THE GEOMETRY OF LEARNING UNDER AI DELEGATION

By

Lingxiao Huang and Nisheeth K. Vishnoi

March 2026

COWLES FOUNDATION DISCUSSION PAPER NO. 2499

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

YALE UNIVERSITY

Box 208281

New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

The Geometry of Learning Under AI Delegation

Lingxiao Huang
Nanjing University

Nisheeth K. Vishnoi
Yale University

Abstract

As AI systems shift from tools to collaborators, a central question is how the skills of humans relying on them change over time. We study this question mathematically by modeling the joint evolution of human skill and AI delegation as a coupled dynamical system. In our model, delegation adapts to relative performance, while skill improves through use and decays under non-use; crucially, both updates arise from optimizing a single performance metric measuring expected task error. Despite this local alignment, adaptive AI use fundamentally alters the global stability structure of human skill acquisition. Beyond the high-skill equilibrium of human-only learning, the system admits a *stable* low-skill equilibrium corresponding to persistent reliance, separated by a sharp basin boundary that makes early decisions effectively irreversible under the induced dynamics. We further show that AI assistance can strictly improve short-run performance while inducing persistent long-run performance loss relative to the no-AI baseline, driven by a negative feedback between delegation and practice. We characterize how AI quality deforms the basin boundary and show that these effects are robust to noise and asymmetric trust updates. Our results identify stability, not incentives or misalignment, as the central mechanism by which AI assistance can undermine long-run human performance and skill.

Contents

1 Introduction

AI systems are increasingly used to support cognitive tasks such as writing, coding, and education [2, 15, 3, 16]. A key reason for this growing reliance is their ability to deliver immediate performance gains: improving accuracy, speed, or output quality over a short horizon [3, 16, 15]. Accordingly, much of the current discourse treats AI use as a one-step decision, evaluated primarily through its short-term impact on task performance. Moreover, AI performance is often uneven across instances, alternating between strong and weak outputs even within the same task [6]. While such variability shapes users' immediate experiences, it remains unclear whether long-run effects are driven by stochastic performance fluctuations or by more structural features of repeated reliance.

Crucially, repeated use of AI tools turns users into learners: agents whose future performance depends on how current task execution shapes underlying skill. However, when such tools are used repeatedly, their impact has been shown to extend beyond immediate performance. Emerging longitudinal studies suggest that sustained AI use can alter users' underlying abilities, leading to reduced independent performance, over-dependence, or skill atrophy once assistance is reduced or removed [13, 3, 2, 8, 1]. For example, writers who rely on AI assistance exhibit weaker recall of their own text, reduced lexical diversity, and diminished sense of ownership, with these deficits persisting even after access to the AI is withdrawn [13]. Related field experiments in coding and mathematics similarly find that AI assistance improves immediate performance and throughput, but leads to diminished human skill acquisition [1] and worse post-test performance once the assistance is removed [3, 16]. Moreover, empirical studies report substantial heterogeneity in outcomes across different subjects, with some users benefiting from AI assistance while others stagnate or regress [29, 16, 32].

These findings point to a fundamental tension. AI assistance is often adopted because it improves short-term task performance, and these gains are likely to grow as AI systems become more capable. Yet repeated reliance may reshape human capability in ways that are invisible to one-step evaluations and short-run benchmarks. This raises a central question: *How do repeated AI use and growing capability shape human skill and performance over time, and when do short-term gains become long-term losses?*

Much of the existing literature is empirical, documenting patterns of reliance and skill change in specific tasks. To our knowledge, the long-term effects of repeated AI assistance on human ability have not been studied in a mathematical framework. Existing learning and decision-theoretic models can be applied in restricted settings, but they do not capture the interaction between short-term reliance, evolving human capability, and increasing AI strength.

Related work. Work on AI-assisted learning documents substantial short-term gains in productivity and output quality alongside declines in retention, transfer, or post-test performance once assistance is removed, with heterogeneous effects across learners with distinct reliance [13, 3, 16, 29, 30]. A recent survey [8] reports a negative relationship between the frequency of AI usage and cognitive load. These findings are often interpreted through behavioral and human–automation lenses, including cognitive offloading, automation bias, and incentive-driven task substitution [21, 20, 27], but remain largely descriptive and do not model long-run dynamics.

Separately, mathematical models of human learning, ranging from learning-forgetting curves and stochastic approximation to Bayesian updating, reinforcement learning, and reinforced urn processes, capture error-driven improvement and use-it-or-lose-it decay under fixed task execution

[7, 19, 12, 11, 25, 22]. In the absence of delegation, such models typically admit a unique stable learning equilibrium. Classical results in stochastic approximation show that many such learning rules track limiting ordinary differential equations whose global behavior governs convergence [4], but these analyses do not account for settings with adaptive delegation or the joint evolution of human skill and AI reliance under a shared performance objective.

Related work on performative prediction and feedback loops studies how deployed models reshape data distributions [23, 17], focusing on model learning under distribution shift rather than the evolution of human skill under adaptive reliance.

Our contributions. We provide a mathematical framework for studying the long-term effects of AI assistance on human performance. Our key abstraction models repeated AI use as a coupled dynamical system in which performance-driven delegation decisions shape the future evolution of human skill. A central feature is that delegation and skill updates are locally aligned, both optimize the same instantaneous objective, yet interact to produce qualitatively different global learning dynamics. Our main contributions are:

- *A unified dynamical model of AI-assisted learning.* We formulate repeated AI delegation and human learning as a coupled dynamical system driven by a single performance objective. Despite local alignment, delegation adapts to relative performance and skill evolves through use and non-use, the resulting global dynamics differ qualitatively from human-only learning.

- *New equilibria under AI assistance.* In contrast to human-only settings without AI assistance, which admit a unique high-skill equilibrium with no delegation, we prove that AI assistance introduces additional equilibria, including a stable low-skill equilibrium corresponding to persistent delegation and an interior saddle point whose skill matches that of the AI (Theorem 3.1).

- *Geometry and irreversibility.* We show that the stable manifold of the interior saddle point forms a one-dimensional global basin boundary separating the state space into exactly two basins of attraction, yielding path dependence and irreversible learning outcomes (Theorem 3.2). This irreversibility does not arise from decay alone, which is reversible in human-only learning, but from the interaction between skill dynamics and adaptive delegation.

- *Short-run gains, long-run losses.* We prove that AI assistance can strictly improve performance in the short run while inducing persistent long-run performance loss relative to the no-AI baseline, even when evaluation is based solely on task performance and the AI is highly capable but imperfect (Theorem 3.4).

- *Mechanism of degradation.* We identify a negative feedback mechanism in which higher initial delegation reduces practice, lowers skill, and induces further delegation through performance-driven updates. We show that this feedback strengthens as AI quality improves (Lemma 3.5).

- *Effect of AI quality.* We characterize how AI capability deforms the basin boundary, showing a monotonic dependence on AI quality (Theorem 3.3). Improving AI capability expands the low-skill basin and can dramatically extend the duration of short-term performance gains, even as it exacerbates long-run skill degradation (Figure 2c).

- *Extensions.* We show that the qualitative phase structure of the dynamics, including multiple equilibria, basin boundaries, and irreversibility, persists under jagged AI performance, noisy or asymmetric delegation updates, and alternative performance objectives (Sections 4 and 7).

Together, our results show that improving AI capability can amplify short-term gains while inducing persistent long-run losses in both human skill and task performance, even under fully rational, performance-driven behavior. These effects arise from the global dynamics induced by adaptive delegation, not from incentive misalignment.

2 Model

We study a repeated task-execution scenario in which a user chooses between performing a task independently or delegating it to AI and receives performance feedback on the resulting output. When such feedback is observed repeatedly, the user acquires an evolving internal capability; we therefore model the user as a *learner* with a time-varying skill parameter. Performance feedback is provided by an external evaluator, which in the repeated-feedback regime effectively plays the role of a *teacher*. This setting isolates the effects of endogenous delegation under AI assistance. We focus on *error-driven* learners [25, 9], whose updates, covering both skill and delegation, are governed by observed performance loss. Accordingly, we begin by specifying the performance loss that drives the learning dynamics.

Skill representation. We represent task-specific ability as a latent scalar parameter that governs performance quality on the task. This parameter summarizes task-specific performance relative to the target. For analytical convenience, we represent skill as a scalar $\theta \in [0, 1]$, where $\theta = 1$ corresponds to optimal task performance. This scalar representation provides a minimal summary of task-specific performance and can be viewed as a projection of richer latent skill spaces often modeled geometrically in machine learning. Such geometric abstractions are common in settings such as semi-supervised learning and deep generative models [26, 14, 28]. Let $\theta(t) \in [0, 1]$ denote the learner's skill at time t. Let $\theta_a \in [0, 1]$ denote the AI's *effective skill*, defined as the expected performance of the AI under the task loss. Importantly, this parameter does not assume that AI performance is stable or deterministic. In practice, AI systems are often jagged, producing highly accurate outputs on many instances while occasionally making severe errors [6]. As shown in Section 4, such variability admits a performance-equivalent deterministic representation under the expected loss.

Performance loss. Task performance is evaluated by an external evaluator against a fixed target. We represent task-specific ability by a scalar skill parameter $\theta \in [0, 1]$, where $\theta = 1$ corresponds to optimal task performance. Let $\theta(t)$ denote the learner's skill at time t, and let θ_a denote the AI's effective skill, defined as its expected performance under the task loss. Let $X(t) \in \{0, 1\}$ be the delegation indicator, where $X(t) = 1$ corresponds to delegating the task to AI. We define the instantaneous performance loss as $\ell(\theta(t), X(t)) := (1 - X(t))(1 - \theta(t))^2 + X(t)(1 - \theta_a)^2$. Let $p(t) \in [0, 1]$ denote the delegation level, i.e., $\Pr[X(t) = 1] = p(t)$. The expected instantaneous performance loss is

$$\ell(\theta(t), p(t)) = (1 - p(t))(1 - \theta(t))^2 + p(t)(1 - \theta_a)^2. \tag{1}$$

When $(1-\theta(t))^2 \geq (1-\theta_a)^2$, increasing delegation strictly improves short-run performance, creating an incentive to rely on AI.

Skill update. The learner's skill $\theta(t) \in [0,1]$ evolves in response to observed performance loss and decays under non-use. To model bounded skill dynamics, we adopt a standard multiplicative-update metric, $F(\theta) = \frac{1}{\theta(1-\theta)}$, which induces updates proportional to $\theta(1-\theta)$ and is widely used in learning and evolutionary dynamics [18]. Conditioned on the realized delegation decision $X(t)$, minimizing the instantaneous loss $\ell(\theta(t), X(t))$ induces the learning drift

$$g_t := F(\theta(t))^{-1}\, \partial_\theta \ell(\theta(t), X(t)),$$

corresponding to replicator-style (MWU) dynamics. When the task is delegated to AI ($X(t) = 1$), the learner does not practice and skill decays toward a default level $\theta_d \in [0,1]$ [10, 25]. We model this as a decay drift

$$d_t := X(t)\, F(\theta(t))^{-1}\, \partial_\theta (\theta(t) - \theta_d)^2.$$

Combining learning and decay yields the stochastic update

$$\theta(t+1) - \theta(t) = -\eta\,(g_t + \Delta d_t)\,,$$

where $\eta > 0$ is the learning rate and $\Delta > 0$ controls the relative strength of decay.

Taking expectations and noting $\Pr[X(t) = 1] = p(t)$ gives

$$\mathbb{E}[\theta(t+1) - \theta(t) \mid \theta(t)] = 2\eta\theta(t)(1-\theta(t))\left((1-p(t))(1-\theta(t)) + \Delta p(t)(\theta_d - \theta(t))\right).$$

As delegation increases, the expected skill update transitions from improvement toward the task optimum to decay toward the default level.

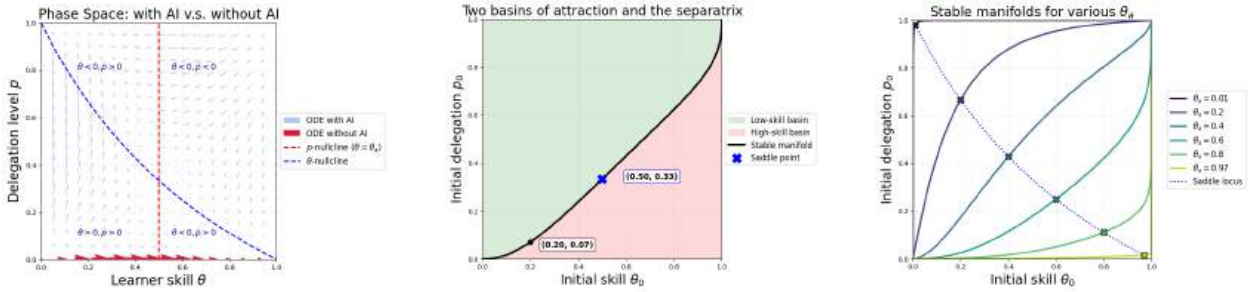
Delegation update. We assume that the learner can estimate its own performance loss and the AI-induced loss under the task objective. Accordingly, the delegation level $p(t) \in [0,1]$ adapts to minimize the expected performance loss $\ell(\theta(t), p(t))$. Since $p(t)$ is a probability, we model its evolution using a multiplicative update. Taking a gradient step on $\ell(\theta(t), p(t))$ yields

$$p(t+1) - p(t) := -\kappa \cdot 2\eta F(p(t))^{-1} \partial_p \ell(\theta(t), p(t)) = \kappa \cdot 2\eta\, p(t)(1-p(t))\left((1-\theta(t))^2 - (1-\theta_a)^2\right),$$

where $\kappa > 0$ controls the relative adaptation rate of delegation. When the AI outperforms the learner, i.e., $(1-\theta(t))^2 \geq (1-\theta_a)^2$, delegation increases; otherwise it decreases. This captures the empirically observed tendency of lower-skill users to rely more heavily on AI assistance [3, 30].

Dynamics. The update rules above define a discrete-time stochastic process. Following standard stochastic-approximation arguments [4], we show that, as the step size $\eta \to 0$, the expected trajectories of this process are well approximated by the following system of ordinary differential equations:

$$\begin{aligned} \dot{\theta} &= \theta(1-\theta)((1-p)(1-\theta) + \Delta p(\theta_d - \theta)), \\ \dot{p} &= \kappa\, p(1-p)((1-\theta)^2 - (1-\theta_a)^2). \end{aligned} \tag{2}$$



(a) Comparison of (3) and (4) (b) Illustration of Theorem 3.2 (c) Effects of θ_a

Figure 1: Plots illustrating the outcomes of ODE (3) and effects of AI skill, with default settings of $(\theta_a, \kappa, \Delta) = (0.5, 3, 2)$.

