI skill $\widetilde{\theta}_a$ rather than the true effective skill θ_a. This replaces $(1-\theta_a)^2$ by $(1-\widetilde{\theta}_a)^2$ in the delegation update, while realized performance is still evaluated using θ_a. For $\widetilde{\theta}_a \in (0,1)$, overestimation expands the low-skill basin; underestimation has the opposite effect.

Stochastic performance-gap feedback (Section D.1.3). To capture noisy performance comparisons, we use $dp = \kappa p(1-p)\big([(1-\theta)^2-(1-\theta_a)^2]dt + \sigma dW_t\big)$, where W_t is standard Brownian motion and $\sigma \ge 0$, while keeping the skill update unchanged. Simulated finite-horizon probabilities of ending in the high-skill, low-reliance region change sharply near the deterministic separatrix, which therefore remains informative about outcomes under noise.

Power-z performance and non-use losses (Section D.1.4). The squared loss can be replaced by $(1-\theta)^z$ for $z \ge 1$. This changes both the performance comparison in the delegation update and the error-driven learning and non-use decay terms in the skill update. The nullclines deform with z, but the two-basin phase structure is observed to persist. The interior equilibrium remains a saddle for every $z \ge 1$.

Skill formation and assisted production.

Learning from AI during delegation (Section D.2.1). Delegation may provide a learning signal when the learner studies AI-generated outputs. Adding an AI-learning term to the skill update pulls θ toward θ_a. In the main specification, the low-skill equilibrium shifts from $(0,1)$ to $\left(\frac{\beta\theta_a}{\beta+\Delta}, 1\right)$, but a low-skill, high-delegation regime is observed to persist.

Within-task AI assistance (Section D.2.2). To capture shared human–AI work, let $u \in [0,1]$ denote AI contribution, with loss $L = (1-u)(1-\theta)^2 + u(1-\theta_a)^2$ and skill potential $\Phi_s = (1-u)(1-\theta)^2 +$

$\Delta u\theta^2$. The same multiplicative updates recover ODE (7) with $p = u$, so the two-basin structure also describes this additive form of collaboration.

Skill-dependent AI performance (Section D.2.3). In some tasks, AI quality depends on the learner's ability to prompt or contextualize the output. We model this by replacing the AI-side loss in Φ_d with $(1 - w\theta - (1 - w)\theta_a)^2$, $w \in [0, 1]$, while leaving the skill update unchanged. As w increases, the delegation signal depends more strongly on learner skill, reshaping the basin boundary. For $w < 1$, the resulting dynamics interpolate between AI performance that is independent of the learner and AI performance that depends on the learner's skill; the two-basin structure is observed to persist across this range and the interior equilibrium also remains a saddle.

Richer learner and AI states.

Multiple skills (Section D.3.1). Skill can be multidimensional. For example, replacing θ by (θ_1, θ_2) and letting performance depend on $\bar{\theta} = w_1\theta_1 + w_2\theta_2$ yields a system over (θ_1, θ_2, p). In the simulations, the basin boundary becomes a separating surface rather than a curve, but bistability and path dependence persist. Local analysis establishes attraction of both extreme equilibria and, when feasible, an interior saddle with a two-dimensional local stable manifold (Proposition D.1).

Improving AI capability (Section D.3.2). If AI capability improves over time, the delegation update becomes time-dependent through $(1-\theta_a(t))^2$, while the skill update is unchanged. As $\theta_a(t)$ increases, delegation becomes attractive over a larger range of skill levels. The frozen-capability separatrix shifts as AI improves; it is not an invariant boundary of the evolving system.

5 Conclusion, limitations, and future work

We developed a dynamical framework for repeated AI-assisted learning in which human skill and AI reliance co-evolve. In the main specification, adaptive delegation changes the global geometry of learning: fixed delegation yields a one-dimensional learning-decay process with a single attractor, whereas adaptive delegation can produce two attracting regimes separated by the stable manifold of an interior saddle. This boundary determines which initial skill-reliance states lead to sustained learning and which lead to persistent delegation and skill loss. Increasing AI capability reshapes the boundary, expanding the low-skill basin even as assistance appears beneficial for longer.

The resulting risk is not AI delegation itself, but the feedback loop between short-run performance gains and reduced opportunities for practice. In this sense, the model formalizes a mechanism for cognitive debt: assistance can improve immediate outcomes while shifting the learner toward a region of the state space from which independent skill is harder to recover. Interventions that reduce early reliance, preserve opportunities for practice, or modify the feedback signal associated with delegation can be understood as attempts to move the learner across the basin boundary or shift the boundary itself toward sustained skill development.

The model is intentionally stylized. It isolates repeated adaptive delegation rather than the full space of human-AI workflows, and it is meant to characterize qualitative dynamics rather than serve as a calibrated structural model of human behavior. Across alternative specifications, the phase structure is observed to persist under jagged or misperceived AI performance, alternative losses, multiple skills, learning from AI, and skill-dependent AI performance. With changing AI capability, frozen-capability boundaries describe the instantaneous comparison rather than the long-run basin structure. Section E outlines how the model parameters can be estimated from repeated-task data, illustrates the calculation on a synthetic example, and states empirical predictions that would support or challenge the basin mechanism. Future work should connect these dynamics to richer workflows that separate generation, review, repair, verification, and effort allocation to test which activities provide practice and preserve independent skill.

Acknowledgments and Disclosure of Funding. We thank Mariia Eremeeva for useful discussions. LH is supported in part by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM118), NSFC Grant No. 625707396. NKV was supported in part by NSF Grant CCF-2112665. The authors received no other relevant third-party funding or support and report no other relevant financial relationships during the 36 months preceding submission. LH began working part-time at Sytrix after submission; the company provided no funding or other support for this research.

References

[1] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Comput.*, 10(2):251–276, February 1998.

[2] Long Bai, Xiangfei Liu, and Jiacan Su. ChatGPT: The cognitive effects on learning and memory. *Brain-X*, 2023.

[3] Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences of the United States of America*, 122, 2025.

[4] Yoram Ben-Porath. The production of human capital and the life cycle of earnings. *Journal of Political Economy*, 75(4, Part 1):352–365, 1967.

[5] Michel Benaïm. Dynamics of stochastic approximation algorithms. In Jacques Azéma, Michel Émery, Michel Ledoux, and Marc Yor, editors, *Séminaire de Probabilités XXXIII*, pages 1–68, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.

[6] Shreyan Biswas, Alexander Erlei, and Ujwal Gadiraju. Belief updating and delegation in multi-task human–ai interaction: Evidence from controlled simulations. *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 2026.

[7] Hermann Ebbinghaus. *Memory: A Contribution to Experimental Psychology*. Teachers College, Columbia University, 1913. Transl. of 1885 German edition.

[8] Asma Ejaz, Muhammad Farhan, Francesco Ernesto, and Alessi Longa. AI and cognitive load: How reliance on AI tools (ChatGPT, etc.) affects critical thinking. *Research Journal of Psychology*, 2025.

[9] Susan E. Embretson and Steven P. Reise. *Item Response Theory for Psychologists*. Psychology Press, 1 edition, 2000.

[10] Ido Erev and Alvin E. Roth. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *The American Economic Review*, 88(4):848–881, 1998.

[11] Thomas L. Griffiths, Charles Kemp, and Joshua B. Tenenbaum. Bayesian models of cognition. In Ron Sun, editor, *The Cambridge Handbook of Computational Psychology*, pages 59–100. Cambridge University Press, 2008.

[12] Andrew Heathcote, Scott Brown, and D. J. K. Mewhort. The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, 7(2):185–207, Jun 2000.

[13] Morris W. Hirsch, Stephen Smale, and Robert L. Devaney. *Differential Equations, Dynamical Systems, and an Introduction to Chaos*. Academic Press, 3 edition, 2013.

[14] Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998.

[15] Lingxiao Huang, Wenyang Xiao, and Nisheeth K. Vishnoi. Delegation and verification under ai. *To appear in ICML*, 2026.

[16] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. 2025.

[17] Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *CHI 2025*, April 2025.

[18] Matthias Lehmann, Philipp B. Cornelius, and Fabian J. Sting. AI meets the classroom: When do large language models harm learning?, 2024.

[19] Grace Liu, Brian Christian, Tsvetomira Dumbalska, Michiel A. Bakker, and Rachit Dubey. Ai assistance reduces persistence and hurts independent performance. *arXiv preprint arXiv:2604.04721*, 2026.

[20] Frederic M. Lord. Applications of item response theory to practical testing problems. 1980.

[21] Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 230–239, New York, NY, USA, 2019. Association for Computing Machinery.

[22] Allen Newell and Paul S. Rosenbloom. Mechanisms of skill acquisition and the law of practice. In John R. Anderson, editor, *Cognitive Skills and Their Acquisition*, pages 1–55. Lawrence Erlbaum, Hillsdale, NJ, 1981.

[23] Donald A. Norman. *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine*. Addison-Wesley, 1994.

[24] R. Parasuraman, T.B. Sheridan, and C.D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3):286–297, 2000.

[25] Philip I. Pavlik Jr. and John R. Anderson. Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, 29(4):559–586, 2005.

[26] Robin Pemantle. A survey of random processes with reinforcement. *Probability Surveys [electronic only]*, 4:1–79, 2007.

[27] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7599–7609. PMLR, 13–18 Jul 2020.

[28] Lawrence Perko. *Differential equations and dynamical systems*. New York: Springer, 2001.

[29] R. A. Rescorla and A. R. Wagner. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black and W. F. Prokasy, editors, *Classical Conditioning II: Current Research and Theory*, pages 64–99. Appleton-Century-Crofts, New York, 1972.

[30] Evan F. Risko and Sam J. Gilbert. Cognitive offloading. *Trends in Cognitive Sciences*, 20(9):676–688, 2016.

[31] William H. Sandholm. *Population Games and Evolutionary Dynamics*. MIT Press, 2010.

[32] Judy Hanwen Shen and Alex Tamkin. How AI impacts skill formation. 2026.

[33] Joongi Shin, Anna Polyanskaya, Andrés Lucero, and Antti Oulasvirta. No evidence for LLMs being useful in problem reframing. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, page 1–25. ACM, April 2025.

[34] Hal L. Smith. *Monotone Dynamical Systems: An Introduction to the Theory of Competitive and Cooperative Systems*. American Mathematical Society, 1995.

[35] Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv*, abs/2306.07899, 2023.

[36] Chunpeng Zhai, Santoso Wibowo, and Lily D. Li. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments*, 11, 2024.

Contents

1 Introduction **1**

2 Model **3**

3 Theoretical results **5**

3.1 Analyzing performance loss across time 6

3.2 Overview of the proofs 7

4 Alternative specifications and robustness **9**

5 Conclusion, limitations, and future work **10**

A Model derivation and basic properties **15**

A.1 From task completion to window-level updates 15

A.2 Conditional drift and finite-horizon ODE approximation 17

A.3 Invariance and convergence of the general ODE 17

A.4 Equilibria with a positive default skill 19

B Global dynamics and comparison systems **20**

B.1 Feedback directions do not imply bistability 20

B.2 Proof of Theorem 3.1: equilibria and stability 21

B.3 A comparison principle for skill gap and reliance 22

B.4 Proof of Theorem 3.2: the global separatrix 23

B.5 Proof of Theorem 3.3: AI capability and basin ordering 24

B.6 Proof of Theorem 3.5: performance reversal 25

B.7 Effects of the remaining parameters on the separatrix 26

B.8 A local-to-global approximation of the separatrix 27

C Reliance adaptation and interventions **28**

C.1 A global result for centered monotone responses 29

C.2 Conservative reliance updates 30

C.3 Adjustment friction and external constraints 30

C.4 Delegation costs and learner-specific thresholds 31

C.5 Detection-based delegation penalties 32

C.6 Non-performance benefits of delegation 32

D Alternative performance, learning, and state specifications **33**

D.1 Performance and perception 33

D.1.1 Jagged AI performance 33

D.1.2 Biased perceived AI skill 34

D.1.3 Stochastic performance-gap feedback 34

D.1.4 Power-z performance and non-use losses 35

D.2 Skill formation and assisted production 36

D.2.1 Learning from AI during delegation 36

D.2.2 Within-task AI assistance 36

D.2.3 Skill-dependent AI performance 36

D.3 Richer learner and AI states 37

D.3.1 Multiple skills . . . 37
D.3.2 Improving AI capability . . . 40
D.4 Assumptions and scope of the conclusions . . . 41

E Empirical calibration and falsifiable predictions **43**
E.1 Measurements and scale-free identification . . . 43
E.2 Measurement-aware trajectory fitting and uncertainty . . . 44
E.3 Synthetic recovery of parameters and basin membership . . . 45
E.4 A prospective longitudinal test . . . 45

A Model derivation and basic properties

We derive the discrete updates, establish finite-time tracking of ODE (6), and analyze its equilibria and convergence with positive default skill. Table 1 summarizes the notation used in this paper.

Table 1: Notation used in the model and its analysis. Rates in the normalized ODE are relative to the skill-acquisition time scale.

Symbol	Domain	Description
θ_d	$[0,1]$	Default skill
θ_a	$[0,1]$	AI skill
θ_0	$[0,1]$	Initial skill of learner
$\theta(t)$	$[0,1]$	Learner skill at time t
$\theta(t;\theta_0,p_0)$	$[0,1]$	Learner skill at time t with initial state (θ_0,p_0)
p_0	$[0,1]$	Initial delegation level of learner
$p(t)$	$[0,1]$	Delegation level at time t
$p(t;\theta_0,p_0)$	$[0,1]$	Delegation level at time t with initial state (θ_0,p_0)
Δ	$(0,\infty)$	Relative non-use decay rate
κ	$(0,\infty)$	Relative delegation-adjustment rate
$d(\theta,\theta')$	$[0,1]^2 \to \mathbb{R}_{\geq 0}$	Distance between two skills θ and θ'
$\ell(\theta,p)$	$[0,1]^2 \to \mathbb{R}_{\geq 0}$	Instantaneous performance loss w.r.t. skill θ and delegation level p
$\psi(\theta)$	$[0,1] \to [0,1]$	Stable-manifold height; endpoint values refer to its closure
$G_\ell(t)$	$\mathbb{R}_{\geq 0} \to \mathbb{R}$	Performance gap with versus without AI at time t
t_c	$\mathbb{R}_{\geq 0}$	Final crossing time for performance loss
$\ell(\theta)$	$[0,1] \to \mathbb{R}_{\geq 0}$	Independent expected performance loss
$g(\theta,\theta_d)$	$[0,1]^2 \to \mathbb{R}_{\geq 0}$	Non-use potential measuring distance from default skill
$\varphi_t(z_0)$	$[0,1]^2$	ODE flow from initial state z_0
η, m	$(0,\infty), \mathbb{N}$	Discrete update step and tasks per observation window

A.1 From task completion to window-level updates

We derive the dynamics corresponding to the repeated-task model in Section 2. Time is indexed by short windows $k = 0, 1, 2, \ldots$, and we write $z_k = (\theta_k, p_k)$. Within each window k, the learner faces m comparable tasks. The learner's skill θ_k and delegation level p_k are held fixed during this window, while the task-level delegation decisions vary across tasks.

For each task $r \in [m]$, let $X_{k,r} \in \{0,1\}$ denote the delegation indicator, where $X_{k,r} = 0$ means that the learner performs the task independently and $X_{k,r} = 1$ means that the task is delegated to AI. We model

$$X_{k,r} \mid \mathcal{F}_k \sim \mathrm{Bern}(p_k), \qquad r = 1, \ldots, m,$$

independently conditional on the history $\mathcal{F}_k$ before window k, which determines z_k. Let

$$\bar{X}_k := \frac{1}{m} \sum_{r=1}^{m} X_{k,r}$$

denote the realized fraction of tasks delegated to AI in window k.

In particular,

$$\mathbb{E}[\bar{X}_k \mid \mathcal{F}_k] = p_k, \qquad \mathrm{Var}(\bar{X}_k \mid \mathcal{F}_k) = \frac{p_k(1-p_k)}{m}.$$

The learning mechanism within one time window is as follows.

• **(Task completion stage)** For each task $r \in [m]$, the learner delegates the task to AI if $X_{k,r} = 1$ and performs it independently if $X_{k,r} = 0$.

• **(Evaluation stage)** The learner submits the resulting output to the teacher and receives task-level feedback. Under the baseline conditional-loss specification, if the task is performed independently, the performance loss is $\ell(\theta_k)$; if the task is delegated to AI, the performance loss is $\ell(\theta_a)$. Thus the realized task-level performance loss is

$$\ell_{k,r} = (1 - X_{k,r})\ell(\theta_k) + X_{k,r}\ell(\theta_a).$$

• **(Skill update)** Independent task execution produces an error-driven learning signal, while delegation produces non-use decay toward the default skill θ_d. We encode these two effects by the task-level skill-update potential

$$\Phi_{s,k,r} = (1 - X_{k,r})\ell(\theta_k) + \Delta X_{k,r} g(\theta_k, \theta_d).$$

Averaging over the m tasks in the window gives

$$\bar{\Phi}_s(\theta_k) = \frac{1}{m} \sum_{r=1}^{m} \Phi_{s,k,r}.$$

Using the multiplicative-update metric $F_\theta(\theta) = 1/(\theta(1-\theta))$, the skill update is

$$\theta_{k+1} = \theta_k - \eta \theta_k (1 - \theta_k) \partial_\theta \bar{\Phi}_s(\theta_k).$$

• **(Delegation update)** The delegation level is updated according to the relative performance of AI-assisted and independent work. The delegation-update potential is the expected performance loss

$$\Phi_d(\theta_k, p_k) = (1 - p_k)\ell(\theta_k) + p_k \ell(\theta_a).$$

Using the multiplicative-update metric $F_p(p) = 1/(p(1-p))$, the delegation update is

$$p_{k+1} = p_k - \eta \kappa p_k (1 - p_k) \partial_p \Phi_d(\theta_k, p_k).$$

Equivalently,

$$p_{k+1} = p_k + \eta \kappa p_k (1 - p_k) \big(\ell(\theta_k) - \ell(\theta_a)\big).$$

Combining the skill and delegation updates, the discrete-time stochastic dynamics are

$$\begin{aligned} &X_{k,r} \sim \mathrm{Bern}(p_k), \qquad r = 1, \dots, m, \\ &\theta_{k+1} = \theta_k - \eta \theta_k (1 - \theta_k) \left[(1 - \bar{X}_k) \partial_\theta \ell(\theta_k) + \Delta \bar{X}_k \partial_\theta g(\theta_k, \theta_d) \right], \\ &p_{k+1} = p_k + \eta \kappa p_k (1 - p_k) \big(\ell(\theta_k) - \ell(\theta_a)\big). \end{aligned} \tag{9}$$

Thus task allocation is random, but the reliance update uses the conditional expected performance gap. Sections D.1.1 and D.1.3 separately study random feedback in the reliance update.

For the quadratic potentials

$$\ell(\theta) = (1 - \theta)^2, \qquad g(\theta, \theta_d) = (\theta - \theta_d)^2,$$

we have

$$\partial_\theta \ell(\theta) = -2(1 - \theta), \qquad \partial_\theta g(\theta, \theta_d) = 2(\theta - \theta_d).$$

Substituting these derivatives into (9) yields

$$\begin{aligned} \theta_{k+1} - \theta_k &= 2\eta \theta_k (1 - \theta_k) \big[(1 - \bar{X}_k)(1 - \theta_k) + \Delta \bar{X}_k (\theta_d - \theta_k) \big], \\ p_{k+1} - p_k &= \eta \kappa p_k (1 - p_k) \big[(1 - \theta_k)^2 - (1 - \theta_a)^2 \big]. \end{aligned} \tag{10}$$

For the explicit discrete updates, a sufficient step-size condition is

$$\eta \max\{2 \max(1, \Delta), \kappa\} < 1.$$

It preserves $[0,1]^2$ and its interior. Indeed, $u^+ = u + \eta u(1-u)v$ stays in $[0,1]$ whenever $u \in [0,1]$ and $|\eta v| \le 1$. The continuous-time invariance proved below does not require this discrete step-size condition.

Remark A.1 (Belief-based interpretation of the multiplicative factor). A belief-based interpretation also motivates the multiplicative factor: if $\mathrm{logit}\, p = \alpha + \beta b$, then $dp/db = \beta p(1-p)$. Feedback can change beliefs and delegation [6]; proportionality between belief changes and the performance gap remains a modeling assumption.

A.2 Conditional drift and finite-horizon ODE approximation

We show that, as the step size $\eta \to 0$, the stochastic trajectory of Dynamics (10) tracks the solution of ODE (6) on finite time intervals. The key point is that each time window contains task-level decisions, whose average delegation rate is $\bar{X}_k$.

Lemma A.2 (Expected discrete dynamics). *Let $\mathcal{F}_k$ denote the history before window k, including the current state (θ_k, p_k). For Dynamics* (10)*, we have*

$$\begin{aligned} \mathbb{E}[\theta_{k+1} \mid \mathcal{F}_k] &= \theta_k + 2\eta\theta_k(1-\theta_k)\big[(1-p_k)(1-\theta_k) + \Delta p_k(\theta_d - \theta_k)\big], \\ p_{k+1} &= p_k + \eta\kappa p_k(1-p_k)\big[(1-\theta_k)^2 - (1-\theta_a)^2\big]. \end{aligned} \tag{11}$$

Proof. The update for p_{k+1} is deterministic conditional on $\mathcal{F}_k$. The skill update depends on the realized fraction of delegated tasks $\bar{X}_k = m^{-1}\sum_{r=1}^{m} X_{k,r}$. Since the $X_{k,r}$ are independent $\mathrm{Bern}(p_k)$ variables conditional on $\mathcal{F}_k$,

$$\mathbb{E}[\bar{X}_k \mid \mathcal{F}_k] = p_k.$$

The update for θ_{k+1} in (10) is affine in $\bar{X}_k$. Substituting this conditional expectation gives the first identity in (11). □

The drift field associated with (11) is

$$F(\theta, p) = \big(2\theta(1-\theta)\big((1-p)(1-\theta) + \Delta p(\theta_d - \theta)\big),\ \kappa p(1-p)\big((1-\theta)^2 - (1-\theta_a)^2\big)\big),$$

which is exactly the right-hand side of ODE (6).

The noise term and time scale. Write $z_{k+1} = z_k + \eta F(z_k) + \eta\xi_{k+1}$, where

$$\xi_{k+1} = \begin{pmatrix} 2\theta_k(1-\theta_k)[\Delta(\theta_d - \theta_k) - (1-\theta_k)](\bar{X}_k - p_k) \\ 0 \end{pmatrix},$$

$$\mathbb{E}[\xi_{k+1} \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}[\|\xi_{k+1}\|^2 \mid \mathcal{F}_k] \le C/m.$$

Let $z^\eta(s)$ be the linear interpolation through z_k at continuous times $s = k\eta$, and let $z(s)$ solve $\dot{z} = F(z)$ from the same initial state. For every fixed horizon $S_0 < \infty$, the accumulated martingale satisfies

$$\mathbb{E} \max_{j \le S_0/\eta} \left\| \eta \sum_{k<j} \xi_{k+1} \right\|^2 \le C_{S_0} \frac{\eta}{m}.$$

On the invariant compact square, F is bounded and Lipschitz. The interpolation remainder is $O(\eta)$; the preceding martingale bound and Gronwall's inequality therefore give

$$\mathbb{E} \sup_{0 \le s \le S_0} \|z^\eta(s) - z(s)\|^2 \le C_{S_0}\left(\eta^2 + \frac{\eta}{m}\right) \longrightarrow 0.$$

This proves finite-time tracking within the stochastic-approximation framework [5]. It does not justify interchanging $s \to \infty$ and $\eta \to 0$, or equating finite-sample stochastic outcomes with ODE basins.

A.3 Invariance and convergence of the general ODE

We establish basic well-posedness properties of the general ODE (6). For this subsection write $a = \theta_a$ and $d = \theta_d$, and define

$$\begin{aligned} \dot{\theta} &= 2\theta(1-\theta)A_d(\theta, p), \qquad \dot{p} = \kappa p(1-p)D_a(\theta), \\ A_d &= (1-p)(1-\theta) + \Delta p(d-\theta), \qquad D_a = (1-\theta)^2 - (1-a)^2. \end{aligned} \tag{12}$$

The factor 2 comes from the quadratic potentials. Rescaling time t to $2t$ removes it and changes κ to $\kappa/2$. Both changes are needed to obtain the normalized model.

Theorem A.3 (Existence, invariance, and convergence). *Assume $\kappa, \Delta > 0$ and $a, d \in [0,1]$. For every initial state $(\theta_0, p_0) \in [0,1]^2$, ODE* (6) *admits a unique global solution $(\theta(t), p(t))$ defined for all $t \ge 0$. Moreover,*

$$(\theta(t), p(t)) \in [0,1]^2 \qquad \textit{for all } t \ge 0.$$

If the initial condition lies in the interior $(0,1)^2$, then the trajectory remains in $(0,1)^2$ for every finite time. In addition, when the equilibria are isolated, every trajectory converges to an equilibrium.

Proof. Let $x = (\theta, p)$ and let $F(x)$ denote the vector field in (12). Each component of F is a polynomial in (θ, p), so F is smooth and hence locally Lipschitz on $\mathbb{R}^2$. Therefore, by the Picard-Lindelöf theorem [13, 28], for every initial condition $x_0 \in \mathbb{R}^2$ there exists a unique maximal solution $x(t)$ on some interval $[0, T_{\max})$.

We first show positive invariance of $[0,1]^2$. Each coordinate has the logistic-barrier form

$$\dot{u}(t) = u(t)(1 - u(t))v(t),$$

where $v(t)$ is continuous along the trajectory. For such an equation, the interval $[0,1]$ is positively invariant: if $u(t)$ reaches 0 or 1, the right-hand side vanishes at that boundary, and uniqueness prevents the trajectory from crossing it. Applying this argument to both $\theta(t)$ and $p(t)$ gives $\theta(t), p(t) \in [0,1]$ for all $t < T_{\max}$. An interior solution satisfies

$$\operatorname{logit} u(t) = \operatorname{logit} u(0) + \int_0^t v(s)\, ds.$$

Boundedness of v on the square excludes finite-time contact with its boundary. The same reasoning applies in backward time when continuing an interior invariant curve.

Since $[0,1]^2$ is compact and F is continuous, the vector field is bounded on $[0,1]^2$. Hence a solution starting in $[0,1]^2$ cannot blow up in finite time. Because F is smooth on all of $\mathbb{R}^2$, there is also no singularity of the vector field. Thus the maximal existence time satisfies $T_{\max} = +\infty$.

It remains to rule out periodic orbits in the interior. On $(0,1)^2$, define the Dulac function

$$\mathcal{D}(\theta, p) = \frac{1}{\theta(1-\theta)p(1-p)}.$$

Using A_d and D_a from (12), we have

$$\mathcal{D}\dot{\theta} = \frac{2A_d(\theta, p)}{p(1-p)}, \qquad \mathcal{D}\dot{p} = \frac{\kappa D_a(\theta)}{\theta(1-\theta)}.$$

Therefore,

$$\frac{\partial}{\partial\theta}(\mathcal{D}\dot{\theta}) + \frac{\partial}{\partial p}(\mathcal{D}\dot{p}) = \frac{2\partial_\theta A_d(\theta, p)}{p(1-p)} = -\frac{2[(1-p) + \Delta p]}{p(1-p)} < 0$$

for all $(\theta, p) \in (0,1)^2$. By the Bendixson-Dulac criterion [28], there are no periodic orbits in the interior. The same sign excludes interior homoclinic loops, but cycle exclusion alone does not establish convergence, because limit sets may meet the boundary. The following sign argument also covers such trajectories.

Case $0 \le d < a < 1$**.** If an interior trajectory remains below d, then $\dot{\theta} > 0$ and

$$\frac{d}{dt} \operatorname{logit} p = \kappa D_a(\theta) \ge \kappa D_a(d) > 0.$$

Hence $p \to 1$, and the monotone skill limit must be d. Otherwise it enters the forward-invariant strip $d < \theta < 1$. There,

$$A_d = K_d(\theta)[h_d(\theta) - p], \qquad K_d = 1 - \theta + \Delta(\theta - d) > 0, \quad h_d = \frac{1-\theta}{K_d}, \quad h_d' = -\frac{\Delta(1-d)}{K_d^2} < 0.$$

Put $q = h_d(a)$ and $v = p - h_d(\theta)$. The four sectors have the sign pattern

	$v < 0$		$v > 0$	
$\theta < a$	$(\dot{\theta}, \dot{p}) = (+,+)$	$(-,+)$	$(\mathcal{L}_d)$	(13)
$\theta > a$	$(+,-)$	$(\mathcal{H}_d)$	$(-,-)$	

On $v = 0$, $\dot{v} = \dot{p}$; on $\theta = a$, $\operatorname{sign}\dot{\theta} = \operatorname{sign}(q - p)$. Therefore $\mathcal{H}_d = \{\theta > a, v < 0\}$ and $\mathcal{L}_d = \{d < \theta < a, v > 0\}$ are forward invariant. Coordinatewise monotonicity gives

$$z_0 \in \mathcal{H}_d \Longrightarrow z(t) \to (1, 0), \qquad z_0 \in \mathcal{L}_d \Longrightarrow z(t) \to (d, 1).$$

In either remaining sector, both coordinates are monotone until the trajectory enters $\mathcal{H}_d$ or $\mathcal{L}_d$. If it never enters, it converges to (a, q): any limit must be an equilibrium, and the direction of motion excludes the other equilibria in that sector's closure. Points on either nullcline, except (a, q), enter the adjacent invariant sector. This proves convergence in this case, including $d = 0$.

Case $d > a$**.** Choose $b \in (a, d)$. For $\theta \le b$,

$$A_d(\theta, p) \ge c_b := \min\{1 - b, \Delta(d - b)\} > 0.$$

Thus $\theta = b$ is crossed upwards in finite time if necessary and cannot subsequently be crossed downwards. For all later times,

$$\frac{d}{dt} \operatorname{logit} p \le \kappa D_a(b) < 0 \quad \Longrightarrow \quad p(t) \to 0.$$

Given $\varepsilon > 0$ with $b < 1 - \varepsilon$, eventually $A_d \ge (1 - \theta)/2 > 0$ on $[b, 1 - \varepsilon]$. The trajectory reaches and stays above $1 - \varepsilon$, yielding $\theta \to 1$. This also covers $d = 1$ when $a < 1$.

Case $d = a \in (0, 1)$**.** If $\theta(t) < a$ forever, both coordinates increase and their only possible equilibrium limit is $(a, 1)$. Eventually θ stays in a compact neighborhood below a, so for some $c > 0$,

$$\dot\theta \ge 2\Delta p\theta(1 - \theta)(a - \theta) \ge c(a - \theta), \qquad a - \theta(t) \le Ce^{-ct}.$$

But $D_a(\theta) = (a - \theta)(2 - a - \theta)$ is then integrable, contradicting $p \to 1$ through the logit identity. Hence θ crosses a. Thereafter p decreases and is bounded by some $P < 1$, while

$$A_a(a, p) = (1 - p)(1 - a) \ge (1 - P)(1 - a) > 0.$$

Continuity gives an $\varepsilon > 0$ such that skill eventually stays above $a + \varepsilon$. The preceding argument then gives $(\theta, p) \to (1, 0)$.

If $a = 0$ or $a = 1$, respectively the edge $\theta = 0$ or $\theta = 1$ consists of equilibria, so these cases are excluded by the isolated-equilibrium hypothesis. All boundary-edge dynamics are one-dimensional and converge. This completes the theorem. □

We next list the equilibria of the general system. This is useful for relating the general ODE (6) to the normalized system (7) analyzed in the main text.

A.4 Equilibria with a positive default skill

Write $a = \theta_a$ and $d = \theta_d$.

Theorem A.4 (Equilibria of the general ODE). *Assume* $0 \le d < a < 1$ *and* $\Delta, \kappa > 0$*. Besides the four corners, the equilibria of* (6) *are*

$$e_d = (d, 1), \qquad e^\dagger = (a, q_d), \qquad q_d = \frac{1 - a}{1 - a + \Delta(a - d)}.$$

When $d = 0$*,* $e_d = (0, 1)$*. The points* $(1, 0)$ *and* e_d *are locally asymptotically stable relative to the square;* $(0, 0)$ *and* $(1, 1)$ *are repellers;* $e^\dagger$ *is a hyperbolic saddle. For* $d > 0$*,* $(0, 1)$ *is an additional boundary saddle.*

Proof. The factors in (12) give the corners. On $p = 1$, $\dot\theta = 2\Delta\theta(1 - \theta)(d - \theta)$ adds e_d. In the interior, $D_a = 0$ gives $\theta = a$, and $A_d = 0$ gives $p = q_d$.

The stability claims follow from the Jacobian. Since $A_d = D_a = 0$ at $e^\dagger = (a, q_d)$, the Jacobian has entries

$$\partial_\theta\dot\theta = 2a(1 - a)\partial_\theta A_d < 0, \qquad \partial_p\dot\theta = 2a(1 - a)\partial_p A_d < 0,$$
$$\partial_\theta\dot p = \kappa q_d(1 - q_d)D_a'(a) < 0, \qquad \partial_p\dot p = 0.$$

Thus the determinant is

$$-(\partial_p\dot\theta)(\partial_\theta\dot p) < 0,$$

so the interior equilibrium is a saddle.

At the corners, the Jacobian is diagonal, with spectra

$(0, 0)$	$(2,\ \kappa[1 - (1 - a)^2])$
$(1, 1)$	$(2\Delta(1 - d),\ \kappa(1 - a)^2)$
$(0, 1)$	$(2\Delta d,\ -\kappa[1 - (1 - a)^2])$
$(1, 0)$	$(0,\ -\kappa(1 - a)^2)$.

For $d > 0$, the Jacobian at e_d is upper triangular and has eigenvalues $-2\Delta d(1-d)$ and $-\kappa D_a(d)$, both negative.

The zero eigenvalues at $(1,0)$ and, when $d = 0$, $(0,1)$ require nonlinear stability. Choose $r > a$ and $0 < P < h_d(r)$. The rectangle $[r,1] \times [0,P]$ is forward invariant and contains only $(1,0)$ as an equilibrium. Theorem A.3 implies convergence within it. Such rectangles give arbitrarily small relative neighborhoods of $(1,0)$. For $d = 0$, rectangles $[0,r] \times [P,1]$ with $r < a$ and $P > h_0(r)$ similarly establish attraction and stability of $(0,1)$. □

Setting $d = 0$ and applying the time rescaling described above recovers the normalized equilibrium structure. The next appendix identifies its global basin boundary, not merely its stationary points.

B Global dynamics and comparison systems

In this section, we provide proofs for the results in Section 3 and show how to derive an approximation of the stable manifold ψ. We first show why the feedback directions alone do not imply two basins.

B.1 Feedback directions do not imply bistability

Fixed delegation is already a monostable control in Section 2. To retain adaptive reliance as well, fix $\theta_a \in (0,1)$ and $\kappa > 0$ and consider

$$\dot{\theta} = \theta(1-\theta)(1-\theta-p) =: f_c, \qquad \dot{p} = \kappa p(1-p)(\theta_a - \theta + 1 - 2p) =: h_c. \tag{14}$$

The skill equation is exactly the baseline equation at $\Delta = 1$. The reliance signal combines AI's skill advantage $\theta_a - \theta$ with the mean-reverting term $1 - 2p$. It preserves the baseline's strict interior cross-effects:

$$\partial_p f_c = -\theta(1-\theta) < 0, \qquad \partial_\theta h_c = -\kappa p(1-p) < 0, \qquad \partial_{\theta_a} h_c = \kappa p(1-p) > 0. \tag{15}$$

Proposition B.1 (A globally monostable adaptive control). *Every interior solution of* (14) *converges to* $z^\star = (1-\theta_a, \theta_a)$*, its unique interior equilibrium and only asymptotically stable equilibrium relative to* $[0,1]^2$.

Proof. Define $z = (\theta, p)^\top$ and

$$V_{\theta_a}(z) = \frac{\theta^2}{2} + \theta p - \theta + p^2 - (1+\theta_a)p, \qquad \dot{z} = -\operatorname{diag}(\theta(1-\theta), \kappa p(1-p))\nabla V_{\theta_a}(z). \tag{16}$$

Its Hessian $H = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \succ 0$, and $\nabla V_{\theta_a}(z^\star) = 0$. Strict convexity gives a unique interior equilibrium. Since the update factors vanish at the boundary, convergence still needs to be proved. Let $K(u\|v) = u\log(u/v) + (1-u)\log((1-u)/(1-v))$ be Bernoulli relative entropy. With

$$W(z) = K(1-\theta_a\|\theta) + \kappa^{-1}K(\theta_a\|p), \qquad \partial_v K(u\|v) = \frac{v-u}{v(1-v)},$$

we obtain

$$\dot{W} = -(z-z^\star)^\top \nabla V_{\theta_a}(z) = -(z-z^\star)^\top H(z-z^\star) \le -\lambda_{\min}(H)\|z-z^\star\|^2. \tag{17}$$

Because $z^\star \in (0,1)^2$, $W(z) \to \infty$ as $z \to \partial[0,1]^2$. Every interior trajectory is confined to the compact sublevel set $\{W \le W(z(0))\} \subset (0,1)^2$. The strict Lyapunov inequality implies convergence and asymptotic stability [28]. No boundary equilibrium can be asymptotically stable relative to the square, since every relative neighborhood contains interior points converging to $z^\star$. □

Under the same multiplicative factors, a common potential for the baseline would require equality of mixed partials:

$$(1-\theta) + \Delta\theta = 2(1-\theta) \quad \Longleftrightarrow \quad \theta = \frac{1}{1+\Delta}. \tag{18}$$

This cannot hold on an open set for constant $\Delta > 0$. The counterexample shows that feedback signs alone do not imply bistability; the absence of this common-potential representation does not establish it either.