A complete derivation and an analysis of the stochastic trajectories are provided in Section 5. Importantly, both θ and p evolve to locally minimize the same instantaneous performance loss, so the system exhibits no incentive misalignment. Despite this local alignment, the resulting global dynamics differ qualitatively from human-only learning, giving rise to multiple equilibria and path-dependent outcomes. Throughout the analysis, we compare the long-run behavior of (2) across different initial conditions (θ_0, p_0) and against the no-AI baseline (4). Unless otherwise stated, we consider the regime $\theta_a, \theta_0 \in [\theta_d, 1]$. Without loss of generality, we normalize $\theta_d = 0$, yielding the simplified system

$$\begin{aligned} \dot{\theta} &= \theta(1-\theta)\left((1-p)(1-\theta) - \Delta p\,\theta\right), \\ \dot{p} &= \kappa\, p(1-p)\left((1-\theta)^2 - (1-\theta_a)^2\right). \end{aligned} \tag{3}$$

Analysis of the general case is deferred to Section 5.3.

Comparison with no-AI case. When the initial delegation level is $p_0 = 0$, we have $p(t) \equiv 0$, yielding a one-dimensional skill dynamics

$$\dot{\theta} = \theta(1-\theta)(1-\theta), \tag{4}$$

which corresponds to learning without AI assistance. This ODE converges to a unique globally stable equilibrium at $\theta = 1$, consistent with classical learning models [19, 4]. This agreement serves as a sanity check for (3).

Figure 1a visualizes the phase spaces with and without AI assistance. The introduction of adaptive delegation substantially alters the dynamics, partitioning the state space $[0, 1]^2$ into multiple behavioral regions separated by the nullclines $\dot{\theta} = 0$ and $\dot{p} = 0$.

Remark 2.1 (**Usability of the model**). We provide an explicit procedure for estimating the parameters θ_a, κ, and Δ from data in Section 8, together with a worked numerical example.

3 Theoretical results

We analyze the qualitative long-run behavior of the learning dynamics (3), with a focus on how adaptive delegation reshapes learning outcomes. Our results proceed in three steps: (i) we characterize the fixed points introduced by AI assistance, (ii) we show that the resulting saddle structure

partitions the state space into two basins of attraction, and (iii) we quantify the resulting tradeoff between short-run performance gains and long-run losses.

Throughout, we study the flow of ODE (3) on the interior $(0,1)^2$. A standard stochastic-approximation argument implies that the discrete-time process from Section 2 tracks this ODE in the small-step regime. Basic well-posedness facts (forward invariance and global existence) are stated and proved in Section 5.3.

We begin by comparing (3) with the no-AI baseline (4). Adaptive delegation changes the fixed-point structure: in addition to the high-skill equilibrium, the system admits a stable low-skill equilibrium corresponding to persistent delegation.

Theorem 3.1 (**Equilibria of ODE** (3))**.** *Let θ_a, κ, Δ be the parameters of ODE* (3)*. When $\theta_a \in (0,1)$, the ODE system exhibits two* stable *nodes (sinks): a high-skill equilibrium at $(\theta^\infty, p^\infty) = (1,0)$ and a low-skill equilibrium $(\theta^\infty, p^\infty) = (0,1)$; two unstable nodes (sources) at $(0,0)$ and $(1,1)$; and an additional* interior saddle point $(\theta^\dagger, p^\dagger) = (\theta_a, \frac{1-\theta_a}{1-(1-\Delta)\theta_a})$.

The key observation is that AI assistance introduces an additional stable low-skill equilibrium $(0,1)$, in contrast to the no-AI baseline (4), which admits a unique stable high-skill equilibrium $\theta^\infty = 1$. The key qualitative change is the emergence of an additional stable equilibrium at $(0,1)$, which has no analog in the no-AI baseline (4). The interior saddle point mediates this shift, producing a phase portrait with competing long-run outcomes. This provides a dynamical explanation for heterogeneous effects of AI assistance across learners reported in recent empirical studies [13, 16].

As $\theta_a \to 0$, the saddle $(\theta^\dagger, p^\dagger)$ approaches the low-skill equilibrium $(0,1)$; as $\theta_a \to 1$, it approaches $(1,0)$. Thus, AI quality directly shifts the global geometry of learning.

The remaining question is which initial conditions converge to the low-skill equilibrium. This is determined not by local improvement, but by which side of the saddle's stable manifold the learner starts on.

Theorem 3.2 (**Two basins divided by saddle point**)**.** *Let θ_a, κ, Δ be the parameters of ODE* (3)*. Let $(p^\dagger, \theta^\dagger)$ be the interior saddle point guaranteed by Theorem 3.1. There exists a monotonically increasing 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 state space $[0,1]^2$ into two distinct 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 $(1,0)$.*

Note that the stable manifold of the saddle point serves as a boundary that partitions the state space into exactly two basins of attraction. In particular, a slight increase in the initial delegation level p_0 can shift the system from $p_0 < \psi(\theta_0)$ to $p_0 > \psi(\theta_0)$, causing the learner to converge to the low-skill equilibrium. Thus, adaptive delegation creates a feedback loop: increased delegation reduces practice, which lowers skill and in turn increases future delegation through performance-driven updates. Although each update is locally aligned with the same performance objective, their interaction alters the global stability structure. In particular, the low-skill outcome is not explained by decay alone, which is reversible when $p \equiv 0$, but by the coupling between decay and adaptive delegation. Figure 1b visualizes the stable manifold $\psi(\cdot)$ and the two basins of attraction.

The monotonicity of ψ with respect to θ indicates that AI delegation is more risky for early-stage learners: even small initial levels of delegation can place them in the low-skill basin, from which recovery requires sustained reductions in delegation. For example, when $\theta_0 = 0.2$, a slight initial delegation with $p_0 > 0.07$ pushes the learner into the low-skill basin. Moreover, since the stable manifold terminates at $(1, 1)$, we have $\psi(\theta) < 1$ for any $\theta < 1$. This implies that even highly skilled learners may face the risk of skill degradation under sufficiently heavy reliance on AI.

To enable actionable predictions of whether a learner lies in the low-skill basin, we propose a piecewise-polynomial approximation $\widetilde{\psi}(\cdot)$ of the basin boundary; see Equation (16) and Figure 4a in Section 6.6.

Effects of AI quality on the basin boundary. We next ask whether a better AI can reduce the risk of skill degradation for learners. To address this, we propose the following theorem that characterize how the basin boundary ψ varies with AI skill θ_a.

Theorem 3.3 (**Effects of θ_a on ψ**). *For all $\theta \in (0, 1)$, $\psi(\theta)$ is monotonically decreasing in θ_a and continuously differentiable with respect to θ_a. Moreover, for any $(\theta_0, p_0) \in (0, 1)^2$, there exists $\theta_a \in (0, 1)$ such that $p_0 = \psi_{\theta_a}(\theta_0)$.*

Theorem 3.3 implies that increasing AI skill shifts the basin boundary downward, expanding the set of initial conditions that converge to the low-skill equilibrium. Equivalently, as θ_a grows, the threshold of "safe" delegation becomes more stringent.

Figure 1c illustrates this dependence. For low AI quality (e.g., $\theta_a = 0.01$), the low-skill basin is small. For very high AI quality (e.g., $\theta_a = 0.99$), the high-skill basin is small, corresponding to near-complete task replacement.

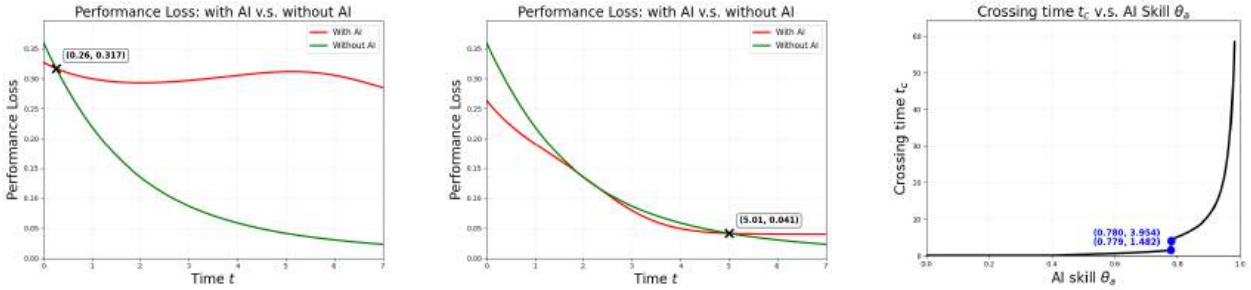
As θ_a increases from 0 to 1, the interior saddle point moves continuously, approaching $(0, 1)$ as $\theta_a \to 0$ and $(1, 0)$ as $\theta_a \to 1$. This motion induces a shifting threshold of safe delegation: beyond the boundary $p = \psi(\theta)$, trajectories are drawn into the low-skill basin.

A key implication is that no learner is inherently safe as AI capability increases: for sufficiently large θ_a, any interior initial condition may fall into the low-skill basin. Sustained practice therefore remains necessary even when AI systems are highly capable. Providing early-stage learners with access to near-solution-level assistance can be particularly risky, since it increases delegation precisely when the learner's skill is still forming.

These predictions are consistent with recent evidence [32, 3]. In particular, settings closer to "feedback"-style assistance tend to show weaker learning losses than settings closer to "solution"-style assistance, which aligns with our model's dependence on effective AI skill.

A further implication is that the timing of access matters: introducing AI after partial skill acquisition can lead to different outcomes than introducing it from the outset, even with the same total exposure. This non-commutativity is evidenced in [13], where participants who trained without LLMs and were later granted access (termed as "Brain-to-LLM") outperformed those who received immediate access, both in quotation accuracy and in self-reported essay ownership. These findings support a dynamic view of delegation, in which AI tools interact with the learner's evolving state rather than acting as context-free substitutes for effort. In addition, we study how the basin boundary varies with other parameters, namely κ and Δ, thereby validating the robustness of the convergence behavior under AI assistance. See Section 6.7 for the corresponding results.

We next translate these asymptotic predictions about skill into consequences for task performance over time.



(a) Loss with and without AI (b) Varying $\theta_a = 0.8$ (c) Crossing time v.s. θ_a

Figure 2: Plots illustrating the instantaneous performance loss across time with and with AI, and how the crossing point varies with respect to AI skill θ_a, with default settings of $(\theta_a, \kappa, \Delta, \theta_0, p_0) = (0.5, 3, 2, 0.4, 0.3)$.

3.1 Analyzing performance loss across time

Although AI assistance may lead to long-run skill degradation, it can reduce instantaneous performance loss in the short run. This effect is particularly pronounced for early-stage learners with $\theta_0 < \theta_a$, for whom the instantaneous loss $\ell(\theta_0, p_0)$ is monotonically decreasing in the delegation level p_0. As discussed in Section 1, related experiments find that AI assistance improves immediate performance, but that this advantage may diminish or even reverse as interaction time increases [3, 13]. For example, in essay-writing tasks, the average accuracy of the LLM-assisted group is consistently higher than that of the brain-only group in the first section, while the brain-only group achieves higher average accuracy by the third section; see Figure 41 in [13]. These observations raise a natural question: can the performance gains from AI assistance persist over time?

To answer this question, we first define the difference in instantaneous performance loss over time between the AI-assisted and no-AI settings. 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. In particular, $\theta(t; \theta_0, 0)$ corresponds to the learner's skill at time t in the absence of AI, yielding a performance loss of $(1 - \theta(t; \theta_0, 0))^2$. We define the *performance gap* at time t as

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

which is the difference in instantaneous performance loss at time t between an AI-assisted learner with initial state (θ_0, p_0) and a no-AI learner with initial skill θ_0. A negative value $G_\ell(t) < 0$ means that delegation yields lower loss at time t than learning without AI. We ask whether $G_\ell(t) < 0$ can hold for all $t \geq 0$.

Theorem 3.4 (Short-term gains, long-term losses). *Let θ_a, κ, Δ be the parameters of ODE (3). Given an initial state $(\theta_0, p_0) \in (0, 1)^2$ with $\theta_0 < \theta_a$, let $t_c := \inf\{t \geq 0 : \forall t' > t,\ G_\ell(t') > 0\}$ denote the crossing time after which AI delegation incurs higher loss than the no-AI baseline. Then $t_c < \infty$. Moreover, $t_c = 0$ when $\theta_0 \geq \theta_a$ and $\theta(t_c; \theta_0, 0) \leq \theta_a$ when $\theta_0 < \theta_a$.*

This result shows that AI assistance yields a short-term performance gain but leads to persistent performance loss after a crossing time t_c; see Figure 2a for an illustration. The crossing time t_c reflects a balance between two competing effects: (1) a performance gain induced by delegating to an AI with higher skill than the learner, i.e., $\theta_a > \theta(t; \theta_0, p_0)$ for $t < t_c$, and (2) a performance loss

induced by skill degradation due to delegation, i.e., $\theta(t;\theta_0,p_0) \le \theta(t;\theta_0,0)$ for all $t \ge 0$, as guaranteed by Lemma 3.5. Notably, the long-run loss reflects both decay under non-use and foregone learning opportunities during delegation, effects that are not internalized under myopic, performance-driven updates. This mechanism yields what is often called the *cognitive debt* phenomenon [13]: short-term performance gains from AI delegation are eventually outweighed by structurally foregone practice under adaptive delegation.

Effects of AI quality on the performance crossing time. We next study how the crossing time t_c varies with AI skill θ_a. This question is of practical interest to learners, for example, those who aim to ensure that the period of performance gains covers the duration of a course when grades are prioritized over skill acquisition.

To address this question, we visualize how the crossing time t_c varies with θ_a in Figure 2c. The plot shows that the duration of performance gains can be substantially prolonged as AI quality improves. When $\theta_a \le \theta_0 = 0.4$, we have $t_c = 0$, since the no-AI learner consistently outperforms the AI-assisted learner in this regime. In contrast, when θ_a increases to a high capability level (e.g., $\theta_a = 0.78$), in the plotted example, the crossing time can increase sharply from $t_c = 1.482$ to $t_c = 3.954$. This occurs because multiple solutions to $G_\ell(t) = 0$ appear, as illustrated in Figure 2b. Intuitively, a learner without AI requires substantially more time to surpass a highly capable AI through independent learning, thereby extending the period during which AI assistance yields a performance advantage.

Higher AI capability can dramatically extend the window of short-run gains while expanding the low-skill basin (Theorem 3.3). As a result, improvements in AI quality can increase the misalignment between optimizing short-term performance and optimizing long-term skill. This tension underscores the risks of incorporating highly capable AI into learning processes and the need to redesign learner incentives and reward structures.