B.2 Proof of Theorem 3.1: equilibria and stability

Theorem 3.1 (Equilibria of ODE (7); restated). *Let θ_a, κ, Δ be the parameters of ODE (7). When $\theta_a \in (0,1)$, the system has two corner equilibria, $(1,0)$ and $(0,1)$, that are locally asymptotically stable relative to $[0,1]^2$; two repelling corner equilibria, $(0,0)$ and $(1,1)$; and one hyperbolic interior saddle $(\theta^\dagger, p^\dagger) = \left(\theta_a, \frac{1-\theta_a}{1-(1-\Delta)\theta_a}\right)$. The stable corner equilibria are non-hyperbolic because of the boundary factors in the vector field.*

Throughout the baseline proofs, fix $a := \theta_a \in (0,1)$ and $\Delta, \kappa > 0$. Write

$$f(\theta,p) = \theta(1-\theta)[(1-p)(1-\theta) - \Delta p\theta], \qquad h(\theta,p) = \kappa p(1-p)D_a(\theta),$$
$$D_a(\theta) = (1-\theta)^2 - (1-a)^2.$$

Proof. To establish the equilibria and their stability properties, we first identify the fixed points of the system and then analyze the Jacobian matrix of the linearized system around these points.

Equilibrium locations. The stationary equations are

$$\theta(1-\theta)[(1-p)(1-\theta) - \Delta p\theta] = 0, \tag{19}$$
$$\kappa p(1-p)D_a(\theta) = 0. \tag{20}$$

From the factors $\theta(1-\theta)$ in (19) and $p(1-p)$ in (20), we immediately identify the four corner equilibria:

$$(\theta,p) \in \{(0,0),(0,1),(1,0),(1,1)\}.$$

To find the interior equilibrium $(\theta^\dagger, p^\dagger) \in (0,1)^2$, we require the non-trivial factors to vanish:

$$D_a(\theta) = 0 \iff \theta = a, \tag{21}$$
$$(1-p)(1-\theta) - \Delta p\theta = 0 \iff p = q(\theta) := \frac{1-\theta}{1-\theta+\Delta\theta}. \tag{22}$$

Since $\theta, a \in [0,1]$, Equation (21) gives the unique skill solution $\theta^\dagger = a$. Substituting this into (22) yields $p^\dagger = q(a)$. Write $e^\dagger = (a,q)$ with $q = q(a)$.

Stability analysis. Let J be the Jacobian matrix of the system:

$$J(\theta,p) = \begin{bmatrix} \partial_\theta f & \partial_p f \\ \partial_\theta h & \partial_p h \end{bmatrix}.$$

Corner equilibria. At the corner points where $(\theta,p) \in \{0,1\}^2$, the off-diagonal terms vanish (i.e., $\partial_p f = 0$ and $\partial_\theta h = 0$). The Jacobian becomes diagonal:

$$J = \mathrm{diag}(\lambda_\theta, \lambda_p),$$

where the eigenvalues are given by the partial derivatives evaluated at the equilibrium $(\theta^\infty, p^\infty)$:

$$\lambda_\theta = (1-2\theta^\infty)\left((1-p^\infty)(1-\theta^\infty) - \Delta p^\infty\theta^\infty\right),$$
$$\lambda_p = \kappa(1-2p^\infty)\left((1-\theta^\infty)^2 - (1-\theta_a)^2\right).$$

We examine two cases:

- **Sources ($\theta^\infty = p^\infty$):**
- At $(0,0)$: $\lambda_\theta = 1 > 0$ and $\lambda_p = \kappa(1-(1-\theta_a)^2) > 0$. Both eigenvalues are positive.
- At $(1,1)$: $\lambda_\theta = \Delta > 0$ and $\lambda_p = \kappa(1-\theta_a)^2 > 0$. Both eigenvalues are positive.

Thus, $(0,0)$ and $(1,1)$ are unstable nodes.

- **Nonhyperbolic attracting corners ($\theta^\infty \neq p^\infty$):**
- At $(1,0)$: $\lambda_\theta = 0$ and $\lambda_p = -\kappa(1-\theta_a)^2 < 0$. The negative λ_p implies exponential decay in the delegation direction (transverse stability). Along the invariant boundary $p = 0$, the skill dynamics reduce to $\dot\theta = \theta(1-\theta)^2$, which is positive for $\theta \in (0,1)$. This implies that θ monotonically increases toward 1.

• At $(0,1)$: $\lambda_\theta = 0$ and $\lambda_p = -\kappa[1-(1-\theta_a)^2] < 0$. The negative λ_p implies exponential decay toward full delegation. Along the invariant boundary $p = 1$, the skill dynamics reduce to $\dot{\theta} = -\Delta\theta^2(1-\theta)$, which is negative for $\theta \in (0,1)$. This implies that θ monotonically decays toward 0.

Local asymptotic stability relative to the square follows from the invariant-rectangle argument in Appendix A.4, specialized to $d = 0$; the time normalization does not change stability.

The interior saddle. At $e^\dagger$, write

$$J^\dagger = \begin{pmatrix} A & B \\ C & 0 \end{pmatrix}, \qquad \begin{aligned} A &= -a(1-a)(1-q+\Delta q) < 0, \\ B &= -a(1-a)(1-a+\Delta a) < 0, \\ C &= -2\kappa q(1-q)(1-a) < 0. \end{aligned} \tag{23}$$

Hence $\det J^\dagger = -BC < 0$, so the equilibrium has one negative and one positive eigenvalue. In particular,

$$\lambda_s = \frac{A - \sqrt{A^2 + 4BC}}{2} < 0, \qquad v_s = (1, m)^\top, \qquad m = \frac{\lambda_s - A}{B} > 0.$$

This proves Theorem 3.1 and supplies the tangent used in the global construction. □

B.3 A comparison principle for skill gap and reliance

For preparation, we give the following lemma that characterizes a monotone coupling between skill gap and delegation.

Lemma B.2 (Monotone coupling between skill gap and delegation). *Let $x = 1 - \theta$ denote the skill gap. For fixed $\kappa, \Delta > 0$, the system in variables (x, p) is cooperative:*

$$\dot{x} = x(1-x)\big[\Delta p(1-x) - (1-p)x\big], \qquad \dot{p} = \kappa p(1-p)\big[x^2 - (1-\theta_a)^2\big].$$

Consequently, if $x_0 \le \widetilde{x}_0$ and $p_0 \le \widetilde{p}_0$, then for all $t \ge 0$,

$$x(t; x_0, p_0) \le x(t; \widetilde{x}_0, \widetilde{p}_0), \qquad p(t; x_0, p_0) \le p(t; \widetilde{x}_0, \widetilde{p}_0).$$

Moreover, for fixed initial condition, increasing θ_a weakly increases $p(t)$ and $x(t)$ for all $t \ge 0$.

For distinct ordered interior initial states, both inequalities are strict for $t > 0$.

Proof. In the variables $x = 1 - \theta$ and p, write the vector field as

$$F_1(x,p) = x(1-x)\big[\Delta p(1-x) - (1-p)x\big],$$

$$F_2(x,p;\theta_a) = \kappa p(1-p)\big[x^2 - (1-\theta_a)^2\big].$$

On the interior $(0,1)^2$,

$$\frac{\partial F_1}{\partial p} = x(1-x)\big[\Delta(1-x) + x\big] \ge 0,$$

and

$$\frac{\partial F_2}{\partial x} = 2\kappa p(1-p)x \ge 0.$$

Thus the system is cooperative in (x, p). By Kamke's comparison theorem for monotone dynamical systems [34], larger initial skill gap x_0, larger initial delegation p_0, or larger AI skill θ_a weakly increase future skill gap and delegation (Lemma B.2).

For the parameter θ_a, observe that

$$\frac{\partial F_1}{\partial \theta_a} = 0, \qquad \frac{\partial F_2}{\partial \theta_a} = 2\kappa p(1-p)(1-\theta_a) \ge 0.$$

Therefore, increasing θ_a increases the vector field in the p coordinate and does not decrease it in the x coordinate. Applying the same comparison theorem to the parameterized systems gives the claimed monotonicity in AI skill.

Strict interior order follows because the variational system has strictly positive off-diagonal coefficients: any nonzero nonnegative difference immediately propagates to both coordinates. □

B.4 Proof of Theorem 3.2: the global separatrix

Theorem 3.2 (Basin partition by the stable manifold; restated). *Let θ_a, κ, Δ be the parameters of ODE (7).*

Let $(\theta^\dagger, p^\dagger)$ be the interior saddle point guaranteed by Theorem 3.1.

There exists a nondecreasing differentiable function $\psi : (0,1) \to (0,1)$ that extends continuously to $[0,1]$ with $\psi(0) = 0$ and $\psi(1) = 1$, whose graph coincides with the one-dimensional stable manifold of $(\theta^\dagger, p^\dagger)$ within $(0,1)^2$.

The curve $p = \psi(\theta)$ separates the two basins of attraction:

- *For any initial state with $p_0 > \psi(\theta_0)$ (higher delegation), the learner converges to the low-skill equilibrium $(0,1)$.*
- *For any initial state with $p_0 < \psi(\theta_0)$ (lower delegation), the learner converges to the high-skill equilibrium $(1,0)$.*

Proof.

Local stable manifold. From Theorem 3.1, the interior equilibrium $(\theta^\dagger, p^\dagger)$ is a hyperbolic saddle point. The stable-manifold theorem [28] gives a unique smooth local curve $W^s_{\mathrm{loc}}(e^\dagger)$ through the saddle, tangent to the stable eigenvector $(1, m)$, where $m > 0$ is given above. Let

$$W^s = \bigcup_{t \geq 0} \varphi_{-t}(W^s_{\mathrm{loc}}) = \{z \in (0,1)^2 : \varphi_t(z) \to e^\dagger\}.$$

Backward continuation is well defined by invariance and uniqueness. Locally W^s is increasing in (θ, p) and decreasing in $(x, p) = (1 - \theta, p)$.

No folding and smooth graph representation. Suppose two distinct points $z, z' \in W^s$ were weakly ordered in (x, p). Lemma B.2 gives

$$z \leq z',\ z \neq z' \quad \Longrightarrow \quad \varphi_t(z) \ll \varphi_t(z') \quad (t > 0),$$

where $\ll$ denotes strict coordinatewise order. For large t, both points lie on W^s_{loc}, a decreasing graph in (x, p) containing no ordered pair. This is a contradiction. Consequently,

$$(\theta, p), (\theta', p') \in W^s,\ \theta < \theta' \quad \Longrightarrow \quad p < p',$$

and no two distinct points have the same θ. Connectedness of W^s therefore gives a continuous strictly increasing graph $p = \psi(\theta)$ on an interval I containing a.

Since $q'(\theta) = -\Delta/(1 - \theta + \Delta\theta)^2 < 0$, the increasing graph ψ and the decreasing skill nullcline q intersect only at $e^\dagger$. Thus

$$f(\theta, \psi(\theta)) \neq 0 \quad (\theta \neq a), \qquad \psi'(\theta) = \frac{h(\theta, \psi(\theta))}{f(\theta, \psi(\theta))}.$$

The flow is tangent to W^s, so $f \neq 0$ excludes vertical tangencies away from the saddle. The stable eigenvector has nonzero θ component at the saddle. Hence ψ is C^1 on I.

Boundary endpoints. The two branches of $W^s(\theta^\dagger, p^\dagger)$ have alpha-limit sets as time goes to $-\infty$. On the lower stable branch, $\theta < a$ and $p < q(a) < q(\theta)$, so $(f, h) > (0, 0)$ and both coordinates decrease in backward time. On the upper branch, $\theta > a$ and $p > q(a) > q(\theta)$, so both coordinates increase in backward time. Each branch is bounded and monotone; its backward limit is an equilibrium. The coordinate bounds leave only

$$\lim_{t \to -\infty} \varphi_t(z_-) = (0,0), \qquad \lim_{t \to -\infty} \varphi_t(z_+) = (1,1).$$

Therefore, the graph connects $(0,0)$ to $(1,1)$, which gives

$$\lim_{\theta \downarrow 0} \psi(\theta) = 0, \qquad \lim_{\theta \uparrow 1} \psi(\theta) = 1.$$

Consequently $I = (0,1)$, and these endpoint values define the continuous extension. The corners belong to $\overline{W^s}$, not to W^s itself.

Identification of the two basins. Take $p_0 > \psi(\theta_0)$ and compare with the reference trajectory from $(\theta_0, \psi(\theta_0)) \in W^s$. The comparison lemma gives

$$\theta(t; \theta_0, p_0) \le \theta(t; \theta_0, \psi(\theta_0)), \qquad p(t; \theta_0, p_0) \ge p(t; \theta_0, \psi(\theta_0)).$$

Every trajectory converges by Theorem A.3; the reference limit is (a, q). Hence the actual limit has $\theta_\infty \le a$ and $p_\infty \ge q$. Among the equilibria, only $(0, 1)$ and (a, q) satisfy these bounds. The latter is impossible because $z_0 \notin W^s$. The reversed comparison applies below the graph. We have proved

$$B_{\text{low}} = \{(\theta, p) \in (0, 1)^2 : p > \psi(\theta)\}, \qquad B_{\text{high}} = \{(\theta, p) \in (0, 1)^2 : p < \psi(\theta)\}. \tag{24}$$

Since ψ is continuous, (24) directly identifies the common boundary:

$$\partial B_{\text{low}} \cap (0, 1)^2 = \partial B_{\text{high}} \cap (0, 1)^2 = W^s(\theta^\dagger, p^\dagger) \cap (0, 1)^2.$$

Every neighborhood of a graph point meets both basins, while each off-graph point is interior to one basin. Thus W^s is their separatrix. □

B.5 Proof of Theorem 3.3: AI capability and basin ordering

Theorem 3.3 (AI skill monotonically sweeps the basin boundary; restated). *For every $\theta \in (0, 1)$, the boundary value $\psi_{\theta_a}(\theta)$ is nonincreasing and continuously differentiable in θ_a.*

Moreover, the family $\{\psi_{\theta_a}\}_{\theta_a \in (0,1)}$ sweeps the state space in the following sense: for every $(\theta_0, p_0) \in (0, 1)^2$, there exists $\theta_a \in (0, 1)$ such that $p_0 = \psi_{\theta_a}(\theta_0)$.

Write ψ_a for the separatrix at AI capability a, with Δ and κ fixed.

Proof. We prove the properties of the separatrix ψ_{θ_a} sequentially.

Ordering. Let $\theta_{a,1} < \theta_{a,2}$ be two distinct AI skill levels, and let ψ_1 and ψ_2 denote the corresponding separatrices. Suppose, for the sake of contradiction, that there exists some $\theta^* \in (0, 1)$ such that $\psi_2(\theta^*) > \psi_1(\theta^*)$. We can select a test point (θ^*, p_0) such that

$$\psi_1(\theta^*) < p_0 < \psi_2(\theta^*).$$

Since $p_0 > \psi_1(\theta^*)$, the initial state lies in the low-skill basin under $\theta_{a,1}$, and Theorem 3.2 gives $\theta_1^\infty = 0$. Since $p_0 < \psi_2(\theta^*)$, it lies in the high-skill basin under $\theta_{a,2}$, giving $\theta_2^\infty = 1$. However, Lemma B.2 states that for any time t, the skill $\theta(t)$ is nonincreasing in θ_a. Taking the limit $t \to \infty$ implies $\theta_2^\infty \le \theta_1^\infty$, a contradiction. Therefore $\psi_2(\theta) \le \psi_1(\theta)$ for all $\theta \in (0, 1)$.

Smooth parameter dependence. The vector field and $e^\dagger(a)$ are smooth, and the saddle is hyperbolic for every $a \in (0, 1)$. Parameter-dependent stable manifolds [28] and the finite-time flow therefore give smooth parametrizations of every local piece of W_a^s. Its tangent has nonzero θ component by the preceding proof. The implicit function theorem then gives C^1 dependence of its intersection with each fixed section $\theta = \theta_0$, including a section through the saddle.

Sweeping property and endpoint limits. We show that for any point $(\theta_0, p_0) \in (0, 1)^2$, there exists a θ_a such that the separatrix passes through it. To prove this, it suffices to establish that, for every fixed $\theta_0 \in (0, 1)$,

$$\lim_{\theta_a \uparrow 1} \psi_{\theta_a}(\theta_0) = 0, \qquad \lim_{\theta_a \downarrow 0} \psi_{\theta_a}(\theta_0) = 1. \tag{25}$$

We prove the two limits by comparison with the limiting systems $\theta_a = 1$ and $\theta_a = 0$.

Fix $0 < r < 1$ and choose $0 < P_- < q(r) < P_+ < 1$. The rectangles

$$R_H = [r, 1] \times [0, P_-], \qquad R_L = [0, r] \times [P_+, 1]$$

are forward invariant and attracted to $(1, 0)$ whenever $a < r$, and to $(0, 1)$ whenever $a > r$, respectively. These parameter-independent traps justify the finite-time parameter-continuity argument below.

First consider $\theta_a = 1$. The delegation equation becomes

$$\dot{p} = \kappa p(1 - p)(1 - \theta)^2 \ge 0.$$

For every interior initial condition with $p_0 > 0$, delegation is nondecreasing. Let $s(P) = (1 - P)/(1 - P + \Delta P)$ denote the fixed-delegation skill equilibrium. Since $p(t) \geq p_0$,

$$\theta(t) \leq M := \max\{\theta_0, s(p_0)\} < 1, \qquad \frac{d}{dt} \operatorname{logit} p \geq \kappa(1 - M)^2 > 0.$$

Hence $p \to 1$. For any $P < 1$, eventually $p \geq P$ and comparison with fixed delegation gives $\limsup \theta(t) \leq s(P)$. Letting $P \uparrow 1$ yields $\theta \to 0$.

Fix $\theta_0 \in (0, 1)$ and any $\varepsilon \in (0, 1)$. Under the limiting system $\theta_a = 1$, the trajectory starting from (θ_0, ε) converges to $(0, 1)$. Hence there exists a finite time T at which this trajectory enters the relative interior of R_L. By continuous dependence of ODE solutions on parameters, for all $\theta_a < 1$ sufficiently close to 1, the trajectory starting from (θ_0, ε) also enters R_L, which remains forward invariant and attracted to $(0, 1)$. Therefore (θ_0, ε) lies in the low-skill basin for all sufficiently large θ_a. Since $\varepsilon > 0$ was arbitrary, the basin threshold satisfies

$$\limsup_{\theta_a \uparrow 1} \psi_{\theta_a}(\theta_0) \leq \varepsilon.$$

Letting $\varepsilon \downarrow 0$ gives

$$\lim_{\theta_a \uparrow 1} \psi_{\theta_a}(\theta_0) = 0.$$

Next consider $\theta_a = 0$. The delegation equation becomes

$$\dot{p} = \kappa p(1 - p)\big((1 - \theta)^2 - 1\big) = -\kappa p(1 - p)\theta(2 - \theta) \leq 0.$$

For every interior initial condition with $p_0 < 1$, delegation is nonincreasing. Since $p(t) \leq p_0 < 1$,

$$\theta(t) \geq m := \min\{\theta_0, s(p_0)\} > 0, \qquad \frac{d}{dt} \operatorname{logit} p = -\kappa\theta(2 - \theta) \leq -\kappa m(2 - m) < 0.$$

Thus $p \to 0$. Comparison with fixed delegation at arbitrarily small $P > 0$ gives $\liminf \theta(t) \geq s(P) \to 1$. The limiting system therefore converges to the high-skill low-delegation regime $(1, 0)$. Fix $\theta_0 \in (0, 1)$ and any $\varepsilon \in (0, 1)$. Under the limiting system $\theta_a = 0$, the trajectory starting from $(\theta_0, 1 - \varepsilon)$ converges to $(1, 0)$. Hence it enters the relative interior of R_H in finite time. Again by continuous dependence on parameters, for all $\theta_a > 0$ sufficiently close to 0, the trajectory starting from $(\theta_0, 1 - \varepsilon)$ also enters R_H and remains in the high-skill basin. Therefore

$$\psi_{\theta_a}(\theta_0) \geq 1 - \varepsilon$$

for all sufficiently small θ_a. Letting $\varepsilon \downarrow 0$ gives

$$\lim_{\theta_a \downarrow 0} \psi_{\theta_a}(\theta_0) = 1.$$

Fix $(\theta_0, p_0) \in (0, 1)^2$. By the endpoint limits,

$$\lim_{\theta_a \downarrow 0} \psi_{\theta_a}(\theta_0) = 1 > p_0, \qquad \lim_{\theta_a \uparrow 1} \psi_{\theta_a}(\theta_0) = 0 < p_0.$$

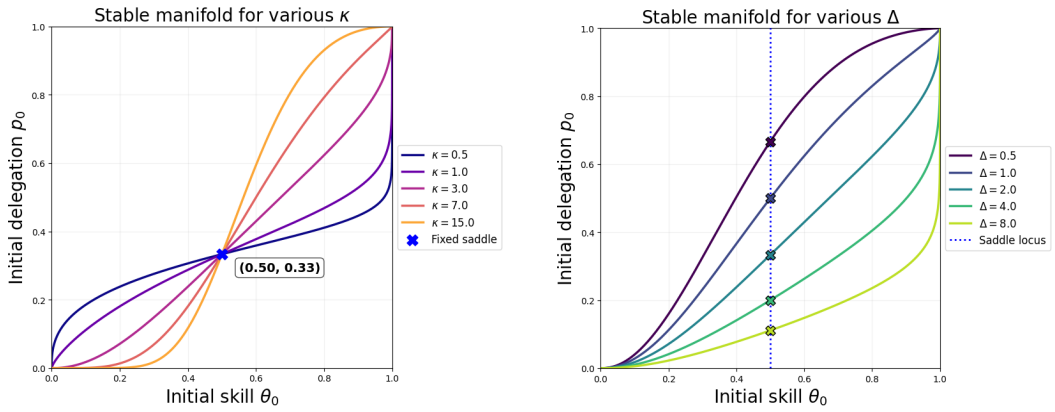
Since $\psi_{\theta_a}(\theta_0)$ is continuous in θ_a, the intermediate value theorem implies that there exists $\theta_a \in (0, 1)$ such that

$$\psi_{\theta_a}(\theta_0) = p_0.$$

This proves the sweeping property. □

B.6 Proof of Theorem 3.5: performance reversal

Theorem 3.5 (Short-term gains, long-term losses; restated). *Let θ_a, κ, Δ be the parameters of ODE (7). Fix an initial state $(\theta_0, p_0) \in (0, 1)^2$ with $\theta_0 < \theta_a < 1$. Let $t_c := \inf\{t \geq 0 : \forall t' > t,\ G_\ell(t') > 0\}$ denote the final crossing time, after which adaptive delegation incurs higher instantaneous task loss than the no-AI baseline from the same initial skill. Then $t_c < \infty$ and $\theta(t_c; \theta_0, 0) \leq \theta_a$.*



(a) Effects of κ (b) Effects of Δ

Figure 4: Plots illustrating the relationship between the basin boundary $\psi(\cdot)$ and model parameters θ_a, κ, Δ, with default setting $(\theta_a, \kappa, \Delta) = (0.5, 3, 2)$.

Proof. We first express the performance gap $G_\ell(t)$ by expanding the loss term $\ell(\theta, p) = (1-p)(1-\theta)^2 + p(1-\theta_a)^2$:

$$\begin{aligned} G_\ell(t) &= \left[(1-p(t;\theta_0,p_0))(1-\theta(t;\theta_0,p_0))^2 + p(t;\theta_0,p_0)(1-\theta_a)^2\right] - (1-\theta(t;\theta_0,0))^2 \\ &= (1-p(t;\theta_0,p_0))\left[(1-\theta(t;\theta_0,p_0))^2 - (1-\theta(t;\theta_0,0))^2\right] \\ &\quad + p(t;\theta_0,p_0)\left[(1-\theta_a)^2 - (1-\theta(t;\theta_0,0))^2\right]. \end{aligned}$$

Since $p_0 > 0$, Lemma B.2 and interior invariance imply that for all $t > 0$, the delegation level remains positive, $p(t;\theta_0,p_0) > 0$, and the AI-assisted skill is no more than the baseline skill:

$$\theta(t;\theta_0,p_0) \le \theta(t;\theta_0,0).$$

Consequently, the first bracketed term (skill loss difference) is non-negative for all $t > 0$:

$$(1-\theta(t;\theta_0,p_0))^2 - (1-\theta(t;\theta_0,0))^2 \ge 0. \tag{26}$$

The baseline skill $\theta(t;\theta_0,0)$ evolves according to $\dot\theta = \theta(1-\theta)^2$, which strictly increases from θ_0 to 1. Since $\theta_0 < \theta_a < 1$, by the Intermediate Value Theorem, there exists a unique finite time $t^* > 0$ such that $\theta(t^*;\theta_0,0) = \theta_a$. For any time $t > t^*$, we have $\theta(t;\theta_0,0) > \theta_a$, which implies:

$$(1-\theta_a)^2 - (1-\theta(t;\theta_0,0))^2 > 0.$$

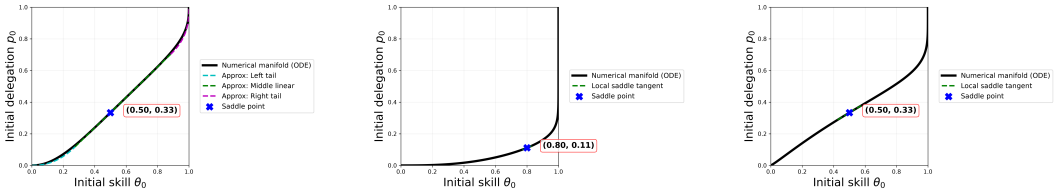
Combining this with (26) and the fact that $p(t;\theta_0,p_0) > 0$, we conclude that $G_\ell(t) > 0$ for all $t > t^*$, because its first term is nonnegative and its second term is strictly positive. The crossing time is defined as $t_c := \inf\{t \ge 0 : \forall t' > t, G_\ell(t') > 0\}$. Since $G_\ell(t) > 0$ holds for all $t > t^*$, it follows that $t_c \le t^* < \infty$. Since the baseline skill increases, $t_c \le t^*$ implies $\theta(t_c;\theta_0,0) \le \theta(t^*;\theta_0,0) = \theta_a$. Initially $G_\ell(0) = p_0[(1-\theta_a)^2 - (1-\theta_0)^2] < 0$, so the initial advantage reverses. The argument does not depend on the attracting basin. It also holds for fixed delegation $p(t) \equiv \bar p \in (0,1]$: scalar comparison again gives skill no greater than the no-AI baseline, and the same loss decomposition is positive after t^*. Thus the reversal does not require adaptive reliance. □

B.7 Effects of the remaining parameters on the separatrix

Beyond Theorem 3.3, we also study how κ and Δ affect the stable manifold. Figure 4 shows the resulting variation in the basin boundary as κ and Δ change. We observe that changing κ does not affect the saddle location, whereas changing Δ shifts the saddle along the line $\theta = \theta_a$.

Lemma B.3 (Dependence on reliance-adjustment speed). *Fix $a \in (0,1)$ and $\Delta > 0$. The separatrix is C^1 in $\kappa > 0$. For $\kappa_2 > \kappa_1$,*

$$\psi_{\kappa_2}(\theta) < \psi_{\kappa_1}(\theta)\ (\theta < a), \qquad \psi_{\kappa_2}(a) = \psi_{\kappa_1}(a) = q(a), \qquad \psi_{\kappa_2}(\theta) > \psi_{\kappa_1}(\theta)\ (\theta > a).$$



(a) Default setting. (b) Varying $\theta_a = 0.8$. (c) Varying $\kappa = 1.5$.

Figure 5: Numerical separatrices and their local approximations, with default parameters $(\theta_a, \kappa, \Delta) = (0.5, 3, 2)$. Panel (a) shows the three-piece smoothly pasted approximation. Panels (b) and (c) vary only the indicated parameter and show the local saddle tangent instead: their left matching points are inadmissible, so no complete three-piece approximation is drawn.

Proof. Smooth dependence follows as in Theorem 3.3. Write the saddle matrix as $J_\kappa = \begin{pmatrix} A & B \\ \kappa C_0 & 0 \end{pmatrix}$ with $A, B, C_0 < 0$. Its stable slope satisfies

$$m_\kappa = \frac{-A - \sqrt{A^2 + 4B\kappa C_0}}{2B}, \qquad \frac{dm_\kappa}{d\kappa} = -\frac{C_0}{\sqrt{A^2 + 4B\kappa C_0}} > 0.$$

The stated ordering therefore holds near the common saddle. It cannot reverse away from it. At any contact of the two stable graphs, their common point lies below the skill nullcline when $\theta < a$, or above it when $\theta > a$; in either case

$$\psi'_{\kappa_2} - \psi'_{\kappa_1} = (\kappa_2 - \kappa_1)\frac{p(1-p)D_a(\theta)}{f(\theta, p)} > 0.$$

On the left, a first contact approached from the saddle would require this difference to be nonpositive; on the right the same is true at a first contact reached while moving rightwards. Both are impossible. □

Lemma B.4 (Dependence on non-use decay). *Fix $a \in (0,1)$ and $\kappa > 0$. The separatrix is C^1 and strictly decreasing in $\Delta > 0$ at each $\theta \in (0,1)$.*

Proof. In skill-gap coordinates, $\partial_\Delta \dot{x} = x(1-x)^2 p \geq 0$ and $\partial_\Delta \dot{p} = 0$. Comparison therefore gives $\psi_{\Delta_2} \leq \psi_{\Delta_1}$ for $\Delta_2 > \Delta_1$. At $\theta = a$, the saddle height $(1-a)/(1-a+\Delta a)$ decreases strictly.

At any other hypothetical contact, both graphs are differentiable and ordered, so their derivatives must agree. But their common delegation drift is $h \neq 0$ and $f_2 < f_1$, giving

$$\psi'_{\Delta_2} - \psi'_{\Delta_1} = h\left(\frac{1}{f_2} - \frac{1}{f_1}\right) = \frac{h(f_1 - f_2)}{f_1 f_2} \neq 0.$$

This contradiction proves strictness. Smooth parameter dependence follows from the same local-manifold and transverse-section argument as before. □

B.8 A local-to-global approximation of the separatrix

In this section, we approximate the stable manifold ψ by the explicit function $\widetilde{\psi}$ in Eq. (27). We construct $\widetilde{\psi}(\theta)$ by approximating the global stable manifold as a composite of three locally valid functions: a linear segment near the saddle point derived from the stable eigenvector, and two power-law segments near the corners $(0,0)$ and $(1,1)$ derived from the local eigenvalues. We determine the transition points θ_l and θ_r by enforcing continuity of both the function and its first derivative (differentiable continuity). Figure 5 illustrates when this construction is admissible.

Local ingredients. Let (a, q) be the saddle and let A, B, C denote the entries of (23). The characteristic equation is $\lambda^2 - A\lambda - BC = 0$, and the stable eigenvalue λ_s is its negative root. The associated eigenvector $(1, m)^\top$ satisfies $(A - \lambda_s) + Bm = 0$, giving the stable tangent

$$L(\theta) = q + m(\theta - a), \qquad m = \frac{-A - \sqrt{A^2 + 4BC}}{2B} > 0.$$

Near the origin $(0,0)$, keeping the lowest-order linear terms gives

$$\dot{\theta} \simeq \theta, \qquad \dot{p} \simeq \kappa[1-(1-a)^2]p.$$

The ratio of the coordinatewise growth rates is $\beta_l = \kappa[1-(1-a)^2]$. The trajectory shape therefore satisfies $dp/d\theta \simeq \beta_l p/\theta$. Integrating gives the power-law segment

$$\psi_{\text{left}}(\theta) = C_l \theta^{\beta_l}.$$

Near $(1,1)$, set $x = 1-\theta$ and $y = 1-p$. The linearized dynamics are

$$\dot{x} \simeq \Delta x, \qquad \dot{y} \simeq \kappa(1-a)^2 y.$$

Their coordinatewise growth-rate ratio is $\beta_r = \kappa(1-a)^2/\Delta$. Thus $dy/dx \simeq \beta_r y/x$, which integrates to $y = C_r x^{\beta_r}$. Transforming back gives

$$\psi_{\text{right}}(\theta) = 1 - C_r(1-\theta)^{\beta_r}.$$

Matching values and slopes. We determine the candidate junctions by matching the values and derivatives of the power-law segments to the saddle tangent. At the left junction, value matching gives $C_l \theta_l^{\beta_l} = L(\theta_l)$, so derivative matching becomes $C_l \beta_l \theta_l^{\beta_l - 1} = \beta_l L(\theta_l)/\theta_l = m$. Similarly, at the right junction, $C_r(1-\theta_r)^{\beta_r} = 1 - L(\theta_r)$. Thus C^1 matching requires

$$\frac{\beta_l L(\theta_l)}{\theta_l} = m, \qquad \frac{\beta_r[1-L(\theta_r)]}{1-\theta_r} = m.$$

Substituting $L(\theta) = q + m(\theta - a)$ and rearranging gives

$$\beta_l(q - ma) = m\theta_l(1-\beta_l), \qquad m\theta_r(1-\beta_r) = m - \beta_r(1-q+ma).$$

When $\beta_l, \beta_r \neq 1$, solving gives

$$\theta_l = \frac{\beta_l(q-ma)}{m(1-\beta_l)}, \qquad \theta_r = \frac{m - \beta_r(1-q+ma)}{m(1-\beta_r)}.$$

The construction is admissible only if

$$0 < \theta_l \leq a \leq \theta_r < 1, \qquad 0 < L(\theta_l) \leq L(\theta_r) < 1.$$

For such junctions define

$$\widetilde{\psi}(\theta) = \begin{cases} L(\theta_l)(\theta/\theta_l)^{\beta_l}, & 0 \leq \theta \leq \theta_l, \\ L(\theta), & \theta_l < \theta < \theta_r, \\ 1 - [1-L(\theta_r)]((1-\theta)/(1-\theta_r))^{\beta_r}, & \theta_r \leq \theta \leq 1. \end{cases} \tag{27}$$

It is continuous on $[0,1]$, increasing, and C^1 in the interior, including at the two junctions. No uniform approximation-error bound is established. If a denominator vanishes or the junctions are inadmissible, use numerical continuation rather than clipping the junctions, which generally breaks derivative matching.

Numerical continuation. A graph-based computation starts at $(a \pm \varepsilon, q \pm m\varepsilon)$ and integrates

$$\frac{dp}{d\theta} = \frac{h(\theta,p)}{f(\theta,p)} \qquad (\theta \neq a).$$

The branch meets no other skill nullcline. Alternatively, integrate the full ODE backward from the stable eigendirection. Check convergence under smaller initial displacement, solver tolerance, and endpoint truncation.

C Reliance adaptation and interventions

The basin characterization gives a geometric way to reason about interventions. A sustained intervention can shift the separatrix $p = \psi(\theta)$ so that more initial states converge to high skill. Changing the response to performance feedback preserves the skill level at which reliance switches direction; changing incentives can move that threshold. A temporary cost must move the learner across the original basin boundary before removal.

C.1 A global result for centered monotone responses

Replace proportional responsiveness to the performance gap by

$$\dot{\theta} = f(\theta, p), \qquad \dot{p} = \kappa p(1-p)\sigma(D_a(\theta)), \qquad D_a(\theta) = (1-\theta)^2 - (1-a)^2, \tag{28}$$

where f is the normalized baseline skill drift, $a = \theta_a \in (0,1)$, and $\sigma \in C^1(\mathbb{R})$ is strictly increasing with $\sigma(0) = 0$ and $\sigma'(0) > 0$. The zero value at equal performance preserves indifference; an uncentered logistic function does not satisfy this condition. Strict increase implies $\sigma' \geq 0$, not necessarily $\sigma' > 0$ everywhere.

Proposition C.1 (Basin geometry under a monotone response). *For $\Delta, \kappa > 0$, system (28) retains the baseline equilibria and their stability types. The interior saddle is $e^\dagger = (a, q(a))$, where $q(\theta) = (1-\theta)/(1-\theta+\Delta\theta)$. Its stable manifold is a strictly increasing C^1 graph $p = \psi_a^\sigma(\theta)$ with endpoint limits 0 and 1. Interior states below, above, and on the graph converge to $(1,0)$, $(0,1)$, and $e^\dagger$, respectively. For fixed σ, Δ, κ, $\psi_a^\sigma(\theta)$ is nonincreasing in a.*

Proof.

Equilibria and local stability. Since $\operatorname{sign}\sigma(D_a) = \operatorname{sign}(a-\theta)$, the nullclines and four-sector sign pattern (13) are unchanged. At the interior equilibrium,

$$J_\sigma^\dagger = \begin{pmatrix} A & B \\ C\sigma'(0) & 0 \end{pmatrix}, \qquad A, B, C < 0 \text{ as in (23)}, \qquad \det J_\sigma^\dagger = -BC\sigma'(0) < 0.$$

The local stable eigenvector has positive slope. The source corners retain positive eigenvalues. The same invariant rectangles used in the baseline establish local attraction of the other two corners.