3.2 Overview of the proofs

We summarize the key proof ideas underlying each main theorem; full proofs are deferred to Section 6.

To Theorem 3.1. The first step is to solve equations $\dot\theta = \dot p = 0$, yielding four corner equilibria in $\{0,1\}^2$ and a unique interior equilibrium $(\theta^\dagger, p^\dagger)$, located at the intersection of the two interior nullclines $(1-p)(1-\theta) - \Delta p\theta = 0$ and $(1-\theta)^2 - (1-\theta_a)^2 = 0$. To analyze stability, we study the eigenvalues of the Jacobian matrix $J = \begin{bmatrix} \partial_\theta\dot\theta & \partial_p\dot\theta \\ \partial_\theta\dot p & \partial_p\dot p \end{bmatrix}$.

For a corner equilibrium $(\theta^\infty, p^\infty) \in \{0,1\}^2$, the Jacobian is diagonal: $J = \mathrm{diag}(\partial_\theta\dot\theta, \partial_p\dot p)$, where $\partial_\theta\dot\theta = (1-2\theta^\infty)\left((1-p^\infty)(1-\theta^\infty) - \Delta p^\infty\theta^\infty\right)$, and $\partial_p\dot p = \kappa(1-2p^\infty)\left((1-\theta^\infty)^2 - (1-\theta_a)^2\right)$. When $\theta^\infty = p^\infty$, both $\partial_\theta\dot\theta$ and $\partial_p\dot p$ are positive, so both eigenvalues of J are positive. Thus, $(0,0)$ and $(1,1)$ are unstable sources. In contrast, when $\theta^\infty \ne p^\infty$, one eigenvalue is zero and the other one is negative, so the corresponding corner equilibria are (non-hyperbolic) stable sinks.

Finally, for the interior equilibrium $(\theta^\dagger, p^\dagger)$, the Jacobian takes the form $J = \begin{bmatrix} \partial_\theta\dot\theta & \partial_p\dot\theta \\ \partial_\theta\dot p & 0 \end{bmatrix}$, where $\partial_p\dot\theta = -\frac{\theta_a^2(1-\theta_a)}{p^\dagger} < 0$, $\partial_\theta\dot p = -2\kappa p^\dagger(1-p^\dagger)\theta_a(1-\theta_a) < 0$. Thus, the Jacobian has one positive and one negative eigenvalue, and $(\theta^\dagger, p^\dagger)$ is a saddle point.

To Theorem 3.2 and Theorem 3.3. The key is to understand why higher initial delegation p_0, lower initial skill θ_0, and greater AI skill θ_a increase susceptibility to long-run skill degradation, shaping the low-skill basin in Theorem 3.2 and its deformation with θ_a (Theorem 3.3). To address this question, we establish the following lemma, which identifies a negative coupling between skill and delegation. This lemma is the main monotonicity principle driving both the basin geometry and the performance results.

Lemma 3.5 (**Negative coupling between skill and delegation**). *Let θ_a, κ, and Δ be the parameters of ODE* (3)*. For any $t > 0$, $\theta(t)$ is non-increasing and $p(t)$ is non-decreasing as functions of each of the following variables: the initial delegation p_0, the initial skill gap $1 - \theta_0$, and θ_a.*

As $t \to \infty$, the lemma implies that the limiting skill level θ^∞ is non-increasing in the initial delegation p_0, the initial skill gap $1 - \theta_0$, and the AI skill θ_a, supporting the monotonic basin structure in our theory. Technically, we exploit the system's monotonicity to lift local saddle stability to a global partition of the phase space.

Beyond its technical role, Lemma 3.5 implies that among learners with identical initial skills, the one with a higher delegation level will have lower skill at any point in time. This provides a theoretical explanation for the negative correlation between delegation level and human understanding, coding proficiency, and cognitive load reported in [1, 8], and predicts that this effect strengthens as AI quality improves.

To Theorem 3.4. The key insight is that once the no-AI learner's skill reaches $\theta(t; \theta_0, 0) \geq \theta_a$, it will thereafter consistently outperform the AI-assisted learner. This follows from Lemma 3.5, which guarantees that $\theta(t; \theta_0, p_0) < \theta(t; \theta_0, 0)$ for all $t > 0$, and from the definition of t_c as the first time at which $\theta(t; \theta_0, 0)$ reaches θ_a. Together, these observations yield the properties of the crossing time t_c established in Theorem 3.4.

4 Model extensions and robustness

Recall that ODE (3) is a minimal dynamical system introduced for analytical tractability. This raises a natural question: does the qualitative phase structure persist in more general settings? To address this, we consider several extensions of the system and examine how they reshape the basins of attraction, thereby validating the robustness of our analytical results. Key results are summarized below, with full derivations and figures in Section 7.

Jagged AI and effective skill. Recall that AI performance is often jagged across instances in practice [6]. Let $s \in [0, 1]$ denote the random per-instance skill of a jagged AI and let μ_a denote a distribution over AI skill on $[0, 1]$. For instance, one may consider a two-point case: $s = 1$ with probability q and $s = \theta_\ell$ with probability $1 - q$. When q is close to 1, this represents a highly capable but occasionally failing AI.

Consequently, the update of the delegation level becomes:

$$\dot{p} = \kappa p(1 - p) \left((1 - \theta)^2 - \mathbb{E}_{s \sim \mu_a}[(1 - s)^2] \right). \tag{5}$$

To analyze the effect of the jagged AI on basins, we define the *effective skill* $\widehat{\theta}_a$ as the deterministic skill that matches expected squared task error: $(1 - \widehat{\theta}_a)^2 = \mathbb{E}\big[(1 - s)^2\big]$. This definition ensures that

a jagged AI and a deterministic AI with ability $\widehat{\theta}_a$ are equivalent under the performance metric governing delegation decisions in our model. In the two-point case, this yields

$$(1-\widehat{\theta}_a)^2 = (1-q)(1-\theta_\ell)^2,\ \widehat{\theta}_a = 1-\sqrt{1-q}\,(1-\theta_\ell).$$

As a result, the phase structure and basin geometry derived in Section 3 remain unchanged under jagged AI. Since $\mathbb{E}_{s\sim\mu_a}\left[(1-s)^2\right] \geq (1-\mathbb{E}_{s\sim\mu_a}[s])^2$, randomness in AI skill increases the effective expected loss of delegation relative to a deterministic AI with the same mean skill, enlarging the high-skill basin.

Noisy update of delegations. ODE (3) assumes accurate beliefs about the AI-induced performance loss $(1-\theta_a)^2$ when updating delegation, which may not hold in practice. For instance, humans may overestimate AI skill due to surface fluency [6]. Let $\widetilde{\theta}_a \in [0,1]$ denote the skill level that the learner believes the AI to have. This misperception affects the delegation update through the loss term used when evaluating delegation decisions, leading to

$$\dot{p} = \kappa p(1-p)\big((1-p)\left((1-\theta)^2-(1-\widetilde{\theta}_a)^2\right) + p\left((1-\theta)^2-(1-\theta_a)^2\right)\big), \tag{6}$$

where the learner evaluates delegation using the perceived loss $(1-\widetilde{\theta}_a)^2$, while realized outcomes depend on the true loss $(1-\theta_a)^2$. Then overestimating AI skill (i.e., $\widetilde{\theta}_a > \theta_a$) increases reliance on AI, enlarging the low-skill basin.

Asymmetric update of delegations. ODE (2) assumes a uniform update rate κ for the delegation level, regardless of whether the learner outperforms the AI (i.e., $(1-\theta)^2 < (1-\theta_a)^2$) or vice versa. This corresponds to a risk-neutral learner, whereas learners in practice often exhibit asymmetric responses to positive and negative evidence [33, 5]. For example, a learner may reduce delegation more aggressively upon realizing that the AI performs worse.

To capture this behavior, we introduce a risk-aversion factor $\alpha \geq 0$ that controls the update rate when the learner outperforms the AI. Let $[x]_+ := \max\{x, 0\}$ for $x \in \mathbb{R}$, and let $f(p) := \kappa p(1-p)\left((1-\theta)^2-(1-\theta_a)^2\right)$ denote the default delegation update in ODE (3). We then derive the following asymmetric update rule for the delegation level:

$$\dot{p} = [f(p)]_+ - \alpha\,[-f(p)]_+. \tag{7}$$

When $\alpha > 1$, negative evidence about AI performance is weighted more heavily than positive evidence, causing delegation to decrease faster once the learner outperforms the AI. Geometrically, this asymmetry shifts the stable manifold outward, enlarging the high-skill basin. Consequently, this extension can be interpreted as an intervention policy for senior learners with $\theta_0 > \theta_a$, encouraging vigilance against over-reliance on AI without restricting access outright.

In addition, we consider an extension with an alternative performance loss in Section 7.4. These extensions show that multiple equilibria, basin boundaries, and long-run degradation are not artifacts of particular modeling choices, but structurally stable consequences of adaptive AI delegation.

5 Omitted details for ODE (2) from Section 2

In this section, we first derive a dynamics from the learning procedure, and then show that the dynamics tracks to ODE (2). Finally, we prove some basic properties of ODE (2), which are omitted in the main body. Table 1 summarizes notations used in this paper.

Table 1: Used notations

Symbol	Domain	Description
1	$[0,1]$	Target skill
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	$\mathbb{R}_{\geq 0}$	Decay rate
κ	$\mathbb{R}_{\geq 0}$	Delegation 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 levels p
$\psi(\theta)$	$[0,1] \to [0,1]$	Mapping for the stable manifold of saddle point from θ to a delegation level
$G_\ell(t)$	$\mathbb{R}_{\geq 0} \to \mathbb{R}$	Performance gap between with and without AI at time t
t_c	$\mathbb{R}_{\geq 0}$	Maximum crossing time for performance loss

5.1 Deriving dynamics under AI assistance

Let $X(t) \in \{0,1\}$ denote the action of the learner at round t, where $X(t)=0$ represents that the learner does the task by itself and $X(t)=1$ represents that the learner delegates the task to AI. Let $p(t) \in [0,1]$ denote the probability of $X(t)=1$ at round t. Let $\theta(t) \in$ denote the skill of the learner at round t. Recall that $\theta_d, \theta_a \in [0,1]$ denote the default skill and AI skill, respectively.

The mechanism of learning with AI assistance is: At round t,

- **(Task completion stage)** Learner delegates the task to AI with probability $p(t)$, which sets $X(t)=1$; and does the task by itself with probability $1-p(t)$, which sets $X(t)=0$.
- **(Evaluation stage)** Learner submits the output to the teacher. The teacher gives a loss ℓ_t to the learner, where
$$\ell_t := (1-X(t))(1-\theta(t))^2 + X(t)\cdot(1-\theta_a)^2.$$
- **(Update of skill)** If $X(t)=0$, the learner skill (intrinsically) updates towards the target 1 according to the loss $\ell_t = (1-\theta(t))^2$:
$$\theta(t+1) := \theta(t) + 2\eta\theta(t)(1-\theta(t))^2$$
Otherwise, if $X(t)=1$, the learner skill decays towards θ_d due to non-practice:
$$\theta(t+1) := \theta(t) + 2\eta\Delta\theta(t)(1-\theta(t))(\theta_d-\theta(t)).$$
- **(Update of delegation level)** If $X(t)=0$, assume that the learner has an accurate belief $(1-\theta_a)^2$ on the loss of AI's output. Otherwise, if $X(t)=1$, assume that the learner has an accurate belief $(1-\theta(t))^2$ on the loss of its output. Together with the obtained loss ℓ_t, the learner updates its delegation level as follows:
$$p(t+1) := p(t) + \kappa\cdot 2\eta p(t)(1-p(t))\left((1-\theta(t))^2 - (1-\theta_a)^2\right).$$

In summary, we derive the following stochastic dynamics under AI assistance:

$$\begin{aligned} X(t) &\sim \text{Bern}(p(t)), \\ \theta(t+1) &= \theta(t) + 2\eta\theta(t)(1-\theta(t))[(1-X(t))(1-\theta(t)) + \Delta \cdot X(t) \cdot (\theta_d - \theta(t))] \\ p(t+1) &= p(t) + \kappa \cdot 2\eta p(t)(1-p(t))[(1-\theta(t))^2 - (1-\theta_a)^2]. \end{aligned} \tag{8}$$

5.2 Dynamics (8) tracks to ODE (2)

We show that as the step size $\eta \to 0$, the stochastic trajectory of Dynamics (8) converges to the solution of the ODE (2). We first derive the expected dynamics of the discrete system.

Lemma 5.1 (**Expected discrete dynamics).** *The expected evolution of the stochastic system* (8)*, conditional on the current state* $(\theta(t), p(t))$*, is given by the deterministic map:*

$$\begin{aligned} \mathbb{E}[\theta(t+1) \mid \mathcal{F}_t] &= \theta(t) + 2\eta\theta(t)(1-\theta(t))[(1-p(t))(1-\theta(t)) + \Delta \cdot p(t) \cdot (\theta_d - \theta(t))], \\ p(t+1) &= p(t) + \kappa \cdot 2\eta p(t)(1-p(t))[(1-\theta(t))^2 - (1-\theta_a)^2]. \end{aligned} \tag{9}$$

Proof. Let $\mathcal{F}_t$ denote the filtration generated by the history up to time t. Note that the update for $p(t+1)$ is deterministic given $\mathcal{F}_t$. The update for $\theta(t+1)$ depends on the random variable $X(t)$. Since $X(t) \sim \text{Bern}(p(t))$, we have $\mathbb{E}[X(t) \mid \mathcal{F}_t] = p(t)$. Because the update rule for $\theta(t+1)$ is affine in $X(t)$, we can compute the conditional expectation by substituting $X(t)$ with its mean $p(t)$:

$$\begin{aligned} \mathbb{E}[\theta(t+1) \mid \mathcal{F}_t] &= \theta(t) + 2\eta\theta(t)(1-\theta(t))[(1-\mathbb{E}[X(t)])(1-\theta(t)) + \Delta\mathbb{E}[X(t)](\theta_d - \theta(t))] \\ &= \theta(t) + 2\eta\theta(t)(1-\theta(t))[(1-p(t))(1-\theta(t)) + \Delta p(t)(\theta_d - \theta(t))]. \end{aligned}$$

This yields the system (9). □

By standard results in stochastic approximation [4, 12], as $\eta \to 0$, the trajectory of the stochastic process (8) concentrates around the solution of the ODE defined by the drift fields in (9), which is exactly ODE (2).

5.3 ODE (2) is well-posed and converges

We show that ODE (2) admits a solution from any initial condition and has no cycles, providing a foundation for the analysis in Section 3.