Global convergence and continuation. Let $v = p - q(\theta)$. On $v = 0$, $\dot{v} = \dot{p}$; on $\theta = a$, $\operatorname{sign}\dot{\theta} = \operatorname{sign}(q(a) - p)$. Hence

$$\mathcal{H} = \{\theta > a, p < q(\theta)\}, \qquad \mathcal{L} = \{\theta < a, p > q(\theta)\}$$

are invariant and converge to $(1,0)$ and $(0,1)$. The two remaining sectors are coordinatewise monotone until entering these sets; if a trajectory never enters, its only possible limit is the saddle. Thus every interior trajectory converges.

For the stable manifold, the rectangles

$$R_- = (0,a) \times (0, q(a)), \qquad R_+ = (a,1) \times (q(a), 1)$$

are backward invariant. On R_- both forward drifts are positive, and on R_+ both are negative. The local stable branches lie in these rectangles. Along backward continuation, both coordinates are monotone, and the branches approach $(0,0)$ and $(1,1)$, respectively. Along either branch, $f \neq 0$ gives $dp/d\theta = \dot{p}/f > 0$; the stable eigendirection gives the C^1 graph at the saddle. The graph spans $(0,1)$, separating two connected regions. Every off-graph trajectory approaches a corner attractor. Since their basins are open, the trapping rectangles identify the lower region as B_{high} and the upper as B_{low}.

AI-capability comparison. In $(x,p) = (1-\theta, p)$ coordinates,

$$\partial_p \dot{x} = x(1-x)[\Delta(1-x) + x] > 0, \qquad \partial_x \dot{p} = 2\kappa p(1-p)x\sigma'(D_a) \geq 0,$$
$$\partial_a \dot{p} = 2\kappa p(1-p)(1-a)\sigma'(D_a) \geq 0.$$

The comparison principle orders both coordinates upward when a increases. Therefore $B_{\text{low}}(a_1) \subseteq B_{\text{low}}(a_2)$ for $a_1 < a_2$, equivalently $\psi_{a_2}^\sigma \leq \psi_{a_1}^\sigma$. □

Positive state-dependent adjustment rates. The same conclusions hold after multiplying $\dot{p}$ by a positive C^1 function $m(p)$ on $[0,1]$, independent of a. The sign chart and backward-invariant rectangles are unchanged, the saddle's lower-left Jacobian entry is multiplied by $m(q) > 0$, and the cooperative cross-effects are retained. Changing only reliance speed can move the separatrix; it does not simply rescale time in both equations.

b	$a = 0.3$	$a = 0.5$	$a = 0.7$
0.5	0.3385	0.6448	0.8720
2	0.3425	0.6450	0.8711
8	0.3704	0.6495	0.8626

Table 2: Low-skill-basin area under normalized sigmoid responses, with $\Delta = 2$ and $\kappa = 3$. Areas are fractions of the unit square, not empirical frequencies.

Normalized sigmoid illustration. We use

$$\sigma_b(u) = \frac{\tanh(bu)}{b} = \frac{2}{b}\left(\frac{1}{1+e^{-2bu}} - \frac{1}{2}\right) \quad (b > 0), \qquad \sigma_0(u) = u, \qquad \sigma_b'(0) = 1.$$

Thus b changes saturation while leaving the saddle linearization fixed. For $\Delta = 2$, $\kappa = 3$, the low-skill-basin fraction is $A_{\text{low}}(a, b) = \int_0^1 [1 - \psi_a^{\sigma_b}(\theta)]\, d\theta$.

Reported RK4 refinement changes all areas by less than 2×10^{-7}; an independent midpoint-grid check differs by less than 1.3×10^{-4} in basin share. These numerical checks illustrate, rather than prove, the global result.

C.2 Conservative reliance updates

To make reliance more conservative after evidence against AI, change only responsiveness, leaving the performance loss and potentials Φ_s, Φ_d unchanged. Feedback-updated beliefs motivate this possibility [6], but do not establish the particular asymmetric rule below. Let

$$D(\theta) := (1-\theta)^2 - (1-\theta_a)^2$$

denote the baseline performance gap driving delegation. Instead of using the symmetric signal $D(\theta)$, we use

$$D_\alpha(\theta) = [D(\theta)]_+ - \alpha[-D(\theta)]_+,$$

where $\alpha > 0$ controls the relative sensitivity to negative evidence about AI. When $\alpha > 1$, the learner reduces reliance faster whenever independent work outperforms AI.

Writing $a = \theta_a$ and $D_a = D$, the dynamics become

$$\dot\theta = f(\theta, p), \qquad \dot p = \kappa p(1-p) D_\alpha(\theta) = \begin{cases} \kappa p(1-p) D_a(\theta), & \theta \le a, \\ \alpha\kappa p(1-p) D_a(\theta), & \theta \ge a. \end{cases} \tag{29}$$

Here $[u]_+ = \max\{u, 0\}$. When $\alpha = 1$, this reduces to the baseline ODE (7).

The nonsmooth threshold. Unlike the smooth response above, D_α need not be differentiable at $\theta = a$. Local Lipschitz continuity still gives unique solutions, and the same sign argument proves convergence. Let ψ_κ be the baseline separatrix with a, Δ fixed. Its lower branch stays in $\theta < a$ and its upper branch in $\theta > a$, so the exact boundary for (29) is

$$\psi_\alpha(\theta) = \begin{cases} \psi_\kappa(\theta), & \theta \le a, \\ \psi_{\alpha\kappa}(\theta), & \theta \ge a. \end{cases}$$

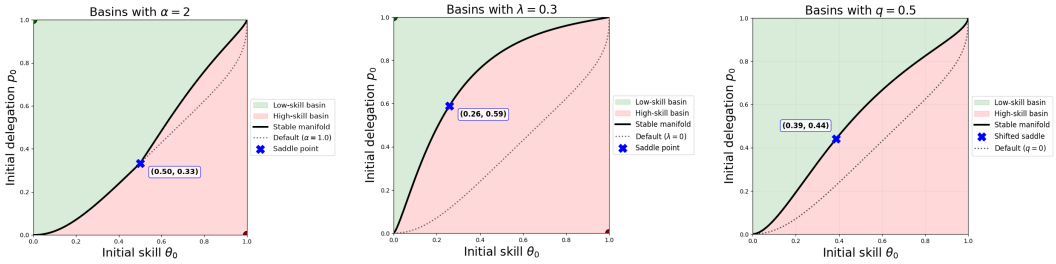
The pieces meet at $(a, q(a))$, approach the saddle forward in time, and approach the two source corners backward in time. They separate the same two attracting regimes. The curve is increasing and piecewise smooth, but its slopes at the saddle differ when $\alpha \neq 1$. The smooth stable-manifold theorem does not apply across this kink.

Lemma B.3 now gives $\psi_\alpha = \psi_1$ on $(0, a]$ and $\psi_\alpha > \psi_1$ on $(a, 1)$ for $\alpha > 1$. Thus the intervention expands the high-skill basin without changing the threshold for learners initially below AI skill. Figure 6a illustrates this learner-side intervention at $\alpha = 2$, without restricting AI access.

C.3 Adjustment friction and external constraints

To model resistance to changing reliance, rather than a change of objective, retain Φ_s, Φ_d and impose the implicit adjustment rule

$$\dot p = \kappa p(1-p) D_a(\theta) - \eta p(1-p)\dot p, \qquad \eta \ge 0.$$



(a) Conservative reliance updates (29) (b) Delegation costs (31) (c) Delegation penalties (33)

Figure 6: Plots illustrating how the basins of attraction change under learner- and teacher-side interventions, with default parameter settings $(\theta_a, \kappa, \Delta) = (0.5, 3, 2)$. The panels use $\alpha = 2$, cost $c = 0.3$ (denoted by λ in the image), and detection probability $q = 0.5$, respectively.

Here η is a friction coefficient, distinct from the discrete step size used in Appendix A.1. Solving gives

$$\dot{\theta} = f(\theta, p), \qquad \dot{p} = \frac{\kappa p(1-p)}{1 + \eta p(1-p)} D_a(\theta). \tag{30}$$

This slows adjustment without adding a velocity state; it is not momentum. For finite η, the multiplier $m(p) = [1 + \eta p(1-p)]^{-1} > 0$ meets the positive-multiplier conditions in Section C.1. Thus the corner stability, interior saddle, and global graph-valued basin partition persist. The saddle stays fixed, but slowing only reliance can move its stable manifold. We do not claim a uniform direction for the basin shift with η.

A hard cap $p \le \bar{p} < 1$ instead removes $(0, 1)$ and requires a separate boundary law; the friction result does not analyze that constrained model or imply convergence to full skill.

C.4 Delegation costs and learner-specific thresholds

To encourage independent practice, introduce a perceived cost $c \ge 0$ per delegated task, in loss units. The learner still adapts to current incentives. Keeping the skill-update potential Φ_s unchanged, add this cost to the delegation-update potential:

$$\Phi_d^c = (1-p)(1-\theta)^2 + p[(1-a)^2 + c].$$

Applying the same multiplicative-update geometry gives

$$\dot{\theta} = f(\theta, p), \qquad \dot{p} = \kappa p(1-p)[D_a(\theta) - c]. \tag{31}$$

When $c = 0$, this reduces to ODE (7).

Equivalently, the intervention makes AI behave as if it had a higher effective loss. For $0 \le c < c_{\max} := 1 - (1-a)^2$, define the effective capability by

$$\widehat{a}(c) = 1 - \sqrt{(1-a)^2 + c} \in (0, a], \qquad (1 - \widehat{a}(c))^2 = (1-a)^2 + c.$$

Thus (31) is exactly the baseline system at $\widehat{a}(c) \le a$. By Theorem 3.3, increasing c raises the separatrix and enlarges the high-skill basin. Figure 6b illustrates this effect at $c = 0.3$, denoted by λ in the image.

Proposition C.2 (Critical sustained cost). *Fix $z_0 = (\theta_0, p_0) \in (0,1)^2$ and $a \in (0,1)$, $\Delta, \kappa > 0$. Define*

$$\Psi_{\theta_0}(c) = \psi_{\widehat{a}(c)}(\theta_0), \qquad c_{\text{crit}} = \inf\{c \in [0, c_{\max}) : p_0 < \Psi_{\theta_0}(c)\}. \tag{32}$$

Then $0 \le c_{\text{crit}} < c_{\max}$. Every $c \in (c_{\text{crit}}, c_{\max})$ places z_0 in the high-skill basin; no $c < c_{\text{crit}}$ does. If $c_{\text{crit}} > 0$, the state lies on the modified separatrix at equality. If it already lies in the unmodified high-skill basin, $c_{\text{crit}} = 0$ and no intervention is needed.

Proof. Theorem 3.3 makes Ψ_{θ_0} continuous and nondecreasing. Moreover,

$$\lim_{c \uparrow c_{\max}} \Psi_{\theta_0}(c) = \lim_{a' \downarrow 0} \psi_{a'}(\theta_0) = 1 > p_0.$$

The costs satisfying the strict inequality therefore form a nonempty upper interval with infimum below $c_{\max}$. Monotonicity gives the claims on either side of the threshold. Continuity gives $\Psi_{\theta_0}(c_{\text{crit}}) = p_0$ when $c_{\text{crit}} > 0$; at equality, the trajectory converges to the saddle, not high skill. □

At $c = c_{\max}$, system (31) is exactly the $a = 0$ limiting system in the proof of Theorem 3.3: the interior saddle no longer exists, and every interior state converges to $(1, 0)$.

Computing the threshold. For fixed initial state and parameters, c_{crit} is obtained by one-dimensional bisection on the monotone condition $\Psi_{\theta_0}(c) > p_0$, using a numerically traced separatrix. For

$$(a, \kappa, \Delta) = (0.5, 3, 2), \quad (\theta_0, p_0) = (0.4, 0.4), \quad \psi_{0.5}(0.4) \simeq 0.238151,$$

the initial state is in the low-skill basin. The reported calculation gives

$$\widehat{a}^* \simeq 0.413142, \qquad c_{\text{crit}} \simeq 0.094403,$$
$$\Psi_{0.4}(0.09) \simeq 0.391993 < 0.4 < 0.410220 \simeq \Psi_{0.4}(0.10).$$

Reported RK4 continuation and bisection refinement changes the baseline boundary by less than 2×10^{-8} and the cost threshold by less than 2×10^{-10}. This is a model-based illustration, not an empirically calibrated policy value.

Temporary intervention. If the cost is removed at time τ, the relevant criterion uses the original system:

$$z_c(\tau) \in B_{\text{high}}(a) \quad \Longleftrightarrow \quad p_c(\tau) < \psi_a(\theta_c(\tau)).$$

At equality, the state lies on the original stable manifold. Thus a temporary cost must move the state across the original boundary before withdrawal; merely entering the cost-modified high-skill basin is insufficient.

C.5 Detection-based delegation penalties

The study of AI-generated outputs [35] motivates our detection-based penalty model. This specification modifies the performance loss and delegation-update potential, but leaves the skill-update potential unchanged. Let $q \in [0, 1]$ be the probability that an AI-generated output is detected. If detected, the AI output incurs the harsher loss $1 - \theta_a$ instead of $(1 - \theta_a)^2$. Writing $a = \theta_a$, the effective evaluated AI-side loss is

$$L_a^{\text{pen}} = (1 - q)(1 - a)^2 + q(1 - a) = (1 - a)^2 + qa(1 - a).$$

The delegation-update potential is

$$\Phi_d^{\text{pen}}(\theta, p) = (1 - p)(1 - \theta)^2 + pL_a^{\text{pen}}.$$

The normalized dynamics are then

$$\begin{aligned} \dot{\theta} &= \theta(1 - \theta)\left((1 - p)(1 - \theta) - \Delta p\theta\right), \\ \dot{p} &= \kappa p(1 - p)[(1 - \theta)^2 - L_a^{\text{pen}}]. \end{aligned} \tag{33}$$

When $q = 0$, this reduces to the baseline ODE (7). The reliance dynamics coincide with the cost model at $c = qa(1 - a)$, so increasing q raises the basin boundary toward greater independent practice by the same separatrix comparison. Unlike a separate usage cost, this intervention changes the evaluated task loss itself. Figure 6c illustrates this effect.

C.6 Non-performance benefits of delegation

Conversely, let $\nu \geq 0$ represent saved time, reduced effort, or other benefits in loss units. Include only benefits not already in ℓ to avoid double counting. Delegation may then be preferred despite worse output quality. With Φ_s unchanged,

$$\Phi_d^{\nu} = (1 - p)(1 - \theta)^2 + p[(1 - a)^2 - \nu],$$

and

$$\dot{\theta} = f(\theta, p), \qquad \dot{p} = \kappa p(1 - p)[D_a(\theta) + \nu]. \tag{34}$$

For $0 \leq \nu < (1 - a)^2$,

$$\widehat{a}(\nu) = 1 - \sqrt{(1 - a)^2 - \nu} \in [a, 1), \qquad \dot{p} = \kappa p(1 - p)D_{\widehat{a}(\nu)}(\theta).$$

This is exactly the baseline system at effective capability $\widehat{a}(\nu)$, with two attractors separated by a separatrix. By Theorem 3.3, increasing ν lowers the boundary and enlarges the low-skill basin. Choosing delegation despite worse output quality need not be irrational when other benefits matter.

The range restriction matters. For $\nu \geq (1-a)^2$, every interior state converges to $(0,1)$:

$$p(t) \geq p_0 > 0, \qquad \theta(t) \leq M := \max\left\{\theta_0, \frac{1-p_0}{1-p_0+\Delta p_0}\right\} < 1,$$

$$\frac{d}{dt}\operatorname{logit} p \geq \kappa(1-M)^2 > 0.$$

Thus $p \to 1$; comparison with the fixed-P skill dynamics and $P \uparrow 1$ then gives $\theta \to 0$. At the threshold the effective capability reaches one, and the interior saddle collides with $(1,0)$. Sufficiently large benefits therefore eliminate the high-skill basin rather than merely shrinking it.

D Alternative performance, learning, and state specifications

Recall that ODE (7) relies on several baseline modeling choices: AI performance is represented by a fixed and correctly perceived effective skill θ_a; performance is measured by squared loss; learner ability is represented by a single scalar skill θ; and delegation provides no direct learning signal. These assumptions make the phase-plane analysis tractable. In this appendix, we relax them and study whether the qualitative conclusions depend on these particular specifications. Table 4 summarizes the assumptions and available results; numerical illustrations do not validate the model on people or establish universal bistability. Unless stated otherwise, $a = \theta_a \in (0,1)$, $\Delta, \kappa > 0$, and f is the normalized skill drift of the main model.

D.1 Performance and perception

D.1.1 Jagged AI performance

With a jagged AI, the performance loss of AI-generated outputs is random. We model this by letting $s_{t,r} \sim \mu_a$ denote the realized AI skill on task r in time window t. Define the effective AI loss

$$L_a := \mathbb{E}_{s\sim\mu_a}[(1-s)^2].$$

Compared with the baseline model, the jagged-AI extension changes only the AI-side performance loss from $\ell(a) = (1-\theta_a)^2$ to L_a. The loss-equivalent effective skill is $a_{\text{eff}} = 1 - \sqrt{L_a}$, so $(1-a_{\text{eff}})^2 = L_a$. As in Section A.1, each time window t contains m comparable tasks. During this window, $\theta(t)$ and $p(t)$ are held fixed, while task-level delegation decisions vary across tasks. For each $r \in [m]$, let

$$X_{t,r} \sim \text{Bern}(p(t)), \qquad s_{t,r} \sim \mu_a,$$

independently conditional on the current state $(\theta(t), p(t))$. Let

$$\bar{X}_t := \frac{1}{m}\sum_{r=1}^{m} X_{t,r}$$

be the realized fraction of delegated tasks. The learning mechanism within one time window is as follows.

• **(Task completion stage)** For each task $r \in [m]$, the learner delegates to AI if $X_{t,r} = 1$ and performs the task independently if $X_{t,r} = 0$.

• **(Evaluation stage)** The teacher provides task-level feedback. If the task is performed independently, the loss is $(1-\theta(t))^2$. If the task is delegated to AI, the loss is $(1-s_{t,r})^2$. Thus

$$\ell_{t,r} = (1-X_{t,r})(1-\theta(t))^2 + X_{t,r}(1-s_{t,r})^2.$$

• **(Skill update)** The skill update is unchanged from the baseline mechanism: independent work generates an error-driven learning signal, while delegation induces non-use decay. Averaging over the m tasks in the window gives

$$\theta(t+1) - \theta(t) = 2\eta\theta(t)(1-\theta(t))\left[(1-\bar{X}_t)(1-\theta(t)) + \Delta\bar{X}_t(\theta_d - \theta(t))\right].$$

• **(Delegation update)** The delegation update uses the realized feedback from delegated tasks and the learner's belief L_a about the expected AI loss on non-delegated tasks. The realized performance gap driving the update is

$$G_t = \frac{1}{m}\sum_{r=1}^{m}\left[(1-X_{t,r})((1-\theta(t))^2 - L_a) + X_{t,r}((1-\theta(t))^2 - (1-s_{t,r})^2)\right].$$

The delegation level updates as $p(t+1) - p(t) = \eta\kappa p(t)(1-p(t))G_t$. The key difference from Dynamics (10) is the introduction of the random AI skill $s_{t,r}$ and the resulting stochasticity in the delegation update.