Theorem 5.2 (**Existence of a solution and convergence).** *Let* θ_a, κ, Δ *be the parameters of ODE* (3)*. Then for any initial state* $(\theta_0, p_0) \in [0,1]^2$*, the following properties hold:*

- *ODE* (3) *has a unique solution* $(p(t), \theta(t)) \in [0,1] \times (0,1)$ *defined for all* $t \geq 0$*;*
- *There always exists* $(\theta^\infty, p^\infty) \in [0,1] \times (0,1)$ *such that* $\lim_{t\to\infty}(p(t), \theta(t)) = (p^\infty, \theta^\infty)$*.*

Proof. We write $x = (\theta, p) \in \mathbb{R}^2$ and denote by $F : \mathbb{R}^2 \to \mathbb{R}^2$ the vector field given by the right-hand side of the system, so that the ODE can be written concisely as

$$\dot{x} = F(x).$$

Local existence and uniqueness. Each component of F is a polynomial in (θ, p), obtained as a product of factors of the form θ, $(1-\theta)$, $(\theta_d - \theta)$, p, $(1-p)$ and fixed parameters in $[0,1]$. In particular, F is C^∞ on $\mathbb{R}^2$, hence locally Lipschitz.

By the Picard–Lindelöf theorem for autonomous systems in $\mathbb{R}^2$, for every initial condition $x^0 \in \mathbb{R}^2$ there exists a unique *maximal* solution $x : [0, T_{\max}) \to \mathbb{R}^2$ with $T_{\max} > 0$ such that $x(0) = x^0$ and $\dot{x}(t) = F(x(t))$ for all $t \in [0, T_{\max})$.

Lemma 5.3 (Logistic barrier lemma). *Let $g : [0, T_{\max}) \to \mathbb{R}$ be continuous, and consider the scalar nonautonomous ODE*

$$\dot{y}(t) \;=\; y(t)\big(1 - y(t)\big)\, g(t), \qquad t \in [0, T_{\max}).$$

Suppose $y : [0, T_{\max}) \to \mathbb{R}$ is a C^1 solution with $y(0) \in [0,1]$. Then

$$y(t) \in [0,1] \quad \textit{for all } t \in [0, T_{\max}).$$

Proof. Define

$$S := \{t \in [0, T_{\max}) : y(s) \in [0,1] \text{ for all } s \in [0,t]\},$$

and let $\tau := \sup S$. Since $y(0) \in [0,1]$ and y is continuous, we have $0 \in S$ so $S \neq \emptyset$ and $\tau \in (0, T_{\max}]$. By continuity of y, we have $y(t) \in [0,1]$ for all $t \in [0, \tau]$ and $y(\tau) \in [0,1]$. We claim that $\tau = T_{\max}$. Suppose, for contradiction, that $\tau < T_{\max}$.

Case 1: $y(\tau) \in (0,1)$. Then, by continuity, there exists $\varepsilon > 0$ such that the solution is defined on $[\tau, \tau + \varepsilon]$ and $y(t) \in (0,1) \subset [0,1]$ for all $t \in [\tau, \tau + \varepsilon]$. Thus $y(t) \in [0,1]$ for all $t \in [0, \tau + \varepsilon]$, contradicting the definition of τ as the supremum of S.

Case 2: $y(\tau) \in \{0,1\}$. Without loss of generality assume $y(\tau) = 0$; the case $y(\tau) = 1$ is identical. Consider the scalar ODE

$$\dot{z}(t) = f(t, z(t)), \qquad f(t,z) := z(1-z)g(t),$$

with initial condition $z(\tau) = 0$. The map f is continuous in (t,z) and, for each fixed t, it is a polynomial in z, hence locally Lipschitz in z on $\mathbb{R}$. By the Picard–Lindelöf theorem for nonautonomous scalar ODEs, there exists a unique solution $z(t)$ in a neighborhood of τ with $z(\tau) = 0$.

Note that the constant function $z(t) \equiv 0$ is a solution, since $f(t,0) = 0$ for all t. Our original function $y(t)$ also solves the same ODE and satisfies $y(\tau) = 0$. By uniqueness of solutions through $(\tau, 0)$, we must have $y(t) \equiv 0$ on some interval $[\tau, \tau + \varepsilon]$ for $\varepsilon > 0$. In particular, $y(t) \in [0,1]$ for all $t \in [0, \tau + \varepsilon]$, again contradicting the definition of τ.

We have reached a contradiction in both cases, so our assumption $\tau < T_{\max}$ was false. Hence $\tau = T_{\max}$ and therefore $y(t) \in [0,1]$ for all $t \in [0, T_{\max})$. □

Invariance of $[0,1]$ for each coordinate. Each coordinate of $x(t)$ can be rewritten in the form

$$\dot{y}(t) = y(t)(1 - y(t))\, g(t),$$

with $g(t)$ continuous. Thus each coordinate satisfies the hypotheses of Lemma 5.3. Since the initial condition lies in $[0,1]^2$, we conclude

$$\theta(t), p(t) \in [0,1] \quad \text{for all } t \in [0, T_{\max}).$$

Hence $[0,1]^2$ is positively invariant.

Since $f(t,y) = y(1-y)g(t)$ is a polynomial in y, it is locally Lipschitz in y, uniformly on compact sets, ensuring the uniqueness needed in Lemma 5.3.

Global existence ($T_{\max} = +\infty$). We now upgrade local solutions to global ones. By Step 2, if $x(0) \in [0,1]^2$ then $x(t) \in [0,1]^2$ for all $t \in [0, T_{\max})$. The set $[0,1]^2$ is compact, and F is continuous, hence

$$M := \sup_{x \in [0,1]^2} \|F(x)\| < \infty.$$

Therefore any solution starting in $[0,1]^2$ is bounded and has bounded derivative:

$$\|x(t)\| \le \sqrt{2} \quad \text{and} \quad \|\dot{x}(t)\| = \|F(x(t))\| \le M \quad \text{for all } t \in [0, T_{\max}).$$

We invoke the Coddington–Levinson theorem for ODEs: for an ODE $\dot{x} = F(x)$ with F locally Lipschitz, a maximal solution $x : [0, T_{\max}) \to \mathbb{R}^2$ can only have $T_{\max} < \infty$ if either

1. $\|x(t)\| \to \infty$ as $t \uparrow T_{\max}$, or
2. $x(t)$ approaches a point where F is not locally Lipschitz (equivalently, the vector field is undefined or blows up).

In our setting, neither can happen:

- $\|x(t)\|$ is uniformly bounded on $[0, T_{\max})$, since $x(t) \in [0,1]^2$;
- F is a polynomial vector field, hence smooth (and therefore locally Lipschitz) on all of $\mathbb{R}^2$.

Thus $T_{\max}$ cannot be finite, and we must have $T_{\max} = +\infty$.

Putting everything together, we have shown:

- for each initial condition $x(0) \in [0,1]^2$ there exists a unique solution $x(t)$ to $\dot{x} = F(x)$ defined on $[0, \infty)$;
- this solution satisfies $x(t) \in [0,1]^2$ for all $t \ge 0$.

This proves both parts (1) and (2) of the theorem. □

We then extend Theorem 3.1 to ODE (2). The key difference is that the low-skill equilibrium shifts to $(\theta_d, 1)$. The proof is identical to that of Theorem 3.1.

Theorem 5.4 (Equilibria of ODE (2)). *Let $\theta_a, \theta_d, \kappa, \Delta$ be the parameters of ODE (2). When $\theta_d \in (0,1)$ and $\theta_a \in (\theta_d, 1)$, the ODE system exhibits two* stable *nodes (sinks): a high-skill equilibrium at $(\theta^\infty, p^\infty) = (1, 0)$ and a low-skill equilibrium $(\theta^\infty, p^\infty) = (\theta_d, 1)$; three unstable nodes (sources) at $(0,0)$, $(1,0)$ and $(1,1)$; and an additional* interior saddle point $(\theta^\dagger, p^\dagger) = (\theta_a, \frac{1-\theta_a}{1-\theta_a+\Delta(\theta_a-\theta_d)})$.

5.4 Analysis of stochastic variant of ODE (3)

The analysis of ODE (3) relies on the limit $\eta \to 0$. In practice, finite step sizes and noisy feedback introduce stochasticity. To study how randomness affects convergence, we consider the following stochastic differential equation (SDE):

$$\begin{aligned} d\theta(t) &= \theta(t)(1-\theta(t))\Big((1-p(t))(1-\theta(t)) - \Delta p(t)\theta(t)\Big)\, dt, \\ dp(t) &= \kappa p(t)(1-p(t))\Big(\big((1-\theta(t))^2 - (1-\theta_a)^2\big)\, dt + \sigma\, dW_t\Big), \end{aligned} \tag{10}$$

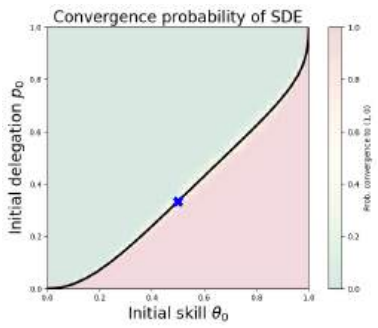


Figure 3: Heatmap of the probability of converging to the high-skill equilibrium $(1, 0)$ for SDE (10) as a function of the initial state (θ_0, p_0), with default settings of $(\theta_a, \kappa, \Delta, \sigma) = (0.5, 3, 2, 0.1)$. Green indicates probability 0, and red indicates probability 1.

where $\sigma \geq 0$ controls the noise level and W_t is a standard Brownian motion. When $\sigma = 0$, this reduces to ODE (3).

Figure 3 shows how the probability of converging to the high-skill equilibrium $(1, 0)$ varies with the initial state $(\theta_0, p_0) \in (0, 1)^2$. We observe that the heatmap closely matches the basin structure in Figure 1b for the ODE, exhibiting a sharp transition in convergence probability. This supports the robustness of our conclusions in Section 3.

6 Omitted proofs and results from Section 3

In this section, we provide proofs for the results in Section 3 and show how to derive an approximation of the stable manifold ψ.

6.1 Proof of Theorem 3.1: equilibrium of ODE (3)

Theorem 6.1 (Restatement of Theorem 3.1). *Let θ_a, κ, Δ be the parameters of ODE (3). When $\theta_a \in (0, 1)$, the ODE system exhibits two* stable *nodes (sinks): a high-skill equilibrium at $(\theta^\infty, p^\infty) = (1, 0)$ and a low-skill equilibrium $(\theta^\infty, p^\infty) = (0, 1)$; two unstable nodes (sources) at $(0, 0)$ and $(1, 1)$; and an additional* interior saddle point $(\theta^\dagger, p^\dagger) = (\theta_a, \frac{1-\theta_a}{1-(1-\Delta)\theta_a})$.

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.

Existence of equilibria. The equilibria are the solutions to the system of algebraic equations $\dot{\theta} = 0$ and $\dot{p} = 0$:

$$\theta(1-\theta)\left((1-p)(1-\theta) - \Delta p\,\theta\right) = 0, \tag{11}$$

$$\kappa\, p(1-p)\left((1-\theta)^2 - (1-\theta_a)^2\right) = 0. \tag{12}$$

From the factors $\theta(1-\theta)$ in (11) and $p(1-p)$ in (12), we immediately identify the four corner equilibria in the domain $\{0, 1\}^2$:

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

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

$$(1-\theta)^2 - (1-\theta_a)^2 = 0, \tag{13}$$
$$(1-p)(1-\theta) - \Delta p\,\theta = 0. \tag{14}$$

Since $\theta, \theta_a \in [0,1]$, Equation (13) implies $1-\theta = 1-\theta_a$, yielding the unique skill solution $\theta^\dagger = \theta_a$. Substituting $\theta^\dagger = \theta_a$ into (14) and solving for p yields:

$$(1-p)(1-\theta_a) = \Delta p \theta_a \implies 1-\theta_a = p\,(1-\theta_a + \Delta\theta_a)\,.$$

Thus, the unique interior equilibrium is

$$(\theta^\dagger, p^\dagger) = \left(\theta_a, \frac{1-\theta_a}{1-(1-\Delta)\theta_a}\right).$$

Stability analysis. We analyze the linear stability via the Jacobian matrix J of the system. Let $f(\theta,p) = \dot{\theta}$ and $g(\theta,p) = \dot{p}$. The Jacobian is given by

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

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 $\partial_\theta g = 0$). 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:

1. **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.

2. **Sinks ($\theta^\infty \neq p^\infty$):**
   - At $(1,0)$: $\lambda_\theta = 0$ and $\lambda_p = -\kappa(1-(1-\theta_a)^2) < 0$. 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. Thus, $(1,0)$ is asymptotically stable.
   - At $(0,1)$: $\lambda_\theta = 0$ and $\lambda_p = -\kappa(1-\theta_a)^2 < 0$. 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. Thus, $(0,1)$ is asymptotically stable.

Thus, the high-skill equilibrium $(1, 0)$ and low-skill equilibrium $(0, 1)$ are stable nodes.

Interior saddle point. At the interior equilibrium $(\theta^\dagger, p^\dagger)$, the diagonal terms vanish because the condition for equilibrium (Equations (13) and (14)) sets the terms multiplying $\theta(1-\theta)$ and $p(1-p)$ to zero. The Jacobian simplifies to:

$$J(\theta^\dagger, p^\dagger) = \begin{bmatrix} \partial_\theta \dot{\theta} & \partial_p \dot{\theta} \\ \partial_\theta \dot{p} & 0 \end{bmatrix}.$$

Evaluating the partial derivatives at $(\theta^\dagger, p^\dagger)$:

$$\partial_p \dot{\theta} = -\frac{\theta_a^2(1-\theta_a)}{p^\dagger} < 0,$$
$$\partial_\theta \dot{p} = -2\kappa p^\dagger (1-p^\dagger)\theta_a(1-\theta_a) < 0.$$

The determinant of the Jacobian is $\det(J) = -(\partial_p \dot{\theta})(\partial_\theta \dot{p})$. Since both off-diagonal terms are negative, their product is positive, making the determinant negative:

$$\det(J) < 0.$$

A negative determinant implies one positive and one negative eigenvalue. Therefore, the interior equilibrium $(\theta^\dagger, p^\dagger)$ is a saddle point. □

6.2 Proof of Theorem 3.2: two basins divided by saddle point

Theorem 6.2 (**Restatement of Theorem 3.2**). *Let θ_a, κ, Δ be the parameters of ODE (3). Let $(p^\dagger, \theta^\dagger)$ be the interior saddle point guaranteed by Theorem 3.1. There exists a monotonically increasing 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 state space $[0,1]^2$ into two distinct 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. The proof proceeds in two steps: first, we establish the existence of the separatrix ψ using the local geometry of the saddle point; second, we use the monotonicity property (Lemma 3.5) to characterize the global basins of attraction defined by this curve.

Existence and properties of the stable manifold. From Theorem 3.1, the interior equilibrium $(\theta^\dagger, p^\dagger)$ is a hyperbolic saddle point. By the Stable Manifold Theorem for planar system [24], there exists a unique one-dimensional stable manifold $W^s(\theta^\dagger, p^\dagger)$ passing through $(\theta^\dagger, p^\dagger)$ tangent to the stable eigenvector. Let this manifold be represented locally as a curve. Since the vector field is smooth and the boundaries of $[0,1]^2$ are invariant, W^s extends to the boundaries of the domain.

Partitioning the basins of attraction. The curve $p = \psi(\theta)$ divides the interior of the state space into two connected components: the set $S_{\text{high}} = \{(\theta, p) : p < \psi(\theta)\}$ and the set $S_{\text{low}} = \{(\theta, p) : p > \psi(\theta)\}$. Consider an initial state (θ_0, p_0) on the curve, i.e., $p_0 = \psi(\theta_0)$. By definition of the stable manifold, the trajectory converges to the saddle point $(\theta^\dagger, p^\dagger)$.

Now consider an initial state $(\theta_0, p_0) \in S_{\text{low}}$, meaning $p_0 > \psi(\theta_0)$. Let (θ_0, p_0') be a reference point on the manifold such that $p_0' = \psi(\theta_0)$. Since $p_0 > p_0'$, Lemma 3.5 implies that for all $t > 0$, the delegation level satisfies $p(t; \theta_0, p_0) \geq p(t; \theta_0, p_0')$ and the skill satisfies $\theta(t; \theta_0, p_0) \leq \theta(t; \theta_0, p_0')$. The reference trajectory on the manifold converges to the saddle $(\theta^\dagger, p^\dagger)$. Because the actual trajectory starts with strictly higher delegation, it cannot cross the invariant manifold and converge to the saddle point $(\theta^\dagger, p^\dagger)$. Since the only stable attractors in the system are $(1, 0)$ and $(0, 1)$ by Theorem 3.1, and it must converge to an equilibrium by Theorem 5.2, it must converge to the low-skill equilibrium $(0, 1)$ such that $\theta^\infty = 0 < \theta^\dagger$.

By a symmetric argument, for any initial state $(\theta_0, p_0) \in S_{\text{high}}$ where $p_0 < \psi(\theta_0)$, Lemma 3.5 implies the learner maintains lower delegation and higher skill relative to the manifold trajectory, forcing convergence to the high-skill equilibrium $(1, 0)$. Thus, ψ is the separatrix dividing the two basins of attraction.

Monotonicity of the stable manifold ψ. We prove that $\psi(\theta)$ is monotonically increasing by contradiction. Suppose, for the sake of contradiction, that ψ is not monotonically increasing. Then there must exist two skill levels $\theta_1 < \theta_2$ such that $\psi(\theta_1) > \psi(\theta_2)$. By the continuity of ψ, we can select a delegation level $\bar{p}$ such that $\psi(\theta_2) < \bar{p} < \psi(\theta_1)$. Now consider two initial states at this same delegation level $\bar{p}$ but different skill levels:

- State $A = (\theta_1, \bar{p})$. Since $\bar{p} < \psi(\theta_1)$, state A lies below the separatrix. It follows from the above analysis that it must converge to the **high-skill equilibrium** $(1, 0)$.
- State $B = (\theta_2, \bar{p})$. Since $\bar{p} > \psi(\theta_2)$, state B lies above the separatrix. It follows from the above analysis that it must converge to the **low-skill equilibrium** $(0, 1)$.

This implies that state A (with lower initial skill θ_1) achieves a better outcome than state B (with higher initial skill θ_2), holding delegation constant. However, Lemma 3.5 states that $\theta(t)$ is non-decreasing in the initial skill gap $-(1 - \theta_0)$, meaning higher initial skill should lead to higher (or equal) limiting skill. State B having higher initial skill but converging to lower limiting skill ($\theta^\infty = 0$) than State A ($\theta^\infty = 1$) is a direct contradiction. Therefore, $\psi(\theta)$ must be monotonically increasing.

Extension of ψ to the boundary. It remains to show that the separatrix ψ extends continuously to the corners, i.e., $\lim_{\theta \to 0} \psi(\theta) = 0$ and $\lim_{\theta \to 1} \psi(\theta) = 1$. The stable manifold W^s is an invariant set. Since trajectories cannot cross, the closure of W^s must contain the limit sets of its trajectories in backward time.

Consider the limit as $\theta \to 0$. Note that points on the boundary segment $\{0\} \times (0, 1)$ are not equilibria; the flow on this boundary satisfies $\dot{p} = \kappa p(1 - p)(1 - (1 - \theta_a)^2) > 0$. Thus, any trajectory starting on the interior of the left boundary converges to the low-skill sink $(0, 1)$ and originates from the source $(0, 0)$. If the separatrix approached a point $(0, p^*)$ with $p^* > 0$, then by continuity of flow, a neighborhood of points near $(0, p^*)$ would follow the boundary flow toward $(0, 1)$. This contradicts the definition of the separatrix as the boundary of the high-skill basin. Specifically, if

$\lim_{\theta \to 0} \psi(\theta) = p^* > 0$, we could pick an initial condition $(\epsilon, p^* - \delta)$ in the high-skill basin (below ψ) that is arbitrarily close to the boundary flow converging to the low-skill sink. This discontinuity is impossible in the smooth vector field. Therefore, the only possible limit point is the source $(0, 0)$.

A symmetric argument applies to the right boundary $\theta = 1$: points on $\{1\} \times (0, 1)$ flow toward the high-skill sink $(1, 0)$ (since $\dot{p} < 0$), preventing the separatrix from attaching anywhere except the source $(1, 1)$. Thus, ψ connects $(0, 0)$ to $(1, 1)$ through the saddle.

□

6.3 Proof of Theorem 3.3: effects of θ_a on ψ

Theorem 6.3 (**Restatement of Theorem 3.3**). *For all $\theta \in (0, 1)$, $\psi(\theta)$ is monotonically decreasing in θ_a and continuously differentiable with respect to θ_a. Moreover, for any $(\theta_0, p_0) \in (0, 1)^2$, there exists $\theta_a \in (0, 1)$ such that $p_0 = \psi_{\theta_a}(\theta_0)$.*

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

Monotonicity with respect to θ_a. We show that for any fixed $\theta \in (0, 1)$, $\psi_{\theta_a}(\theta)$ is decreasing in θ_a. Let $\theta_{a,1} < \theta_{a,2}$ be two distinct AI skill levels. 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 $\psi_2(\theta^*) > \psi_1(\theta^*)$. Since the functions are continuous and connect $(0, 0)$ to $(1, 1)$ (Theorem 3.2), we can select a test point (θ^*, p_0) in the state space such that:

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

Now, consider the long-run convergence of a learner starting at (θ^*, p_0) under the two different AI regimes:

- **Regime 1 ($\theta_{a,1}$):** Since $p_0 > \psi_1(\theta^*)$, the initial state lies in the *Low-Skill Basin* (above the separatrix). By Theorem 3.2, the learner converges to the low-skill equilibrium: $\theta_1^\infty = 0$.
- **Regime 2 ($\theta_{a,2}$):** Since $p_0 < \psi_2(\theta^*)$, the initial state lies in the *High-Skill Basin* (below the separatrix). By Theorem 3.2, the learner converges to the high-skill equilibrium: $\theta_2^\infty = 1$.

Comparing the outcomes, we have $\theta_2^\infty > \theta_1^\infty$. However, Lemma 3.5 states that for any time t, the skill $\theta(t)$ is non-increasing in θ_a. Taking the limit $t \to \infty$, this implies that increasing the AI skill from $\theta_{a,1}$ to $\theta_{a,2}$ must result in $\theta_2^\infty \leq \theta_1^\infty$. The result $\theta_2^\infty = 1 > 0 = \theta_1^\infty$ is a direct contradiction. Therefore, the assumption must be false, and it must hold that $\psi_2(\theta) \leq \psi_1(\theta)$ for all $\theta \in (0, 1)$.

Differentiability. The vector field defined by ODE (3) is C^∞-smooth with respect to the parameter θ_a. Since the saddle point $(\theta^\dagger, p^\dagger)$ is hyperbolic for all $\theta_a \in (0, 1)$, the Stable Manifold Theorem with parameters [24] guarantees that the local stable manifold varies smoothly (is C^1) with respect to θ_a. Since the global separatrix ψ_{θ_a} is obtained by backward integration of the local manifold—and the flow of the ODE is smooth with respect to parameters—the global function $\psi_{\theta_a}(\theta)$ is continuously differentiable with respect to θ_a.

Sweeping Property. We show that for any point (θ_0, p_0), there exists a θ_a such that the separatrix passes through it. We analyze the geometric limits of the saddle point $(\theta^\dagger, p^\dagger)$ as θ_a approaches the boundaries:

- **Limit $\theta_a \to 1$:** The saddle point $(\theta^\dagger, p^\dagger)$ converges to the high-skill corner $(1, 0)$. The stable manifold connecting $(0, 0)$ to the saddle collapses toward the θ-axis ($p = 0$). Thus, $\lim_{\theta_a \to 1} \psi_{\theta_a}(\theta_0) = 0$.
- **Limit $\theta_a \to 0$:** The saddle point converges to the low-skill corner $(0, 1)$ (since $p^\dagger \to 1$). The stable manifold connects the source $(0, 0)$ to the saddle $(0, 1)$ and effectively fills the upper triangle, pushing the boundary toward $p = 1$. Thus, $\lim_{\theta_a \to 0} \psi_{\theta_a}(\theta_0) = 1$.

Since $\psi_{\theta_a}(\theta_0)$ is a continuous function of θ_a and its range includes the interval $(0, 1)$ as θ_a varies, the Intermediate Value Theorem implies that for any $p_0 \in (0, 1)$, there exists a value $\theta_a \in (0, 1)$ such that $\psi_{\theta_a}(\theta_0) = p_0$. □

6.4 Proof of Theorem 3.4: short-term gain, long-term losses

Theorem 6.4 (Restatement of Theorem 3.4). *Let θ_a, κ, Δ be the parameters of ODE (3). Given an initial state $(\theta_0, p_0) \in (0, 1)^2$ with $\theta_0 < \theta_a$, let $t_c := \inf\{t \geq 0 : \ \forall t' > t, \ G_\ell(t') > 0\}$ denote the crossing time after which AI delegation incurs higher loss than the no-AI baseline. Then $t_c < \infty$. Moreover, $t_c = 0$ when $\theta_0 \geq \theta_a$ and $\theta(t_c; \theta_0, 0) \leq \theta_a$ when $\theta_0 < \theta_a$.*

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

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

Since $p_0 > 0$, Lemma 3.5 implies 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:

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

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

$$(1 - \theta(t; \theta_0, p_0))^2 - (1 - \theta(t; \theta_0, 0))^2 \geq 0. \tag{15}$$

Case 1: $\theta_0 \geq \theta_a$. If the learner starts with skill superior to the AI, then $\theta(t; \theta_0, 0) > \theta_0 \geq \theta_a$ for all $t \geq 0$. This implies $(1 - \theta_a)^2 > (1 - \theta(t; \theta_0, 0))^2$. Since $p(t; \theta_0, p_0) > 0$ and the skill difference (15) is positive, every term in the expression for $G_\ell(t)$ is non-negative, and the second term is strictly positive. Thus, $G_\ell(t) > 0$ for all $t > 0$, implying $t_c = 0$.

Case 2: $\theta_0 < \theta_a$. 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 (15) and the fact that $p(t;\theta_0,p_0)>0$, we conclude that $G_\ell(t)$ is a sum of strictly positive terms for all $t>t^*$. The crossing time is defined as $t_c := \inf\{t\geq 0 : \forall t'>t, G_\ell(t')>0\}$. Since $G_\ell(t)>0$ holds for all $t>t^*$, it follows that $t_c\leq t^*<\infty$. Finally, because $G_\ell(t)$ becomes strictly positive whenever $\theta(t;\theta_0,0)>\theta_a$, any period where the AI provides a benefit ($G_\ell(t)\leq 0$) must occur while the baseline skill is still below the AI skill. Therefore, at the crossing time, the baseline skill satisfies $\theta(t_c;\theta_0,0)\leq\theta_a$. □

6.5 Proof of Lemma 3.5: negative coupling between skill and delegation

Lemma 6.5 (**Restatement of Lemma 3.5**). *Let θ_a, κ, and Δ be the parameters of ODE (3). For any $t>0$, $\theta(t)$ is non-increasing and $p(t)$ is non-decreasing as functions of each of the following variables: the initial delegation p_0, the initial skill gap $1-\theta_0$, and θ_a.*

Proof. We prove the lemma using mathematical induction on the Euler-discretized system. Let $\eta>0$ be a sufficiently small step size. The system evolves according to the map $M:[0,1]^2\to[0,1]^2$ defined by:

$$\begin{aligned}\theta_{k+1} &= \Phi_\theta(\theta_k,p_k) := \theta_k+\eta\cdot\theta_k(1-\theta_k)\left[(1-p_k)(1-\theta_k)-\Delta p_k\theta_k\right],\\ p_{k+1} &= \Phi_p(\theta_k,p_k) := p_k+\eta\cdot\kappa p_k(1-p_k)\left[(1-\theta_k)^2-(1-\theta_a)^2\right].\end{aligned}$$

We define the partial order $\preceq$ on the state space such that $(\theta,p)\preceq(\hat\theta,\hat p)$ if and only if:

$$\theta\leq\hat\theta \quad\text{and}\quad p\geq\hat p.$$

Monotonicity of the update map. We first show that the map M preserves this order. That is, if $(\theta,p)\preceq(\hat\theta,\hat p)$, then $(\Phi_\theta(\theta,p),\Phi_p(\theta,p))\preceq(\Phi_\theta(\hat\theta,\hat p),\Phi_p(\hat\theta,\hat p))$. This requires checking the partial derivatives of the update functions.
Skill update Φ_θ. The function Φ_θ is differentiable on $[0,1]^2$. By the Mean Value Theorem, there exists a point $\xi=(\xi_\theta,\xi_p)$ on the line segment connecting (θ,p) and $(\hat\theta,\hat p)$ such that:

$$\Phi_\theta(\hat\theta,\hat p)-\Phi_\theta(\theta,p)=\nabla\Phi_\theta(\xi)\cdot\begin{pmatrix}\hat\theta-\theta\\ \hat p-p\end{pmatrix}=\frac{\partial\Phi_\theta}{\partial\theta}(\xi)(\hat\theta-\theta)+\frac{\partial\Phi_\theta}{\partial p}(\xi)(\hat p-p).$$

We evaluate the signs of the partial derivatives for sufficiently small η:

- $\frac{\partial\Phi_\theta}{\partial\theta}=1+\eta\frac{\partial f}{\partial\theta}$. Since f and its derivative are bounded on the compact set $[0,1]^2$, for small η, this term is strictly positive (>0).
- $\frac{\partial\Phi_\theta}{\partial p}=\eta\frac{\partial f}{\partial p}=\eta\cdot\theta(1-\theta)[-(1-\theta)-\Delta\theta]$. Since $\theta,\Delta\geq 0$, this derivative is strictly non-positive (≤ 0).