Let $\mathcal{F}_t$ denote the information available before the task draws in window t. Since the draws are conditionally independent with $\mathbb{E}[(1-s_{t,r})^2 \mid \mathcal{F}_t] = L_a$, we have

$$\mathbb{E}[\bar{X}_t \mid \mathcal{F}_t] = p(t)$$

and

$$\begin{aligned}\mathbb{E}[G_t \mid \mathcal{F}_t] &= (1-p(t))((1-\theta(t))^2 - L_a) + p(t)((1-\theta(t))^2 - L_a)\\ &= (1-\theta(t))^2 - L_a.\end{aligned}$$

Therefore,

$$\mathbb{E}[\theta(t+1) \mid \mathcal{F}_t] = \theta(t) + 2\eta\theta(t)(1-\theta(t))\left[(1-p(t))(1-\theta(t)) + \Delta p(t)(\theta_d - \theta(t))\right],$$
$$\mathbb{E}[p(t+1) \mid \mathcal{F}_t] = p(t) + \eta\kappa p(t)(1-p(t))[(1-\theta(t))^2 - L_a].$$

After setting $\theta_d = 0$ and applying the baseline time normalization, the mean-field equations are

$$\dot{\theta} = f(\theta, p), \qquad \dot{p} = \kappa p(1-p)[(1-\theta)^2 - L_a] = \kappa p(1-p)D_{a_{\text{eff}}}(\theta).$$

Under the fixed AI-quality distribution and stated conditional independence, the mean-field equations reduce exactly to the baseline, whose basin theorem applies for $a_{\text{eff}} \in (0,1)$; individual random trajectories need not follow ODE orbits.

D.1.2 Biased perceived AI skill

Let $\widetilde{a} \in [0,1]$ be the learner's fixed perceived AI capability. True performance still uses a, but reliance uses $\widetilde{a}$:

$$\ell(\theta, p) = (1-p)(1-\theta)^2 + p(1-a)^2, \qquad \widetilde{\Phi}_d(\theta, p) = (1-p)(1-\theta)^2 + p(1-\widetilde{a})^2.$$

With Φ_s unchanged, applying the same multiplicative-update geometry gives

$$\dot{\theta} = f(\theta, p), \qquad \dot{p} = \kappa p(1-p)[(1-\theta)^2 - (1-\widetilde{a})^2]. \tag{35}$$

For $0 < \widetilde{a} < 1$, the dynamics reduce exactly to the baseline at $\widetilde{a}$; its separatrix is $\psi_{\widetilde{a}}$. For $a, \widetilde{a} \in (0,1)$ and $\theta \in (0,1)$, Theorem 3.3 therefore gives

$$\widetilde{a} > a \implies \psi_{\widetilde{a}}(\theta) \le \psi_a(\theta), \qquad \widetilde{a} < a \implies \psi_{\widetilde{a}}(\theta) \ge \psi_a(\theta).$$

Overestimating AI quality therefore enlarges the high-reliance basin, while underestimating it favors independent practice; realized loss still uses the true a.

D.1.3 Stochastic performance-gap feedback

The ODE analysis in Section 3 describes the averaged dynamics of repeated-task adaptation. In practice, task variation and finite samples make the performance feedback used to update delegation noisy. We consider the following Itô perturbation of the delegation update:

$$d\theta = f(\theta, p)\,dt, \qquad dp = \kappa p(1-p)[D_a(\theta)\,dt + \sigma\,dW_t], \qquad \sigma \ge 0, \tag{36}$$

where W_t is a standard Brownian motion and σ controls the magnitude of persistent noise in the delegation-update signal. The system reduces to the ODE at $\sigma = 0$. This models persistent feedback noise; it is not the diffusion limit of the finite-task process in Appendix A.2. Only the reliance equation is perturbed, isolating uncertainty in relative-performance feedback.

We also consider the state-dependent version

$$d\theta = f(\theta, p)\,dt, \qquad dp = \kappa p(1-p)[D_a(\theta)\,dt + 2\sigma(1-\theta)\,dW_t]. \tag{37}$$

This encodes the hypothesis that feedback is harder to interpret at lower skill; for fixed p, the noise amplitude decreases with θ.

State space and finite-horizon interpretation. For either model, put $r(\theta) = \sigma$ or $2\sigma(1-\theta)$. In the interior, Itô's formula gives

$$d \operatorname{logit} p = \left[\kappa D_a(\theta) + \frac{\kappa^2 r(\theta)^2}{2}(2p-1)\right] dt + \kappa r(\theta)\, dW_t.$$

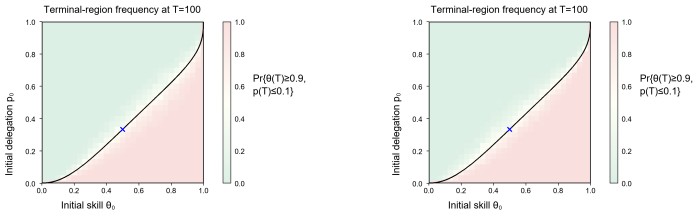
The transformed coefficients are bounded, preventing finite-time boundary hitting; the coordinate boundaries are invariant.

For each initial state z_0, the simulations estimate the terminal-region probability

$$P_{T,\varepsilon}(z_0) = \Pr_{z_0}\{\theta(T) \geq 1-\varepsilon,\ p(T) \leq \varepsilon\},$$

$$\widehat{P}_{T,\varepsilon}(z_0) = \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}\{\theta^{(j)}(T) \geq 1-\varepsilon,\ p^{(j)}(T) \leq \varepsilon\}.$$

The simulations use $T = 100$, $\varepsilon = 0.1$, and $N = 128$ paths at each point of a 31×31 initial-state grid, with time step 0.05. Figure 7 shows a sharp transition near the deterministic separatrix for this horizon and noise level.



(a) Noisy delegation update (36) (b) Noisy delegation update (37)

Figure 7: Finite-horizon frequency of ending in the high-skill terminal region $\theta(T) \geq 1-\varepsilon, p(T) \leq \varepsilon$ in simulations of SDEs (36) and (37), as a function of the initial state (θ_0, p_0), with default settings $(\theta_a, \kappa, \Delta, \sigma) = (0.5, 3, 2, 0.1)$. Green denotes frequency 0 and red frequency 1. This classification does not establish an asymptotic basin partition for the stochastic system.

D.1.4 Power-z performance and non-use losses

We consider a z-th power loss, which changes the sensitivity of feedback to errors:

$$\ell_z(\theta) = (1-\theta)^z, \qquad z \geq 1.$$

To keep the skill-decay term consistent with the same geometry, we also use the power distance $g_z(\theta, 0) = \theta^z$ in the normalized setting $\theta_d = 0$. Thus the skill-update and delegation-update potentials become

$$\Phi_s^z = (1-p)(1-\theta)^z + \Delta p \theta^z, \qquad \Phi_d^z = (1-p)(1-\theta)^z + p(1-a)^z.$$

Absorbing the derivative factor z into time and the relative rate κ gives

$$\dot{\theta} = \theta(1-\theta)[(1-p)(1-\theta)^{z-1} - \Delta p \theta^{z-1}], \qquad \dot{p} = \kappa p(1-p)[(1-\theta)^z - (1-a)^z]. \tag{38}$$

The case $z = 2$ is the baseline. The interior nullclines and their intersection are

$$\theta = a, \qquad p = q_z(\theta) := \frac{(1-\theta)^{z-1}}{(1-\theta)^{z-1} + \Delta \theta^{z-1}}, \qquad e_z^\dagger = (a, q_z(a)).$$

Here q_z denotes a nullcline, not the separatrix. It is strictly decreasing for $z > 1$ and constant, $q_1 = 1/(1+\Delta)$, for $z = 1$.

The interior equilibrium remains a saddle for every stated exponent. At $e_z^\dagger$, $\partial_p \dot{p} = 0$ and

$$\partial_p \dot{\theta} = -a(1-a)[(1-a)^{z-1} + \Delta a^{z-1}] < 0, \qquad \partial_\theta \dot{p} = -\kappa z q_z(a)(1-q_z(a))(1-a)^{z-1} < 0.$$

Thus the determinant is negative. Figure 8a illustrates the global basin deformation at $z = 3$. The nullclines and local stability are analytical results; the global basin comparison is numerical.

D.2 Skill formation and assisted production

Figure 8 illustrates these extensions together with the power-loss model.

D.2.1 Learning from AI during delegation

ODE (7) assumes that delegation provides no learning signal. We now relax this assumption by allowing AI-generated outputs to pull the learner's skill toward the AI skill level. This modifies the skill-update potential Φ_s, while the performance loss and delegation-update potential Φ_d remain unchanged. In the normalized setting $\theta_d = 0$, we replace the delegation-side decay term in Φ_s by a combination of non-use decay and learning from AI:

$$\Phi_s^\beta = (1-p)(1-\theta)^2 + p[\Delta\theta^2 + \beta(\theta - a)^2],$$

where $\beta \geq 0$ controls the strength of learning from AI. Applying the same multiplicative-update geometry and the same time rescaling as in ODE (7) gives

$$\dot{\theta} = \theta(1-\theta)[(1-p)(1-\theta) + p(\beta(a-\theta) - \Delta\theta)], \qquad \dot{p} = \kappa p(1-p)D_a(\theta). \tag{39}$$

The skill bracket can be rewritten as

$$(1-p)(1-\theta) + (\Delta+\beta)p(d_\beta - \theta), \qquad d_\beta = \frac{\beta a}{\beta + \Delta} < a.$$

This is the normalized general-default model in Appendix A.4. The full-delegation attractor shifts from $(0,1)$ to $(d_\beta, 1)$, while $(1,0)$ remains attracting. Its skill nullcline is

$$p = q_\beta(\theta) = \frac{1-\theta}{1-\theta+(\beta+\Delta)\theta - \beta a}, \qquad d_\beta < \theta < 1,$$

and it meets $\theta = a$ at the unchanged saddle $(a, q(a))$. Thus learning from AI raises retained skill under persistent delegation, but the limit remains below AI capability when $\Delta > 0$. For $\theta > a$, the same drift pulls skill downward toward a. Figure 8b illustrates the basin shift numerically.

D.2.2 Within-task AI assistance

To capture shared human–AI work, let $u \in [0,1]$ denote AI contribution intensity, with

$$L_{\text{add}}(\theta, u) = (1-u)(1-\theta)^2 + u(1-a)^2, \qquad \Phi_s^{\text{add}} = (1-u)(1-\theta)^2 + \Delta u\theta^2.$$

Applying the same multiplicative updates to these potentials gives, after the baseline normalization,

$$\dot{\theta} = \theta(1-\theta)[(1-u)(1-\theta) - \Delta u\theta], \qquad \dot{u} = \kappa u(1-u)D_a(\theta). \tag{40}$$

Identifying $u = p$ gives exactly the baseline ODE, so its global two-basin characterization also applies to this additive form of collaboration.

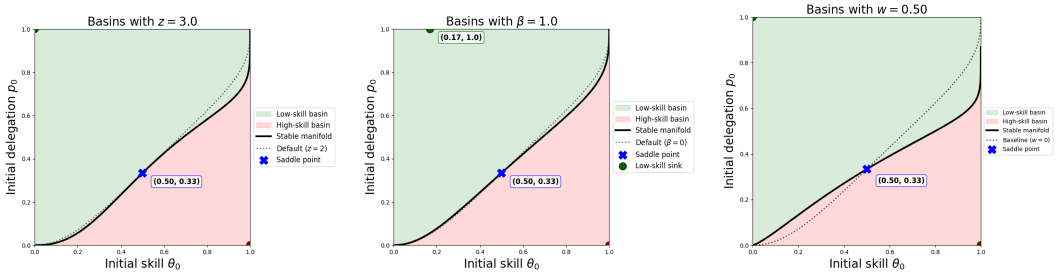
D.2.3 Skill-dependent AI performance

The baseline model assumes that AI performance depends only on the AI's effective skill θ_a. This is appropriate for settings where AI can complete a delegated task with little user input, such as routine homework problems or fixing isolated bugs. In other settings, AI output quality also depends on the user's skill: for example, in writing a paper, a more skilled user can provide better prompts, steer the interaction, and revise AI outputs more effectively.

Nonlinear blended output. For nonlinear blending, fix the AI share $q \in [0,1]$ within an assisted workflow. Let p be the evolving fraction of tasks using it. Blending skill before evaluating loss gives

$$L_c(\theta, q) = [1 - (1-q)\theta - qa]^2 = (1-q)(1-\theta)^2 + q(1-a)^2 - q(1-q)(\theta - a)^2. \tag{41}$$

The last term improves loss relative to random selection of the two outputs, not relative to the better pure output. With $w = 1-q$, the assisted loss becomes $[1 - w\theta - (1-w)a]^2$. Here w measures how much AI-assisted performance depends on the learner's own skill. The case $w = 0$ recovers the baseline model, where AI loss is $(1-\theta_a)^2$. Larger w corresponds to settings where effective AI output quality depends more strongly on the user's ability to guide and evaluate the AI. We keep the



(a) Basins of power-loss model (38) (b) Learning drift from AI (39) (c) Skill-dependent AI loss (42)

Figure 8: Numerical illustrations of power-loss learning, learning from AI, and skill-dependent assisted output (Sections D.1.4–D.2.3), with default parameters $(\theta_a, \kappa, \Delta) = (0.5, 3, 2)$ and respective extension parameters $z = 3$, $\beta = 1$, and $w = 0.5$. Each panel changes the stated model component; these comparisons illustrate the specified extensions rather than a theorem for arbitrary learning or production rules.

skill-update potential Φ_s unchanged to isolate the production effect; learning from assisted work is considered above. The delegation-update potential becomes

$$\Phi_d^{\text{int}} = (1-p)(1-\theta)^2 + p[1 - w\theta - (1-w)a]^2.$$

The dynamics are

$$\dot{\theta} = f(\theta, p), \qquad \dot{p} = \kappa p(1-p)\big[(1-\theta)^2 - [1 - w\theta - (1-w)a]^2\big]. \tag{42}$$

For $0 \le w < 1$, the performance gap factors as

$$(1-w)(a-\theta)[2 - (1+w)\theta - (1-w)a],$$

with positive last factor in the interior. Thus the reliance nullcline and saddle remain $\theta = a$ and $(a, q(a))$. At the saddle,

$$\partial_p \dot{p} = 0, \qquad \partial_\theta \dot{p} = -2\kappa(1-w)q(a)(1-q(a))(1-a) < 0,$$

and $\partial_p \dot{\theta} < 0$ as in the baseline model, so the Jacobian determinant remains negative and the equilibrium is a saddle. When $w = 1$, AI-assisted loss equals the learner's own loss, so the delegation drift vanishes identically and the adaptive-delegation mechanism degenerates.

For $w < 1$, increasing w leaves the saddle location unchanged but weakens the performance gap driving delegation updates.

Figure 8c illustrates this extension with $w = 0.5$. The displayed comparison uses $(a, \kappa, \Delta) = (0.5, 3, 2)$. Observe that the boundary shifts upward for low initial skill, reflecting that low-skill learners obtain less effective AI assistance and therefore have a weaker incentive to delegate. For higher initial skill, the boundary can shift downward, reflecting that skilled users can extract more value from AI and may maintain reliance over a larger region. These boundary shifts are numerical observations, distinct from the local saddle calculation.

D.3 Richer learner and AI states

D.3.1 Multiple skills

ODE (7) represents learner ability by a single scalar skill parameter θ. This scalar representation is useful because it allows a complete phase-space characterization: the equilibria can be solved explicitly, the interior equilibrium is a saddle, and its stable manifold forms a curve that partitions the state space into high- and low-skill basins. We now examine how the same reliance feedback extends beyond the one-dimensional skill case. Consider a two-skill model with skill vector $(\theta_1, \theta_2) \in [0,1]^2$. Let

$$\bar{\theta} = w_1\theta_1 + w_2\theta_2, \qquad w_1, w_2 > 0, \quad w_1 + w_2 = 1,$$

denote the task-relevant aggregate skill. The performance loss depends on the aggregate skill:

$$\ell(\theta_1, \theta_2) = (1 - \bar{\theta})^2.$$

Thus the delegation-update potential becomes

$$\Phi_d = (1-p)(1-\bar{\theta})^2 + p(1-a)^2.$$

The skill-update potential is also modified to reflect coordinate-wise skill decay. In the normalized setting $\theta_d = 0$, we use

$$\Phi_s(\theta_1, \theta_2, p) = (1-p)(1-\bar{\theta})^2 + \frac{p}{2}\sum_{i=1}^{2} \Delta_i \theta_i^2,$$

where $\Delta_i > 0$ is the non-use decay rate of skill i. The factor $1/2$ is a normalization convention and can be absorbed into Δ_i. Applying the multiplicative-update geometry coordinate-wise gives

$$\begin{aligned} \dot{\theta}_i &= \theta_i(1-\theta_i)[2w_i(1-p)(1-\bar{\theta}) - \Delta_i p \theta_i], \qquad i \in \{1,2\}, \\ \dot{p} &= \kappa p(1-p)[(1-\bar{\theta})^2 - (1-a)^2]. \end{aligned} \tag{43}$$

The delegation update is therefore still driven by the comparison between human and AI performance, but human performance now depends on the aggregate skill $\bar{\theta}$, while skill decay can differ across coordinates. This model contains the one-dimensional system as a special case. In particular, when $w_1 = w_2 = 1/2$, $\Delta_1 = \Delta_2 = \Delta$, and $\theta_1(0) = \theta_2(0)$, the diagonal subspace $\theta_1(t) = \theta_2(t)$ is invariant. On this diagonal, $\bar{\theta}(t) = \theta_1(t) = \theta_2(t)$, and (43) reduces to ODE (7).

Equilibrium and basin structure.

We next describe the stationary structure of the two-skill system. The two equilibria

$$(\theta_1, \theta_2, p) = (1, 1, 0), \qquad (\theta_1, \theta_2, p) = (0, 0, 1)$$

correspond respectively to the high-skill and low-skill regimes. The interior reliance nullcline is $\bar{\theta} = a$; intersecting it with the skill nullclines gives the equilibrium candidate characterized below.

Proposition D.1 (Local attraction and saddle structure). *For $a \in (0,1)$ and the positive parameters above, $e_H = (1,1,0)$ and $e_L = (0,0,1)$ are locally asymptotically stable relative to the cube. Let $S = w_1^2/\Delta_1 + w_2^2/\Delta_2$. The only candidate fully interior equilibrium is*

$$\theta_i^\dagger = \frac{w_i a}{\Delta_i S}, \qquad q := p^\dagger = \frac{2(1-a)S}{a + 2(1-a)S}. \tag{44}$$

It is feasible exactly when both skill coordinates are below one. Whenever feasible, it has two negative eigenvalues and one positive eigenvalue, and a smooth local stable manifold of dimension two.

Proof. We prove corner attraction using invariant neighborhoods, then locate the interior equilibrium and use its Jacobian to count stable directions.

Attracting corners. Set $(V_H, r_H) = (2 - \theta_1 - \theta_2, p)$ and $(V_L, r_L) = (\theta_1 + \theta_2, 1 - p)$. Let $w_- := \min_i w_i$, $w_+ := \max_i w_i$, $\Delta_- := \min_i \Delta_i$, and $\Delta_+ := \max_i \Delta_i$. Define

j	δ_j	γ_j	c_j	C_j
H	$\min\{1/2, (1-a)/\sqrt{2}\}$	$\kappa(1-a)^2/4$	$w_-^2/2$	Δ_+
L	$\min\{1/2, a/2\}$	$\frac{\kappa}{2}[(1-a/2)^2 - (1-a)^2]$	$\Delta_-/8$	$2w_+$

On $V_j \le \delta_j$, $r_j \le 1/2$, direct substitution gives

$$\dot{r}_j \le -\gamma_j r_j, \qquad \dot{V}_j \le -c_j V_j^2 + C_j r_j V_j \qquad (j = H, L). \tag{45}$$

For the negative quadratic term, use $1 - \bar{\theta} \ge w_- V_H$ near e_H, and $\theta_1^2 + \theta_2^2 \ge V_L^2/2$ near e_L. The constants also use $\theta_i \ge 1/2$ near e_H, and $1 - \theta_i \ge 1/2$ near e_L.

If $r_j(0) < 1/2$ and $V_j(0) e^{C_j r_j(0)/\gamma_j} < \delta_j$, a first-exit argument and Gronwall's inequality keep the trajectory in this neighborhood. More explicitly,

$$r_j(t) \le r_j(0) e^{-\gamma_j t}, \qquad U_j := V_j e^{-C_j \int_0^t r_j(s)\, ds}, \qquad \dot{U}_j \le -c_j U_j^2.$$

It follows that

$$0 \le V_j(t) \le \frac{V_j(0) e^{C_j r_j(0)/\gamma_j}}{1 + c_j V_j(0) t} \longrightarrow 0.$$

These bounds prove attraction and Lyapunov stability despite the zero eigenvalues at the corners. They also give explicit sufficient terminal neighborhoods for numerical classification.

Interior equilibrium. At an interior equilibrium, $\dot{p} = 0$ gives $\bar{\theta} = a$, and the skill equations give

$$\Delta_i q \theta_i = 2w_i(1-q)(1-a), \qquad a = \sum_i w_i \theta_i = \frac{2(1-q)(1-a)}{q} S.$$

Solving yields (44). Positivity is automatic, as is $0 < q < 1$; feasibility requires $w_i a < \Delta_i S$ for $i = 1, 2$.

Jacobian inertia. Write $\mathbf{w} = (w_1, w_2)^\top$ and

$$D = \operatorname{diag}(\theta_i^\dagger(1-\theta_i^\dagger)), \qquad M = q \operatorname{diag}(\Delta_1, \Delta_2) + 2(1-q)\mathbf{w}\mathbf{w}^\top,$$
$$\alpha = 2(1-a) + a/S, \qquad \beta = 2\kappa q(1-q)(1-a).$$

All diagonal entries of D and both scalars are positive, and $M \succ 0$. The equilibrium Jacobian and its symmetric form are

$$J^\dagger = \begin{pmatrix} -DM & -\alpha D\mathbf{w} \\ -\beta \mathbf{w}^\top & 0 \end{pmatrix} = T \begin{pmatrix} -B & -\mathbf{r} \\ -\mathbf{r}^\top & 0 \end{pmatrix} T^{-1}, \tag{46}$$

where

$$T = \operatorname{diag}(D^{1/2}, \sqrt{\beta/\alpha}), \qquad B = D^{1/2} M D^{1/2} \succ 0, \qquad \mathbf{r} = \sqrt{\alpha\beta}\, D^{1/2}\mathbf{w} \neq 0.$$

The Schur complement of $-B$ is $\mathbf{r}^\top B^{-1}\mathbf{r} > 0$. By symmetric block elimination, the inertia is that of $\operatorname{diag}(-B, \mathbf{r}^\top B^{-1}\mathbf{r})$: two negative and one positive eigenvalue. Similarity transfers these eigenvalues to $J^\dagger$. The stable-manifold theorem [28] gives the asserted local surface. □

Global scope. The local stable surface is not proved to be the entire basin boundary: the skill-gap cross-effects remain negative for $i \neq j$, since $\partial_{\theta_j} \dot{\theta}_i = -2w_i w_j (1-p)\theta_i(1-\theta_i) < 0$, and the planar Dulac argument does not classify this three-dimensional flow.

Boundary equilibria add another issue. On the face $\theta_j = b \in \{0, 1\}$, a candidate with $\theta_i, p \in (0, 1)$ is

$$\theta_i = \frac{a - w_j b}{w_i}, \qquad p = \frac{2w_i^2(1-a)}{2w_i^2(1-a) + \Delta_i(a - w_j b)}.$$

Its transverse stability is separate from the face dynamics. If $a = w_1$, the entire line $(1, 0, p)$ consists of equilibria.

Simulations.

Figure 9 illustrates the convergence behavior of the two-skill dynamics (43). In Figure 9a, trajectories are initialized with the same delegation level but nearby skill vectors. The displayed trajectories approach different equilibria. Figure 9b fixes one coordinate, for example $\theta_{2,0}$, and plots the resulting basins in the $(\theta_{1,0}, p_0)$-plane. The reduced slice exhibits the same qualitative two-basin structure as the one-dimensional model.

The two-skill model also captures heterogeneous decay across skills. For example, when $w_1 = w_2$, $\theta_1(0) = \theta_2(0)$, and $\Delta_1 < \Delta_2$, the first skill decays more slowly under delegation. For common interior initial skill and positive delegation, $\theta_1(t) > \theta_2(t)$ for $t > 0$, producing transient differentiation even when the two coordinates ultimately converge to the same corner.

Finite-grid evidence for a separating surface. The reported heterogeneous checks use all $13^3 = 2{,}197$ initial states in $\{0.05, 0.125, \ldots, 0.95\}^3$ per parameter setting. RK4 integration uses $T = 250$ and step 0.04. A state is classified only after entering one of the sufficient neighborhoods

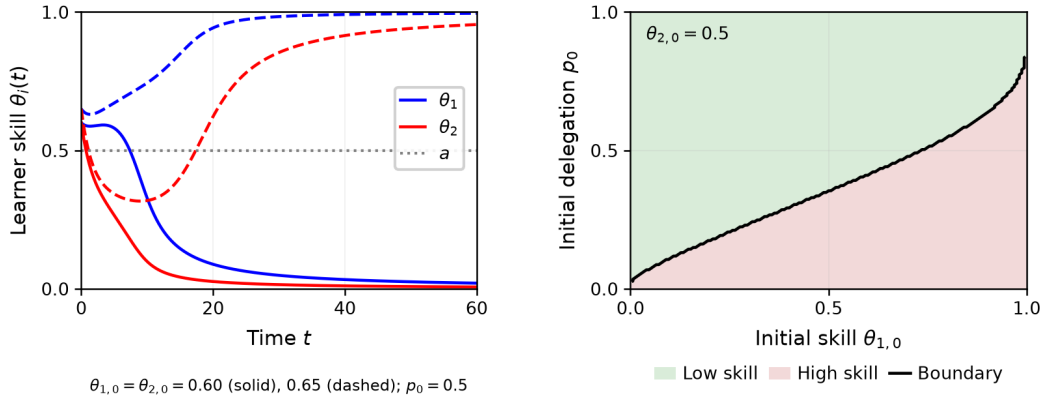
$$\mathcal{N}_j = \{r_j < 1/2,\ V_j e^{C_j r_j/\gamma_j} < \delta_j\}, \qquad j \in \{H, L\},$$

not merely by choosing the nearest corner.

Refining to $T = 300$ and step 0.02 changes no classifications. The recorded connectivity, saddle-perturbation, and slice checks are consistent with a single threshold surface, but do not prove global continuity or connectedness or exclude finer-scale fractal structure. The named implementation, `r06_multiskill_check.py`, is unavailable in the inspected archive; these are recorded results, not rerun checks.

$(a, \kappa, \Delta_1, \Delta_2, w_1, w_2)$	High-skill	Low-skill	Unresolved
$(.55, 2, 1, 2.5, .35, .65)$	615	1,582	0
$(.60, 5, .8, 4, .70, .30)$	633	1,564	0
$(.35, 1, 4, .7, .25, .75)$	1,700	497	0

Table 3: Reported finite-grid outcomes with unequal skill weights and decay rates. All 6,591 sampled trajectories entered an analytically sufficient attracting neighborhood. The grid does not exhaust the continuous state space.



(a) Skill trajectories from equal initial skill coordinates

(b) Basin slice at fixed $\theta_{2,0}$

Figure 9: Updated two-skill trajectories and basin slice from (43) for $(\theta_a, \kappa, \Delta_1, \Delta_2, w_1, w_2) = (0.5, 3, 1, 3, 0.5, 0.5)$. Panel (a) uses $\theta_{1,0} = \theta_{2,0} \in \{0.60, 0.65\}$ and $p_0 = 0.5$; panel (b) fixes $\theta_{2,0} = 0.5$. Nearby initial conditions can have different long-run outcomes. With equal weights, common interior initial skill, and positive delegation, $\Delta_1 < \Delta_2$ gives $\theta_1(t) > \theta_2(t)$ for $t > 0$, although the skills can converge to the same corner. The unequal-weight full-grid checks in Table 3 complement these illustrative slices.

D.3.2 Improving AI capability

ODE (7) assumes a fixed AI skill level θ_a. We now relax this assumption by allowing AI capability to improve according to $\dot{a} = \rho a(1-a)$, $\rho > 0$, with $a(0) \in (0, 1)$. This specification changes the performance loss and delegation-update potential through the time-varying AI loss $(1 - a(t))^2$, while the skill-update potential Φ_s remains unchanged:

$$\dot{\theta} = f(\theta, p), \qquad \dot{p} = \kappa p(1-p)[(1-\theta)^2 - (1-a(t))^2], \qquad \dot{a} = \rho a(1-a). \tag{47}$$

Equivalently, $\Phi_d(\theta, p; t) = (1-p)(1-\theta)^2 + p(1-a(t))^2$, and

$$a(t) = \frac{1}{1 + (a(0)^{-1} - 1)e^{-\rho t}} \longrightarrow 1.$$

The system is non-autonomous in (θ, p) and autonomous in the augmented state (θ, p, a).

Write $a_0 = a(0)$. For this logistic path to perfect capability, let u be the no-AI skill path with $u(0) = \theta(0)$, and define $L_0(t) = (1-u(t))^2$ and $L_A(t) = (1-p(t))(1-\theta(t))^2 + p(t)(1-a(t))^2$.

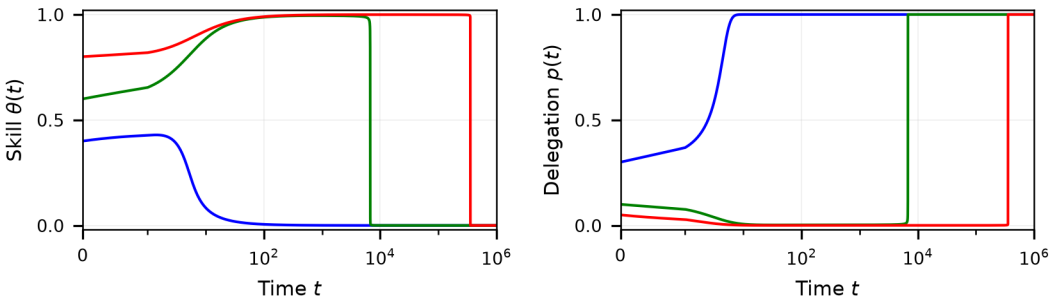
Proposition D.2 (Limit under logistic improvement). *If $\Delta, \kappa, \rho > 0$ and $(\theta(0), p(0), a_0) \in (0, 1)^3$, then $p(t) \to 1$, $\theta(t) \to 0$, and*

$$\theta(t) \sim \frac{1}{\Delta t}, \qquad L_0(t) \sim t^{-2}, \qquad L_A(t) = O(e^{-ct}) \quad \textit{for every } 0 < c < \min\{\kappa, 2\rho\}.$$

Consequently $L_A(t)/L_0(t) \to 0$.

Proof. For $\ell = \log[p/(1-p)]$, we have $\dot{\ell} \geq -\kappa(1-a(t))^2$ and

$$I := \int_0^\infty (1-a(t))^2 \, dt = (a_0 - 1 - \log a_0)/\rho < \infty.$$



(a) Unaided skill (b) AI delegation

Figure 10: Finite-horizon trajectories under logistic AI improvement (47), with $(a_0, \Delta, \kappa, \rho) = (0.5, 2, 3, 0.1)$ and horizon $T = 10^6$. All paths converge to $(\theta, p) = (0, 1)$ (Proposition D.2), although high skill can persist for long transients. Time is linear near zero and logarithmic above $t = 1$.

Thus $\ell(t) \geq \ell(0) - \kappa I$ and $p(t) \geq \varepsilon > 0$. Because $f(\theta, p)$ decreases with p, scalar comparison with fixed delegation ε gives $\theta(t) \leq M := \max\{\theta(0), q(\varepsilon)\} < 1$. Eventually $(1 - a(t))^2 \leq (1 - M)^2/2$, so $\dot{\ell} \geq \kappa(1 - M)^2/2$ and $p \to 1$ exponentially. For each $P < 1$, eventually $p \geq P$, whence the same scalar comparison gives $\limsup \theta \leq q(P)$; letting $P \uparrow 1$ gives $\theta \to 0$. Therefore $\dot{\ell} \to \kappa$ and $1 - p = O(e^{-(\kappa - \eta)t})$ for each $0 < \eta < \kappa$.

Put $v = 1/\theta$. Direct differentiation of the baseline drift gives

$$\dot{v} = \Delta p(1 - \theta) - (1 - p)(1 - \theta)^2 v.$$

Also $\dot{\theta} \geq -\Delta\theta^2$, so $v(t) \leq v(0) + \Delta t$. The second term therefore tends to zero, while the first tends to Δ; integration gives $v(t)/t \to \Delta$. On the no-AI path, $\dot{u} = u(1 - u)^2$ gives $u \to 1$ and $\frac{d}{dt}(1-u)^{-1} = u \to 1$, proving $L_0 \sim t^{-2}$. Finally, $0 \leq L_A \leq (1-p)+(1-a)^2$ and $1-a = O(e^{-\rho t})$, proving the stated exponential bound and loss ratio. □

Both losses vanish, but assisted output improves faster despite the loss of independent skill.

A useful interpretation is quasi-static. As AI capability improves, delegation becomes more attractive and the basin boundary moves accordingly. At each time, $\psi_{a(t)}$ is the boundary for fixed capability, not an invariant boundary of (47). To see the difference, put $R(t) = p(t) - \psi_{a(t)}(\theta(t))$. The fixed-capability invariance equation gives

$$\dot{R}(t)\big|_{R=0} = -\partial_a \psi_{a(t)}(\theta(t))\, \rho a(t)(1 - a(t)) \geq 0.$$

Thus motion of the frozen boundary contributes even when the instantaneous vector field is tangent to it.

Convergence to full delegation and zero independent skill can be very slow for initial states that would retain high skill under frozen $a = a_0$. Figure 10 shows finite-horizon trajectories integrated in logit coordinates. Under frozen $a = 0.5$, $(0.4, 0.3)$ lies on the low-skill side of the baseline separatrix, while $(0.6, 0.1)$ and $(0.8, 0.05)$ lie on its high-skill side.

D.4 Assumptions and scope of the conclusions

Table 4 maps assumptions to their use in our proofs and the relaxations analyzed here, not to necessary conditions for every model with the same conclusions.

Table 4: Model assumptions and available relaxations. "Exact" means a reduction of the specified dynamics, not empirical validation. Local stable-manifold results do not by themselves identify an entire global basin boundary.

Assumption	Type and role	Baseline result using it	Relaxation and status
Scalar skill $\theta \in [0, 1]$.	State/scale restriction; planar dynamics.	Global graph and ordering in Theorems 3.2–3.3.	Section D.3.1: local attractors and feasible interior saddle surface; numerical global basins.
Squared loss $\ell(\theta) = (1 - \theta)^2$.	Functional form; explicit error and nullclines.	Equilibria and performance reversal, Theorems 3.1 and 3.5.	Section D.1.4: power-loss nullclines and local saddle; numerical basin comparison.
Error-driven practice: independent-work term $\ell(\theta)$ in Φ_s.	Behavioral direction: improvement through $-\ell'(\theta)$.	Skill drift (3) and the fixed-delegation limit.	Power losses retain this direction; general non-error-driven learning is not analyzed.
Non-use decay $g = (\theta - \theta_d)^2$, $\Delta > 0$; baseline $\theta_d = 0$.	Behavioral drift/functional form; zero default is not a time normalization.	Baseline and general equilibria, Theorems 3.1 and A.4.	Section A.4: attractor $(\theta_d, 1)$ for $0 < \theta_d < \theta_a$. Power decay: Section D.1.4.
Delegation displaces effective practice.	Coupling: practice share $1 - p$.	Full-delegation skill and feedback loop.	Section D.2.1: retained skill $\beta\theta_a/(\beta + \Delta)$; numerical basin deformation.
Performance-only reliance potential $\Phi_d = (1-p)\ell(\theta)+p\ell(\theta_a)$.	Behavioral objective: current loss, without cost or convenience benefit.	Reliance drift (5) and capability ordering.	Costs/benefits: exact reductions in stated ranges. Monotone responses: Proposition C.1. Asymmetry: piecewise-smooth boundary.
Multiplicative factors $\theta(1 - \theta)$ and $p(1 - p)$.	Geometry: invariant faces and vanishing endpoint mobility.	Invariance, corner equilibria, global basins.	Positive reliance multipliers: Sections C.1–C.3. Arbitrary preconditioners/hard caps are not covered.
Fixed AI capability.	Stationarity: autonomous dynamics, fixed benchmark.	Equilibria and fixed-capability comparative statics.	Fixed-distribution jaggedness: exact expected-loss reduction. Logistic improvement to one: full reliance (Proposition D.2); frozen boundaries are not invariant.
Learner-independent AI loss $\ell_a = \ell(\theta_a)$.	Production: quality independent of human prompting/revision skill.	Delegation nullcline and interior saddle.	Section D.2.3: same nullcline/saddle for $w < 1$, numerical global comparison; no reliance adaptation at $w = 1$.
Binary task allocation $X \sim \mathrm{Bern}(p)$.	Aggregation: whole-task independent/delegated choices.	Expected potentials and the task-level derivation.	Affine output/practice effects with the same adaptation law: exact continuous-contribution reduction. Nonlinear blending is separate.
Many comparable tasks per short window.	Mean field: averaged feedback, slowly changing state.	Finite-horizon tracking (Section A.2); ODE theorems.	Persistent noise: Section D.1.3 (numerical). Finite-task measurement error: Section E. Neither is the ODE limit.
Accurate human–AI evaluation on a common loss scale.	Information: reliance uses the actual expected gap.	Performance comparison and skill measurement.	Constant perceived-AI bias: exact substitution; noisy signals: simulations. General evaluator bias/incomparable tasks are not analyzed. Identification uses randomized matched probes.