Substituting these signs back into the MVT equation:

- Term 1: $\frac{\partial\Phi_\theta}{\partial\theta}(\xi)>0$ and $(\hat\theta-\theta)\geq 0 \implies$ Product is ≥ 0.
- Term 2: $\frac{\partial\Phi_\theta}{\partial p}(\xi)\leq 0$ and $(\hat p-p)\leq 0$ (since $p\geq\hat p$) $\implies$ Product is ≥ 0.

Thus, the sum is non-negative:

$$\Phi_\theta(\hat{\theta},\hat{p}) - \Phi_\theta(\theta,p) \geq 0 \implies \Phi_\theta(\theta,p) \leq \Phi_\theta(\hat{\theta},\hat{p}).$$

Delegation update Φ_p. Similarly, applying the Mean Value Theorem to Φ_p, there exists a point ζ on the segment such that:

$$\Phi_p(\hat{\theta},\hat{p}) - \Phi_p(\theta,p) = \frac{\partial \Phi_p}{\partial \theta}(\zeta)(\hat{\theta}-\theta) + \frac{\partial \Phi_p}{\partial p}(\zeta)(\hat{p}-p).$$

Evaluating the derivatives:

- $\frac{\partial \Phi_p}{\partial p} = 1 + \eta\frac{\partial g}{\partial p}$. For small η, this is strictly positive (> 0).
- $\frac{\partial \Phi_p}{\partial \theta} = \eta\frac{\partial g}{\partial \theta} = \eta \cdot \kappa p(1-p)[-2(1-\theta)]$. This derivative is strictly non-positive (≤ 0).

Analyzing the terms:

- Term 1: $\frac{\partial \Phi_p}{\partial \theta}(\zeta) \leq 0$ and $(\hat{\theta}-\theta) \geq 0 \implies$ Product is ≤ 0.
- Term 2: $\frac{\partial \Phi_p}{\partial p}(\zeta) > 0$ and $(\hat{p}-p) \leq 0 \implies$ Product is ≤ 0.

Thus, the sum is non-positive:

$$\Phi_p(\hat{\theta},\hat{p}) - \Phi_p(\theta,p) \leq 0 \implies \Phi_p(\hat{\theta},\hat{p}) \leq \Phi_p(\theta,p).$$

Combining the above analysis, we have proved that M preserves the order $\preceq$.

Inductive proof. We now apply the order-preserving property of the map M to the specific variables.

Case A: initial delegation p_0. Let two learners have states (θ_0, p_0) and $(\theta_0, \hat{p}_0)$ with $p_0 > \hat{p}_0$.

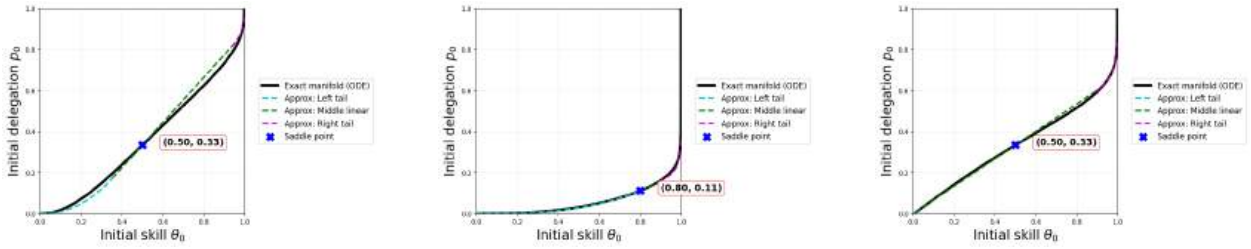
- *Base Case:* $(\theta_0, p_0) \preceq (\theta_0, \hat{p}_0)$ holds immediately.
- *Inductive Step:* Assume $(\theta_k, p_k) \preceq (\hat{\theta}_k, \hat{p}_k)$. By the monotonicity of M proven above, the next state satisfies $(\theta_{k+1}, p_{k+1}) \preceq (\hat{\theta}_{k+1}, \hat{p}_{k+1})$.

Case B: initial skill gap $1-\theta_0$. Let two learners have states (θ_0, p_0) and $(\hat{\theta}_0, p_0)$ where $\theta_0 < \hat{\theta}_0$.

- *Base Case:* $(\theta_0, p_0) \preceq (\hat{\theta}_0, p_0)$ holds immediately.
- *Inductive Step:* By the monotonicity of M, the order is preserved.

Case C: AI skill θ_a. Let system 1 have θ_a and system 2 have $\hat{\theta}_a$ with $\theta_a > \hat{\theta}_a$. Here, the update maps M_1 and M_2 differ slightly. Specifically, in the delegation update Φ_p, the drift term involves $-(1-\theta_a)^2$. Since $\theta_a > \hat{\theta}_a$, we have $-(1-\theta_a)^2 > -(1-\hat{\theta}_a)^2$. Thus, for any fixed state (θ, p), we have $\Phi_p(\theta, p; \theta_a) > \Phi_p(\theta, p; \hat{\theta}_a)$.

- *Base Case:* $(\theta_0, p_0) = (\hat{\theta}_0, \hat{p}_0)$.



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

Figure 4: Plots illustrating the closeness between the stable manifold $\psi(\cdot)$ and its approximation $\widetilde{\psi}$, with default parameter settings $(\theta_a, \kappa, \Delta) = (0.5, 3, 2)$.

- *Step 1:* $\theta_1 = \hat{\theta}_1$ (skill update independent of θ_a given same state), but $p_1 > \hat{p}_1$ due to the parameter difference. Thus $(\theta_1, p_1) \preceq (\hat{\theta}_1, \hat{p}_1)$.
- *Inductive Step:* Assume $(\theta_k, p_k) \preceq (\hat{\theta}_k, \hat{p}_k)$. Apply the map M_1 to the LHS: $p_{k+1} = \Phi_p(\theta_k, p_k; \theta_a) \geq \Phi_p(\theta_k, p_k; \hat{\theta}_a)$ (Parameter monotonicity). Then use State monotonicity: $\Phi_p(\theta_k, p_k; \hat{\theta}_a) \geq \Phi_p(\hat{\theta}_k, \hat{p}_k; \hat{\theta}_a) = \hat{p}_{k+1}$. Similarly for skill: $\theta_{k+1} = \Phi_\theta(\theta_k, p_k) \leq \Phi_\theta(\hat{\theta}_k, \hat{p}_k) = \hat{\theta}_{k+1}$. Thus, the order is preserved.

Taking $\eta \to 0$, the discrete order preservation implies the continuous trajectories satisfy the monotonicity properties. □

6.6 Approximation of the stable manifold of the saddle point

In this section, we construct an explicit approximation $\widetilde{\psi}$ to the stable manifold ψ (Eq. (16)) to enable actionable predictions. Figure 4a plots $\widetilde{\psi}$ under different parameter choices and shows that it closely tracks ψ.

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 (C^1 continuity).

1. Linear approximation at the saddle. We first linearize the system around the saddle point $(\theta^\dagger, p^\dagger)$. The Jacobian matrix J evaluated at the saddle is given by:

$$J = \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix} = \begin{bmatrix} \partial_\theta \dot{\theta} & \partial_p \dot{\theta} \\ \partial_\theta \dot{p} & \partial_p \dot{p} \end{bmatrix}.$$

At the saddle, we have $\partial_p \dot{p} = 0$, so $J_{22} = 0$. The other entries are:

$$\begin{aligned} J_{11} &= (1 - 2\theta^\dagger)\left[(1 - p^\dagger)(1 - \theta^\dagger) - \Delta p^\dagger \theta^\dagger\right] + \theta^\dagger(1 - \theta^\dagger)[-(1 - p^\dagger) - \Delta p^\dagger] \\ &= -\theta^\dagger(1 - \theta^\dagger)(1 - p^\dagger + \Delta p^\dagger) \\ J_{12} &= \theta^\dagger(1 - \theta^\dagger)[-(1 - \theta^\dagger) - \Delta \theta^\dagger] = -\theta^\dagger(1 - \theta^\dagger)(1 - \theta^\dagger(1 - \Delta)) \\ J_{21} &= -2\kappa p^\dagger(1 - p^\dagger)(1 - \theta^\dagger) = -\frac{a_1}{1 - \theta^\dagger} \cdot 2(1 - \theta^\dagger) \end{aligned}$$

The characteristic equation for eigenvalues λ is $\lambda^2 - J_{11}\lambda - J_{12}J_{21} = 0$. The stable eigenvalue λ_s is the negative root. The associated eigenvector $v_s = [1, m^\dagger]^T$ satisfies $(J_{11} - \lambda_s) + J_{12}m^\dagger = 0$. Solving for the slope $m^\dagger$ yields the expression:

$$m^\dagger = \frac{\lambda_s - J_{11}}{J_{12}} = \frac{-J_{11} - \sqrt{J_{11}^2 + 4J_{12}J_{21}}}{2J_{12}}.$$

Thus, near the saddle, the manifold is approximated by the line:

$$\psi_{\text{mid}}(\theta) = p^\dagger + m^\dagger(\theta - \theta^\dagger).$$

2. Power-law approximation at $(0,0)$. Near the origin $(0,0)$, we approximate the dynamics by keeping only the lowest-order linear terms. For $\theta \approx 0, p \approx 0$:

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

We define the ratio of the growth rates as $\beta_l := \frac{\dot{p}/\rho}{\dot{\theta}/\theta} = \kappa(1-(1-\theta^\dagger)^2)$. The differential equation describing the trajectory shape is $\frac{dp}{d\theta} = \frac{\dot{p}}{\dot{\theta}} = \beta_l \frac{p}{\theta}$. Integrating this yields the power law form:

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

3. Power-law approximation at $(1,1)$. Near the corner $(1,1)$, we transform variables to $x = 1-\theta$ and $y = 1-p$, where $x, y \approx 0$. The dynamics become:

$$\dot{x} = -\dot{\theta} \approx -\theta(1-\theta)(-\Delta p\theta) \approx \Delta x. \quad (\text{since } \theta \to 1, p \to 1).$$
$$\dot{y} = -\dot{p} \approx -\kappa p(1-p)(-(1-\theta_a)^2) \approx \kappa y(1-\theta^\dagger)^2.$$

We define the ratio of decay rates as $\beta_r := \frac{\dot{y}}{\dot{x}} = \frac{\kappa(1-\theta^\dagger)^2}{\Delta}$. The trajectory shape satisfies $\frac{dy}{dx} = \beta_r \frac{y}{x}$. Integrating yields $y = C_r x^{\beta_r}$. Transforming back to original coordinates, we get:

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

4. Smooth pasting at breakpoints. We determine the breakpoints θ_l and θ_r by matching the linear segment ψ_{mid} to the power laws ψ_{left} and ψ_{right}.

- **Left Breakpoint θ_l:** We require $\psi_{\text{left}}(\theta_l) = \psi_{\text{mid}}(\theta_l)$ and $\psi'_{\text{left}}(\theta_l) = \psi'_{\text{mid}}(\theta_l)$. The derivative condition gives:

$$C_l\beta_l\theta_l^{\beta_l - 1} = m^\dagger \implies \frac{\beta_l}{\theta_l}(C_l\theta_l^{\beta_l}) = m^\dagger \implies \frac{\beta_l}{\theta_l}\psi_{\text{mid}}(\theta_l) = m^\dagger.$$

  Substituting the linear form $\psi_{\text{mid}}(\theta_l) = p^\dagger + m^\dagger(\theta_l - \theta^\dagger)$:

$$\beta_l(p^\dagger + m^\dagger\theta_l - m^\dagger\theta^\dagger) = m^\dagger\theta_l.$$

  Rearranging terms to solve for θ_l:

$$\beta_l(p^\dagger - m^\dagger\theta^\dagger) = m^\dagger\theta_l(1-\beta_l) \implies \theta_l = \frac{\beta_l(p^\dagger - m^\dagger\theta^\dagger)}{m^\dagger(1-\beta_l)}.$$

- **Right Breakpoint θ_r:** Similarly, we match the function and derivative at θ_r. The derivative condition for the right segment is:

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

Setting this equal to $m^\dagger$ at θ_r:

$$\frac{\beta_r}{1-\theta_r}(1-\psi_{\text{mid}}(\theta_r)) = m^\dagger.$$

Substituting $\psi_{\text{mid}}(\theta_r) = p^\dagger + m^\dagger(\theta_r - \theta^\dagger)$:

$$\beta_r(1-p^\dagger - m^\dagger\theta_r + m^\dagger\theta^\dagger) = m^\dagger(1-\theta_r).$$

Rearranging terms to solve for θ_r:

$$\beta_r(1-p^\dagger + m^\dagger\theta^\dagger) - \beta_r m^\dagger\theta_r = m^\dagger - m^\dagger\theta_r$$

$$m^\dagger\theta_r(1-\beta_r) = m^\dagger - \beta_r(1-p^\dagger + m^\dagger\theta^\dagger)$$

$$\theta_r = \frac{m^\dagger - \beta_r(1-p^\dagger + m^\dagger\theta^\dagger)}{m^\dagger(1-\beta_r)}.$$

To ensure $\theta_l \le \theta_a \le \theta_r$, we let

$$\theta_l := \min\{\theta^\dagger, \frac{\beta_l(p^\dagger - m^\dagger\theta^\dagger)}{m^\dagger(1-\beta_l)}\}, \text{ and } \theta_r := \min\{1, \frac{m^\dagger - \beta_r(1-p^\dagger + m^\dagger\theta^\dagger)}{(1-\beta_r)m^\dagger}\}.$$

In summary, the approximation $\widetilde{\psi}$ is:

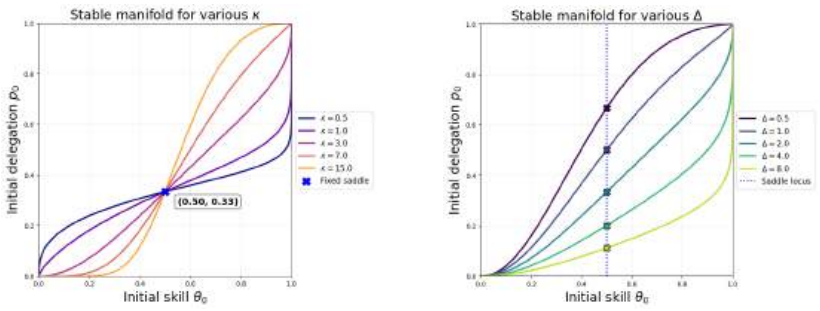
$$\widetilde{\psi}(\theta) := \begin{cases} (p^\dagger + m^\dagger\cdot(\theta_l - \theta^\dagger))\cdot(\frac{\theta}{\theta_l})^{\beta_l} & \text{if } 0 \le \theta \le \theta_l \\ p^\dagger + m^\dagger\cdot(\theta - \theta^\dagger) & \text{if } \theta_l < \theta < \theta_r \\ 1-(1-p_m(\theta_r))\cdot(\frac{1-\theta}{1-\theta_r})^{\beta_r} & \text{if } \theta_r \le \theta \le 1. \end{cases} \tag{16}$$

6.7 Effects of parameters on the stable manifold

Beyond Theorem 3.3, we also study how κ and Δ affect the stable manifold. Figure 5 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 6.6 (Effects of κ on stable manifold). *Fix θ_a. For $\theta \in (0, \theta_a)$, $\psi(\theta)$ is monotonically decreasing in κ and continuously differentiable with respect to κ. For $\theta \in (\theta_a, 1)$, $\psi(\theta)$ is monotonically increasing in κ and continuously differentiable with respect to κ.*

Proof. The lemma establishes the behavior of the stable manifold $\psi(\theta)$ (separatrix) as the delegation speed parameter κ varies. We treat differentiability and monotonicity separately.