E Empirical calibration and falsifiable predictions

We outline how the parameters of the main specification can be estimated from repeated-task data and how the resulting phase portrait could be tested. The goal is not to provide a fully calibrated psychometric procedure, but to make clear which observables correspond to the model variables and which empirical patterns would support or challenge the basin mechanism.

E.1 Measurements and scale-free identification

Consider an experimental setup with $T+1$ equally spaced observation windows, $t = 0, \dots, T$. In each window t, the learner faces m_t comparable tasks indexed by $r = 1, \dots, m_t$. For each task, we observe a delegation indicator $X_{t,r} \in \{0, 1\}$, where $X_{t,r} = 1$ means that the task is delegated to AI, and an evaluation loss $\ell_{t,r} \in [0, 1]$ assigned by the evaluator to the realized output. We also assume access to an AI-only benchmark loss $\ell^a_{t,r} \in [0, 1]$, obtained either by evaluating the AI on the same task or from a held-out benchmark of comparable tasks. All losses are normalized to lie in $[0, 1]$.

The empirical delegation level in window t is estimated by the fraction of tasks delegated to AI:

$$\widehat{p}_t := \frac{1}{m_t} \sum_{r=1}^{m_t} X_{t,r}.$$

An elicited willingness to delegate can provide a separate measure, but its agreement with observed delegation must be checked.

Estimating AI effective skill.

Under the squared-loss specification, the AI effective skill is defined through its expected loss. We estimate it by

$$\widehat{L}_a := \frac{1}{\sum_{t=0}^{T} m_t} \sum_{t=0}^{T} \sum_{r=1}^{m_t} \ell^a_{t,r}, \qquad \widehat{\theta}_a := 1 - \sqrt{\widehat{L}_a}.$$

If AI performance is jagged, this estimate should be interpreted as the loss-equivalent effective skill satisfying $(1 - \widehat{\theta}_a)^2 = \widehat{L}_a$.

Estimating learner skill.

If a window contains no independently completed tasks, or if independently completed tasks are selected non-randomly by the learner, θ_t is not identified from submitted outputs alone.

We therefore estimate learner skill using randomized no-AI probes $\mathcal{P}_t$ sampled from the same task distribution, independently of delegation choices, possibly combining nearby windows or using smoothing. With $\widehat{L}^H_t = |\mathcal{P}_t|^{-1} \sum_{r \in \mathcal{P}_t} \ell_{t,r}$ and an independent matched AI benchmark, the operational estimates are

$$\widehat{p}_t = \frac{1}{m_t} \sum_{r=1}^{m_t} X_{t,r}, \qquad \widehat{\theta}_t = 1 - \sqrt{\widehat{L}^H_t}, \qquad \widehat{\theta}_a = 1 - \sqrt{\widehat{L}_a}. \tag{48}$$

These quantities express task-specific skill through expected loss, not general psychometric ability.

Estimating scale-free rates.

The change across an empirical window depends on its learning scale. We fit coefficients a_1, a_2, a_3 for independent learning, non-use decay, and reliance adjustment. The normalized parameters $\Delta = a_2/a_1$ and $\kappa = a_3/a_1$ measure decay and reliance adaptation relative to learning; these ratios are invariant to a common rescaling of time. The scale a_1 must be fitted when unknown, but does not affect the phase portrait. Define $\delta\widehat{z}_t = \widehat{z}_{t+1} - \widehat{z}_t$ and

$$\begin{aligned} x_{1t} &= \widehat{\theta}_t(1 - \widehat{\theta}_t)^2(1 - \widehat{p}_t), \qquad x_{2t} = \widehat{\theta}^2_t(1 - \widehat{\theta}_t)\widehat{p}_t, \\ x_{3t} &= \widehat{p}_t(1 - \widehat{p}_t)[(1 - \widehat{\theta}_t)^2 - \widehat{L}_a]. \end{aligned} \tag{49}$$

For short windows, the Euler equations are $\delta\widehat{\theta}_t \simeq a_1 x_{1t} - a_2 x_{2t}$ and $\delta\widehat{p}_t \simeq a_3 x_{3t}$. An initial estimate is

$$\widehat{\mathbf{a}} \in \operatorname*{arg\,min}_{a_1, a_2, a_3 \geq 0} \sum_{t=0}^{T-1} \left[(\delta\widehat{\theta}_t - a_1 x_{1t} + a_2 x_{2t})^2 + (\delta\widehat{p}_t - a_3 x_{3t})^2 \right], \qquad \widehat{\Delta} = \frac{\widehat{a}_2}{\widehat{a}_1}, \quad \widehat{\kappa} = \frac{\widehat{a}_3}{\widehat{a}_1}. \tag{50}$$

This is nonnegative least squares. The ratios require $\widehat{a}_1 > 0$; otherwise the fit does not recover the normalized parameters. When each window is known to equal one normalized time unit, fix $a_1 = 1$. The corresponding unconstrained least-squares formulas are

$$\widehat{\Delta} = \frac{\sum_t x_{2t}(x_{1t} - \delta\widehat{\theta}_t)}{\sum_t x_{2t}^2}, \qquad \widehat{\kappa} = \frac{\sum_t x_{3t}\delta\widehat{p}_t}{\sum_t x_{3t}^2},$$

using reliable skill estimates, $x_{2t} > 0$ for the first fit, and x_{3t} away from zero for the second.

Structural identification conditions.

With exact states and the specified discrete transition law, define the skill design matrix X_s with rows $(x_{1t}, -x_{2t})$. The rates are uniquely identified under

$$\operatorname{rank} X_s = 2, \qquad \sum_{t=0}^{T-1} x_{3t}^2 > 0.$$

The first condition separates independent learning from non-use decay; the second identifies reliance responsiveness. For exact interior states,

$$\frac{x_{2t}}{x_{1t}} = \frac{p_t\theta_t}{(1-p_t)(1-\theta_t)}, \qquad x_{3t} = p_t(1-p_t)[(1-\theta_t)^2 - (1-\theta_a)^2].$$

Identification therefore needs variation in practice relative to delegation. Precision depends on the elapsed normalized horizon $H = a_1 T$, observation and task counts, and regressor conditioning: nearly proportional skill regressors, reliance near zero or one, or skill near AI capability can leave little information even in a long record. For ODE observations, these are short-window conditions, not finite-sample guarantees.

E.2 Measurement-aware trajectory fitting and uncertainty

The initial regression has noisy regressors. Even independent measurement errors e_t give negatively correlated differences $e_{t+1} - e_t$, so ordinary regression intervals based on residual variance are not justified. We instead fit a latent trajectory to observed losses and delegation counts.

Observation model and objective. For the synthetic study, the latent state follows the stated Euler recursion with coefficients (a_1, a_2, a_3). Conditional on that trajectory, assume independent measurements

$$N_t := \sum_r X_{t,r} \sim \mathrm{Bin}(m_t, p_t), \qquad \widehat{L}_t^H \sim \mathcal{N}((1-\theta_t)^2, \sigma^2/h_t), \qquad \widehat{L}_a \sim \mathcal{N}(L_a, \sigma^2/m_a).$$

Here h_t and m_a are the numbers of probes and benchmark observations, and $L_a = (1-\theta_a)^2$. Up to a data-dependent constant, twice the negative log-likelihood is

$$\mathcal{Q} = \sum_{t=0}^{T} \left[\frac{h_t}{\sigma^2}(\widehat{L}_t^H - (1-\theta_t)^2)^2 + 2m_t D_{\mathrm{Ber}}(\widehat{p}_t \| p_t) \right] + \frac{m_a}{\sigma^2}(\widehat{L}_a - L_a)^2, \tag{51}$$

where $D_{\mathrm{Ber}}(u\|v) = u\log(u/v) + (1-u)\log((1-u)/(1-v))$, with the usual $0\log 0$ convention. Gaussian observations are assumed for this recovery experiment; individual human task losses need not be Gaussian.

Computation. Initialize from (50) and (48), and fit

$$\gamma = (\log a_1, \log\Delta, \log\kappa, \operatorname{logit}\theta_0, \operatorname{logit} p_0, \operatorname{logit}\theta_a), \qquad a_2 = a_1\Delta, \quad a_3 = a_1\kappa.$$

Minimize $\mathcal{Q}$ over feasible Euler trajectories, not merely positive rates; use integrated ODE transitions for longer windows to limit transition error.

For the covariance calculation, write $\mathcal{Q} = \|r(\gamma)\|^2$ by stacking

$$r_{H,t} = \frac{\sqrt{h_t}}{\sigma}[\widehat{L}_t^H - (1-\theta_t)^2], \quad r_{p,t} = \operatorname{sign}(\widehat{p}_t - p_t)\sqrt{2m_t D_{\mathrm{Ber}}(\widehat{p}_t\|p_t)}, \quad r_a = \frac{\sqrt{m_a}}{\sigma}(\widehat{L}_a - L_a).$$

This retains initial-state and AI-benchmark uncertainty.

Approximate confidence intervals. Let $J = \partial r/\partial \gamma$ at the fitted parameters. With the specified observation variances and a full-rank local information matrix, use

$$\widehat{\Sigma} = (J^\top J)^{-1}, \qquad \mathrm{CI}_{.95}(\Delta) = \exp(\widehat{\gamma}_2 \pm 1.96\sqrt{\widehat{\Sigma}_{22}}), \qquad \mathrm{CI}_{.95}(\kappa) = \exp(\widehat{\gamma}_3 \pm 1.96\sqrt{\widehat{\Sigma}_{33}}).$$

These intervals use a local covariance approximation under the stated observation model and variances, not an exact finite-sample result. Process noise, evaluation bias, estimated variances, or weak information need further treatment. The coverage study below tests this approximation under the synthetic design.

E.3 Synthetic recovery of parameters and basin membership

Data-generating design. To assess parameter recovery from noisy observations, we generate Euler mean-field trajectories with

$$(\theta_a, \Delta, \kappa) = (0.5, 2, 3), \qquad a_1 = 0.5, \qquad (\theta_0, p_0) = (0.4, 0.3).$$

At each of the $T + 1$ observations, use $h = 20$ no-AI probes, task-loss noise standard deviation $\sigma = 0.05$, and m independent delegation indicators. The AI benchmark has $m_a = 4000$ observations. Average losses are clipped to $[10^{-8}, 1 - 10^{-8}]$ if necessary; no clipping occurred in the reported replications. This is measurement noise around a deterministic trajectory, not noise in learning from finitely many practice events.

Parameter recovery. For the first dataset at $(T, m) = (20, 200)$, the fitted rates are $(\widehat{a}_1, \widehat{a}_2, \widehat{a}_3) = (0.557, 1.039, 1.487)$, with $\mathcal{Q} = 42.170$. Their ratios and covariance intervals give

$$\widehat{\Delta} = 1.865 \quad [1.696, 2.051], \qquad \widehat{\kappa} = 2.668 \quad [2.176, 3.272].$$

Brackets denote nominal 95% confidence intervals; displayed rates are rounded.

Table 5: Reported synthetic recovery over 200 independent replications per design. Estimates are median [interquartile range]; coverage is the proportion of nominal 95% intervals containing the true value. All fits converged.

T	m	$\widehat{\Delta}$	$\widehat{\kappa}$	Coverage (Δ, κ)
10	50	2.032 [1.903,2.159]	3.351 [2.339,4.417]	$(94.5\%, 94.0\%)$
20	200	2.006 [1.933,2.081]	3.027 [2.846,3.273]	$(93.5\%, 92.5\%)$
20	800	1.992 [1.950,2.043]	2.987 [2.856,3.125]	$(93.5\%, 92.5\%)$

Across the three designs in Table 5, both estimates are within 20% of truth in 28.5%, 93.0%, and 99.5% of replications, respectively. Recovery improves across these designs, but the results do not establish a universal window requirement, and interval coverage is below the nominal level.

Basin membership. The quantity of substantive interest is which side of the separatrix a learner occupies. For each of the 200 fits with $(T, m) = (20, 200)$, compute the fitted ODE separatrix $\widehat{\psi}$ and classify the deterministic 41×41 initial-state grid on $[0.05, 0.95]^2$. Agreement with the true-parameter ODE is 99.08% over 336,200 classifications. It is 90.44% in the vertical strip $|p_0 - \psi(\theta_0)| < 0.05$ and 100% outside. Errors concentrate near the boundary, where small state or parameter changes matter most. Test states are known in this calculation; it excludes uncertainty in new learners' states and real-world model error.

These reported results test recovery under the model, not its accuracy as a model of human behavior.

E.4 A prospective longitudinal test

Empirical predictions and falsification.

For learners with common parameters, states above the fitted separatrix should approach high delegation and low independent skill, and states below it the reverse, not necessarily monotonically. Near the boundary, delayed access, required independent practice, or conservative delegation feedback should shift some trajectories toward the high-skill basin. Holding learning and adaptation parameters fixed, increasing AI capability should weakly lower the boundary and enlarge the low-skill basin.

Greater reliance followed by weaker independent performance alone would not distinguish a basin mechanism from a smooth dose–response. Its distinctive prediction is a thresholded long-run phase structure: the deterministic asymptotic outcome map is discontinuous across the separatrix, but finite-horizon trajectories are continuous in initial states, with further smoothing from noise and heterogeneity. The finite-horizon signatures are progressive separation and heightened boundary sensitivity, not a literal discontinuity; bimodality requires sampled states in both basins.

For future validation, we propose following students solving systems of linear equations. One topic supports a scalar task-specific skill measure, with matched difficulty, a common evaluation objective, and fixed AI. The AI benchmark should lie within the learners' skill range, with variation in performance gaps and delegation: consistently small gaps reveal little about κ.

Measurements and practice exposure. Record delegation choices and their access, timing, and feedback conditions. Randomly assign matched no-AI probes independently of voluntary delegation, and use an independent matched AI benchmark, as in (48). Estimate skill from these probes alone, and account for their practice exposure and feedback in the fitted model. Continuing probes require practice-adjusted predictions; baseline follow-up predictions apply only after practice-inducing probes stop, from the resulting state.

Estimation and held-out predictions. Fit early-session measurements using Sections E.1–E.2, then observe optional AI use during follow-up and a final matched no-AI assessment. From the post-probe state $z_0 = (\theta_0, p_0)$, the estimated baseline long-run regime is

$$\widehat{B}(z_0) = \begin{cases} \text{high}, & p_0 < \widehat{\psi}(\theta_0), \\ \text{low}, & p_0 > \widehat{\psi}(\theta_0), \end{cases}$$

leaving near-boundary cases uncertain after accounting for uncertainty in both states and parameters.

Compare held-out observations with fitted finite-horizon trajectories, not equilibrium outcomes, at horizons long enough to distinguish predictions. A model without a long-run basin partition provides a comparison; equal or better held-out predictions would weaken support for the basin mechanism.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction state the main claims of the paper: the framework, the fixed-versus-adaptive delegation contrast, the basin-partition result, the short-run/long-run performance tradeoff, and robustness under alternative specifications. These claims correspond to Theorems 3.1-3.5 and Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: Section 5 discusses the stylized nature of the model, the distinction between qualitative dynamical analysis and calibrated structural modeling, and directions for richer usage patterns and empirical calibration. Section 4 and Section D also discuss robustness to alternative specifications.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes].

   Justification: The main theoretical results state their assumptions in Section 3, and full proofs are provided in Section B. The paper also gives proof sketches in Section 3.2 and well-posedness details in Section A.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: Sections D.3.1, C.1, and E.3 specify the synthetic designs, parameters, grids, numerical settings, and recovery protocol. The supplementary reproducibility archive contains available supporting scripts and recorded outputs; Section D.3.1 identifies the unavailable multiskill-check implementation. These experiments evaluate the stated models and do not constitute human-participant validation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: We provide the illustrative simulation code and a supplementary reproducibility archive for the new synthetic recovery and robustness checks, including available scripts, parameter settings, a README, and recorded outputs. Script coverage is partial, as noted in Section D.3.1. The study uses model-generated data and no participant dataset.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: Sections E.1–E.3 specify the measurement model, fitting objective, initialization, sample sizes, random seed, replications, and basin-classification grid. The robustness appendices state their integration and parameter settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

Justification: Sections E.2–E.3 define approximate 95% confidence intervals, report medians and interquartile ranges over 200 independent replications per design, and report repeated-sample interval coverage. The stated assumptions and observed undercoverage limit the interpretation of these intervals.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No].

   Justification: The supplementary scripts and numerical settings describe the low-dimensional computations, but the manuscript does not record the hardware, memory use, or execution times of the reported synthetic experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work is theoretical, uses no human-subject experiments, and does not release datasets or models. It is intended to clarify risks and mitigation directions for adaptive AI delegation.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: Section 5 discusses the risk that improved current output conceals foregone practice and long-run skill loss. Section C analyzes sustained and temporary interventions, while distinguishing model-based thresholds from a complete welfare comparison.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper does not release data, models, or other assets with misuse risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A].

Justification: The paper does not use existing datasets, codebases, pretrained models, or other assets as part of the research methodology.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes].

Justification: The supplementary reproducibility archive provides a README, available supporting scripts, and recorded outputs for the synthetic recovery and robustness checks. The README identifies dependencies and invocation guidance; the manuscript specifies the designs, limitations, and interpretation. These materials contain model-generated results and no human-participant data.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A].

Justification: The paper does not involve crowdsourcing, human-subject experiments, or collection of human-participant data.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not conduct human-subject research, crowdsourcing studies, or collection of participant data, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: The core methodology is a mathematical model and analysis of dynamical systems; LLMs are not used as an important, original, or non-standard component of the research method.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.