(a) Effects of κ (b) Effects of Δ

Figure 5: Plots illustrating the relationship between the basin boundary $\psi(\cdot)$ and model parameters θ_a, κ, Δ, with default setting $(\theta_a, \kappa, \Delta) = (0.5, 3, 2)$.

1. Differentiability. The vector field defined by ODE (3) is C^∞-smooth with respect to all state variables and the parameter κ. Since the interior equilibrium $(\theta^\dagger, p^\dagger)$ is a hyperbolic saddle point for all $\kappa > 0$ (Theorem 3.1), the Stable Manifold Theorem with parameters guarantees that the local stable manifold varies smoothly (C^1) with respect to κ. Since the global separatrix $\psi(\theta)$ is the backward extension of this local manifold, and the flow is smooth, $\psi(\theta)$ is continuously differentiable with respect to κ.

2. Monotonicity via vector field comparison. We prove the monotonicity by comparing the slopes of the trajectories under two different parameters $\kappa_1 < \kappa_2$. Let $F_\kappa(\theta, p) = (\dot{\theta}, \dot{p}_\kappa)$ denote the vector field, where:

$$\dot{\theta} = f(\theta, p), \quad \dot{p}_\kappa = \kappa \cdot g(\theta, p),$$

with $g(\theta, p) = p(1-p)[(1-\theta)^2 - (1-\theta_a)^2]$. Note that f is independent of κ, and g captures the sign of the delegation drift. The slope of the vector field at any point (θ, p) where $f \neq 0$ is given by:

$$m(\theta, p; \kappa) = \frac{dp}{d\theta} = \frac{\dot{p}_\kappa}{\dot{\theta}} = \kappa \frac{g(\theta, p)}{f(\theta, p)}.$$

Case 1: Region $\theta \in (0, \theta_a)$. In this region, the skill is below the AI skill, so $(1-\theta)^2 > (1-\theta_a)^2$, implying $g(\theta, p) > 0$. The separatrix ψ connects the source $(0, 0)$ to the saddle $(\theta_a, p^\dagger)$. Along this curve, θ is increasing, so $\dot{\theta} = f(\theta, p) > 0$. Since $g > 0$ and $f > 0$, the slope m is positive. Comparing $\kappa_1 < \kappa_2$:

$$m(\theta, p; \kappa_2) = \frac{\kappa_2}{\kappa_1} m(\theta, p; \kappa_1) > m(\theta, p; \kappa_1).$$

The vector field for κ_2 is strictly "steeper" (points more upward) than for κ_1 everywhere in this region. Consider a point $z = (\theta_0, p_0)$ lying exactly on the separatrix ψ_1 for κ_1. Under the flow of κ_1, the trajectory $\gamma_1(t)$ starting at z converges to the saddle. Under the flow of κ_2, the trajectory $\gamma_2(t)$ starting at z has a strictly larger slope at every point than γ_1. Geometrically, this implies γ_2 must rise *above* γ_1. Since γ_1 eventually hits the saddle, the steeper trajectory γ_2 must pass "above" the saddle (i.e., to the left of the stable manifold of the saddle in the local linearization), entering the basin of attraction of the low-skill equilibrium $(0, 1)$. For z to be in the low-skill basin of κ_2, it must lie *above* the separatrix ψ_2. Thus, $p_0 > \psi_2(\theta_0)$. Since $p_0 = \psi_1(\theta_0)$, we have $\psi_1(\theta_0) > \psi_2(\theta_0)$. Therefore, $\psi(\theta)$ is monotonically decreasing in κ on $(0, \theta_a)$.

Case 2: Region $\theta \in (\theta_a, 1)$. In this region, $(1-\theta)^2 < (1-\theta_a)^2$, implying $g(\theta, p) < 0$. The separatrix connects the saddle to the source $(1, 1)$ (in backward time), or equivalently, trajectories flow from the saddle towards the high-skill sink $(1, 0)$. However, technically ψ is defined as the stable manifold. In this region, ψ separates flow to $(0, 1)$ and $(1, 0)$.

Consider the slope again. Note that $g(\theta, p)$ is negative, which means delegation is being suppressed. Increasing κ makes $\dot{p}$ *more negative*. Consider a point z on the separatrix ψ_1 (for κ_1). Under κ_1, it flows to the saddle. Under κ_2, the downward push is stronger. The trajectory starting at z will drop *below* the trajectory of κ_1. Since the κ_1 trajectory hits the saddle, the κ_2 trajectory (being lower) will pass "below" the saddle, entering the basin of attraction of the high-skill equilibrium $(1, 0)$. For z to be in the high-skill basin of κ_2, it must lie *below* the separatrix ψ_2. Thus, $p_0 < \psi_2(\theta_0)$. Since $p_0 = \psi_1(\theta_0)$, we have $\psi_1(\theta_0) < \psi_2(\theta_0)$. Therefore, $\psi(\theta)$ is monotonically increasing in κ on $(\theta_a, 1)$. □

Lemma 6.7 (**Effects of Δ on stable manifold**). *For $\theta \in (0, 1)$, $\psi(\theta)$ is monotonically decreasing in Δ and continuously differentiable with respect to Δ.*

Proof. We analyze the dependence of the stable manifold $\psi(\theta)$ on the drift parameter Δ.

Differentiability. The vector field of ODE (3) is smooth (C^∞) with respect to the state variables (θ, p) and the parameter Δ. The interior equilibrium $(\theta^\dagger, p^\dagger)$ depends smoothly on Δ (specifically, $p^\dagger(\Delta) = \frac{1-\theta_a}{1-\theta_a+\Delta\theta_a}$). Since the equilibrium remains a hyperbolic saddle for all $\Delta > 0$, the Stable Manifold Theorem with parameters [24] guarantees that the local stable manifold varies smoothly (C^1) with respect to Δ. By the smoothness of the flow extending the local manifold to the global separatrix, $\psi(\theta)$ is continuously differentiable with respect to Δ.

Monotonicity via trajectory comparison. We prove that $\psi(\theta)$ decreases as Δ increases by establishing a monotonicity property for the trajectories and using a basin-of-attraction argument.

Let $\Delta_1 < \Delta_2$. Consider two systems starting from the same initial state (θ_0, p_0). Comparing the drift functions:

- The delegation update $\dot{p} = g(\theta, p)$ is independent of Δ.
- The skill update $\dot{\theta} = f(\theta, p; \Delta)$ satisfies:

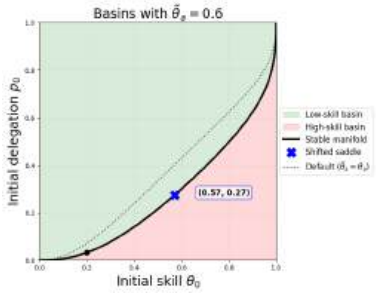
$$\frac{\partial f}{\partial \Delta} = -\theta^2(1-\theta)p < 0 \quad \text{for } \theta, p \in (0, 1).$$

Since a higher Δ strictly reduces the skill growth rate, the system forms a monotone dynamical system with respect to this parameter. Specifically, similar to Lemma 3.5, strictly lower skill growth leads to lower future skill $\theta(t)$, which in turn induces higher delegation $p(t)$. Thus, for any finite $t > 0$, the solutions satisfy:

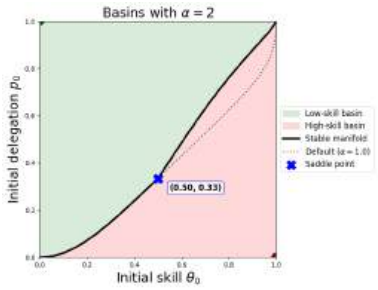
$$\theta(t; \Delta_2) \le \theta(t; \Delta_1) \quad \text{and} \quad p(t; \Delta_2) \ge p(t; \Delta_1).$$

Taking the limit $t \to \infty$, the equilibria reached must satisfy the same ordering: $\theta^\infty(\Delta_2) \le \theta^\infty(\Delta_1)$.

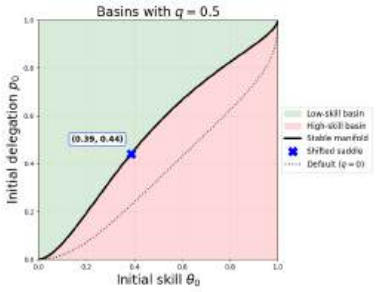
Let ψ_1 and ψ_2 be the separatrices for Δ_1 and Δ_2 respectively. We claim $\psi_2(\theta) < \psi_1(\theta)$ for all $\theta \in (0, 1)$. Suppose, for contradiction, that there exists some θ^* such that $\psi_2(\theta^*) \ge \psi_1(\theta^*)$. We can choose an initial delegation p_0 such that $\psi_1(\theta^*) < p_0 < \psi_2(\theta^*)$. Consider the asymptotic outcome of a learner starting at (θ^*, p_0) under both regimes:



(a) Noisy delegation update (5)



(b) Asymmetric delegation update (7)



(c) Alternative performance loss (17)

Figure 6: Plots illustrating how the basins vary under different model extensions, with default parameter settings $(\theta_a, \kappa, \Delta) = (0.5, 3, 2)$.

- **Under Δ_1:** Since $p_0 > \psi_1(\theta^*)$, the state is in the low-skill basin (above the separatrix). The learner converges to the low-skill equilibrium ($\theta^\infty = 0$).
- **Under Δ_2:** Since $p_0 < \psi_2(\theta^*)$, the state is in the high-skill basin (below the separatrix). The learner converges to the high-skill equilibrium ($\theta^\infty = 1$).

This implies $\theta^\infty(\Delta_2) = 1 > 0 = \theta^\infty(\Delta_1)$. However, this contradicts the trajectory monotonicity, which requires $\theta^\infty(\Delta_2) \le \theta^\infty(\Delta_1)$. Therefore, the assumption is false, and it must hold that $\psi_2(\theta) < \psi_1(\theta)$. Thus, $\psi(\theta)$ is monotonically decreasing in Δ. □

7 Omitted details from Section 4

In this section, we provide detailed derivations of ODEs for the extensions in Section 4. In addition, we present an extension with an alternative performance objective in Section 7.4. Figure 6 plots the basins of attraction under different model extensions.

7.1 Derivation of ODE (5) for jagged AI

With a jagged AI, the main difference is that the performance loss by AI's output is random. Consequently, the mechanism of learning in Section 5.1 extends to be:

- **(Task completion stage)** Learner delegates the task to AI with probability $p(t)$, which sets $X(t) = 1$; and does the task by itself with probability $1 - p(t)$, which sets $X(t) = 0$.
- **(Evaluation stage)** Learner submits the output to the teacher. The teacher gives a loss ℓ_t to the learner. If the output is done by the learner itself, $\ell_t = (1 - \theta(t))^2$. Otherwise, if the output is done by AI, $\ell_t = (1 - s_t)^2$, where $s_t \sim \mu_a$ is a random AI skill drawn from μ_a. Then

$$\ell_t := (1 - X(t))(1 - \theta(t))^2 + X(t) \cdot (1 - s_t)^2.$$

- **(Update of skill)** If $X(t) = 0$, the learner skill (intrinsically) updates towards the target 1 according to the loss $\ell_t = (1 - \theta(t))^2$:

$$\theta(t + 1) := \theta(t) + 2\eta\theta(t)(1 - \theta(t))^2$$

Otherwise, if $X(t) = 1$, the learner skill decays towards θ_d due to non-practice:

$$\theta(t+1) := \theta(t) + 2\eta\Delta\theta(t)(1-\theta(t))(\theta_d - \theta(t)).$$

- **(Update of delegation level)** If $X(t) = 0$, assume that the learner has an accurate belief $\mathbb{E}_s[(1-s)^2]$ on the expected performance loss of AI's output. Otherwise, if $X(t) = 1$, assume that the learner has an accurate belief $(1-\theta(t))^2$ on the loss of its output. Together with the obtained loss ℓ_t, the learner updates its delegation level as follows:

$$\begin{aligned} p(t+1) := p(t) + \kappa \cdot 2\eta p(t)(1-p(t))\big[&(1-X(t))\left((1-\theta(t))^2 - \mathbb{E}_s[(1-s)^2]\right) \\ &+ X(t)\left((1-\theta(t))^2 - (1-s_t)^2\right)\big]. \end{aligned}$$

In summary, we derive the following stochastic dynamics under AI assistance:

$$\begin{aligned} X(t) &\sim \mathrm{Bern}(p(t)), \\ s_t &\sim \mu_a, \\ \theta(t+1) &= \theta(t) + 2\eta\theta(t)(1-\theta(t))[(1-X(t))(1-\theta(t)) + \Delta \cdot X(t) \cdot (\theta_d - \theta(t))] \\ p(t+1) &= p(t) + \kappa \cdot 2\eta p(t)(1-p(t))\big[(1-X(t))\left((1-\theta(t))^2 - \mathbb{E}_s[(1-s)^2]\right) \\ &\quad + X(t)\left((1-\theta(t))^2 - (1-s_t)^2\right)\big]. \end{aligned}$$

The key difference from Dynamics (8) is the introduction of the random variable s_t and the resulting modification to the update rule for p.

Like in Section 5.2, we then track this stochastic dynamics to its expectation. Since the randomness of s_t is independent from $(X(t), \theta(t), p(t))$, we have that

$$\begin{aligned} &\mathbb{E}_{X(t),s_t}[p(t+1) \mid \theta(t), p(t)] \\ &= \mathbb{E}_{X(t)}\left[\mathbb{E}_{s_t}[p(t+1) \mid \theta(t), p(t), X(t)]\right] && \text{(Law of iterated expectations)} \\ &= \mathbb{E}_{X(t)}\left[p(t) + 2\eta\kappa p(t)(1-p(t))\left((1-\theta(t))^2 - \mathbb{E}_{s_t}[(1-s_t)^2]\right)\right] && \text{(Linearity of expectation)} \\ &= p(t) + 2\eta\kappa p(t)(1-p(t))\left((1-\theta(t))^2 - \mathbb{E}_{s_t}[(1-s_t)^2]\right). && \text{(Independence from } X(t)) \end{aligned}$$

As the step size $\eta \to 0$, this expected dynamics tracks to ODE (5).

7.2 Derivation of ODE (6) for noisy update of delegations

With an inaccurate belief $\widetilde{\theta}_a$, the main difference is on the update of the delegation level when the learner does the task by its own, i.e., if $X(t) = 0$, the learner has an accurate belief $(1-\widetilde{\theta}_a)^2$ on the loss of AI's output. Otherwise, if $X(t) = 1$, we still assume that the learner has an accurate belief $(1-\theta(t))^2$ on the loss of its output. Together, the learner updates its delegation level as follows:

$$\begin{aligned} p(t+1) := p(t) + \kappa \cdot 2\eta p(t)(1-p(t))\big[&(1-X(t))\left((1-\theta(t))^2 - (1-\widetilde{\theta}_a)^2\right) \\ &+ X(t)\left((1-\theta(t))^2 - (1-\theta_a)^2\right)\big]. \end{aligned}$$

Then its corresponding expected dynamics is:

$$\begin{aligned} \mathbb{E}[p(t+1) \mid (\theta(t), p(t))] = p(t) + \kappa \cdot 2\eta p(t)(1-p(t))\big[&(1-p(t))\left((1-\theta(t))^2 - (1-\widetilde{\theta}_a)^2\right) \\ &+ p(t)\left((1-\theta(t))^2 - (1-\theta_a)^2\right)\big]. \end{aligned}$$

As the step size $\eta \to 0$, this expected dynamics tracks to ODE (6).

7.3 Derivation of ODE (7) for asymmetric update of delegations

With a risk-averse factor $\alpha > 0$, the update for the delegation level extends to be: If $(1-\theta(t))^2 < (1-\theta_a)^2$, the step size is $\kappa\alpha$ rather than κ. This corresponds to the following dynamics:

$$p(t+1) := p(t) + \kappa \cdot 2\eta p(t)(1-p(t)) \left[[(1-\theta(t))^2 - (1-\theta_a)^2]_+ - \alpha[(1-\theta_a)^2 - (1-\theta(t))^2]_+ \right].$$

As the step size $\eta \to 0$, this dynamics tracks to ODE (7).

7.4 Extension with alternative performance loss

Another extension considers the case in which the teacher penalizes the learner upon detecting that an output was generated by AI, serving as an intervention to reduce AI reliance [31]. To study this effect, let $q \in [0,1]$ denote the probability that AI-generated output is detected, and assume that upon detection the performance loss of AI output is $|1-\theta_a|$. This alternative loss exceeds $(1-\theta_a)^2$ and thus acts as an explicit penalty for AI delegation.

Under this model, the mechanism of learning in Section 5.1 extends to be:

- **(Task completion stage)** Learner delegates the task to AI with probability $p(t)$, which sets $X(t) = 1$; and does the task by itself with probability $1 - p(t)$, which sets $X(t) = 0$.
- **(Evaluation stage)** Learner submits the output to the teacher. The teacher gives a loss ℓ_t to the learner. If the output is done by the learner itself, $\ell_t = (1-\theta(t))^2$. Otherwise, if the output is done by AI, the teacher successfully detects with probability q, which sets $Y(t) = 1$. If $Y(t) = 1$, $\ell_t = |1-\theta_a|$; while if $Y(t) = 0$, $\ell_t = (1-\theta_a)^2$. Then

$$\ell_t := (1-X(t))(1-\theta(t))^2 + X(t) \cdot \big((1-Y(t))(1-\theta_a)^2 + Y(t)|1-\theta_a|\big).$$

- **(Update of skill)** If $X(t) = 0$, the learner skill (intrinsically) updates towards the target 1 according to the loss $\ell_t = (1-\theta(t))^2$:

$$\theta(t+1) := \theta(t) + 2\eta\theta(t)(1-\theta(t))^2$$

  Otherwise, if $X(t) = 1$, the learner skill decays towards θ_d due to non-practice:

$$\theta(t+1) := \theta(t) + 2\eta\Delta\theta(t)(1-\theta(t))(\theta_d - \theta(t)).$$

- **(Update of delegation level)** If $X(t) = 0$, assume that the learner has an accurate belief $(1-q)(1-\theta_a)^2 + q|1-\theta_a|$ on the expected performance loss of AI's output. Otherwise, if $X(t) = 1$, assume that the learner has an accurate belief $(1-\theta(t))^2$ on the loss of its output. Together with the obtained loss ℓ_t, the learner updates its delegation level as follows:

$$\begin{aligned} p(t+1) := p(t) + \kappa \cdot 2\eta p(t)(1-p(t)) \big[&(1-X(t)) \left((1-\theta(t))^2 - (1-q)(1-\theta_a)^2 - q|1-\theta_a|\right) \\ &+ X(t) \left((1-\theta(t))^2 - (1-Y(t))(1-\theta_a)^2 - Y(t)|1-\theta_a|\right) \big] \end{aligned}$$

In summary, we derive the following stochastic dynamics under AI assistance:

$$\begin{aligned}
X(t) &\sim \text{Bern}(p(t)),\\
Y(t) &\sim \text{Bern}(q),\\
\theta(t+1) &= \theta(t) + 2\eta\theta(t)(1-\theta(t))[(1-X(t))(1-\theta(t)) + \Delta\cdot X(t)\cdot(\theta_d - \theta(t))]\\
p(t+1) &= p(t) + \kappa\cdot 2\eta p(t)(1-p(t))\big[(1-X(t))\left((1-\theta(t))^2 - (1-q)(1-\theta_a)^2 - q|1-\theta_a|\right)\\
&\qquad + X(t)\left((1-\theta(t))^2 - (1-Y(t))(1-\theta_a)^2 - Y(t)|1-\theta_a|\right)\big].
\end{aligned}$$

The key difference from Dynamics (8) is the introduction of the random variable $Y(t)$ and the resulting modification to the update rule for p.

Like in Section 5.2, we then track this stochastic dynamics to its expectation. We have that

$$\begin{aligned}
&\mathbb{E}_{X(t),Y(t)}[p(t+1)\mid\theta(t),p(t)]\\
&=\mathbb{E}_{X(t)}\left[\mathbb{E}_{Y(t)}[p(t+1)\mid\theta(t),p(t),X(t)]\right] \quad \text{(Law of iterated expectations)}\\
&=\mathbb{E}_{X(t)}\big[p(t)+\kappa\cdot 2\eta p(t)(1-p(t))\big[(1-X(t))\left((1-\theta(t))^2-(1-q)(1-\theta_a)^2-q|1-\theta_a|\right)\\
&\quad + X(t)\left((1-\theta(t))^2-(1-q)(1-\theta_a)^2-q|1-\theta_a|\right)\big]\big] \quad \text{(Linearity of expectation)}\\
&=p(t)+2\eta\kappa p(t)(1-p(t))\left((1-\theta(t))^2-(1-q)(1-\theta_a)^2-q|1-\theta_a|\right). \quad \text{(Independence from } X(t))
\end{aligned}$$

As the step size $\eta\to 0$, this dynamics tracks to ODE

$$\dot{p} = \kappa p(1-p)\left((1-\theta)^2-(1-q)(1-\theta_a)^2-q|1-\theta_a|\right). \tag{17}$$

Intuitively, increasing q raises the expected cost of delegating to AI by making AI usage more likely to incur a harsher loss, thereby weakening the incentive to delegate and shifting trajectories toward sustained practice. Figure 6c illustrates this effect. This demonstrates the effectiveness of such penalties as an intervention mechanism.

8 Usability of our model

We demonstrate the usability of our framework by providing an explicit procedure for deriving parameters θ_a, κ, Δ, together with a worked numerical example.

Consider an experimental setup in which a learner decides whether to delegate tasks to AI or perform them independently over T sections. For each section t, the following data are observed: an indicator $X(t)\in\{0,1\}$, where $X(t)=1$ denotes delegation to AI; an evaluation loss $\ell_t\in[0,1]$ for the learner's output as assessed by an teacher; and an evaluation loss $\ell_{a,t}\in[0,1]$ for the output generated by AI alone. Here, all loss values are normalized to lie in $[0,1]$. We note that such data have been collected in [13] and are partially reported in Figure 41 of that paper.

In addition, applying our model requires access to the learner's delegation level $p(t)\in[0,1]$, which can be collected by querying learners about their willingness to delegate to AI prior to each section.

Deriving AI skill. Using the evaluation losses $\{\ell_{a,t}\}_{t=1}^{T}$, we estimate the (effective) AI skill as

$$\theta_a := 1 - \sqrt{\frac{1}{T}\sum_{t=1}^{T}(1-\ell_{a,t})^2}.$$

Deriving the learning rate. Let

$$A := \{t \in [T] : X(t) = 0\} = \{t_1, t_2, \ldots, t_m\}$$

be the set of sections in which the learner performs the task independently. Using the observed losses ℓ_t, we estimate the learner's skill at time $t \in A$ by

$$\theta(t) := 1 - \sqrt{\ell_t}.$$

For two consecutive sections $t_j, t_{j+1} \in A$ such that $t_{j+1} = t_j + 1$, our model implies

$$\theta(t+1) = \theta(t) + \eta\,\theta(t)(1-\theta(t))^2.$$

Hence, we estimate the learning rate at time t_j as

$$\eta_{t_j} := \frac{\theta(t_j+1) - \theta(t_j)}{\theta(t_j)(1-\theta(t_j))^2}.$$

Let $B := \{t \in A : t+1 \in A\}$ denote the set of such consecutive self-practice sections. Averaging over B yields the learning-rate estimate

$$\eta := \frac{1}{|B|} \sum_{t \in B} \eta_t.$$

Deriving the delegation rate. For $t \in B$, using the observed losses $\ell_t, \ell_{a,t}$ and reported delegation levels $p(t), p(t+1)$, our model implies $p(t+1) = p(t) + \eta\,\kappa\,p(t)(1-p(t))\,(\ell_t - \ell_{a,t})$. Thus, we estimate the delegation rate at time t by

$$\kappa_t := \frac{p(t+1) - p(t)}{\eta\,p(t)(1-p(t))(\ell_t - \ell_{a,t})}.$$

Averaging over B gives the delegation-rate estimate

$$\kappa := \frac{1}{|B|} \sum_{t \in B} \kappa_t.$$

Deriving the decay rate. For $t_j \in A \setminus (\{t_m\} \cup B)$, the learner delegates tasks to AI between sections t_j and t_{j+1}, during which skill decay occurs. Using a first-order approximation, we obtain $\theta(t_{j+1}) \approx \theta(t_j) + \eta\,\theta(t_j)(1-\theta(t_j))\,((1-\theta(t_j)) - (t_{j+1} - t_j - 1)\Delta\,\theta(t_j))$. Solving for Δ yields the estimate

$$\Delta_j := \frac{\eta\,\theta(t_j)(1-\theta(t_j))^2 - (\theta(t_{j+1}) - \theta(t_j))}{\eta\,\theta(t_j)(1-\theta(t_j))(t_{j+1} - t_j - 1)\theta(t_j)}.$$

Averaging over all such t_j gives the decay-rate estimate

$$\Delta := \frac{1}{|A \setminus (\{t_m\} \cup B)|} \sum_{t_j \in A \setminus (\{t_m\} \cup B)} \Delta_j.$$

Table 2: An illustrative example of the procedure for deriving the parameters θ_a, κ, and Δ.

t	Decision	$X(t)$	ℓ_t	$\ell_{a,t}$	$p(t)$	$\theta(t)$
1	Manual	0	0.36	0.04	0.20	0.40
2	Manual	0	0.25	0.04	0.25	0.50
3	Delegate	1	—	0.04	0.35	—
4	Delegate	1	—	0.04	0.45	—
5	Manual	0	0.30	0.04	0.40	0.45

Working example. To illustrate this procedure, we simulate a short trajectory ($T = 5$) of a learner interacting with a high-performing AI. The data is presented in Table 2.
Step 1: Deriving AI skill θ_a. The AI loss is consistent at $\ell_{a,t} = 0.04$. The implied skill is:

$$\theta_a = 1 - \sqrt{0.04} = 1 - 0.2 = \mathbf{0.80}.$$

Step 2: Deriving learning rate η. We observe consecutive manual practice at $t = 1, 2$.

- Skill growth: $\theta(2) - \theta(1) = 0.50 - 0.40 = 0.10$.
- Theoretical growth term: $\theta(1)(1 - \theta(1))^2 = 0.4(0.6)^2 = 0.144$.
- Estimate: $\eta = 0.10/0.144 \approx \mathbf{0.694}$.

Step 3: Deriving delegation rate κ. Using the transition from $t = 1$ to $t = 2$:

- Change of delegation level: $p(2) - p(1) = 0.25 - 0.20 = 0.05$.
- Performance gap: $\ell_1 - \ell_{a,1} = 0.36 - 0.04 = 0.32$.
- Denominator term: $\eta p(1)(1 - p(1))(\text{gap}) = 0.694 \cdot 0.2 \cdot 0.8 \cdot 0.32 \approx 0.0355$.
- Estimate: $\kappa = 0.05/0.0355 \approx \mathbf{1.41}$.

Step 4: Deriving decay rate Δ. We observe a manual session at $t = 5$ following a delegation block ($t = 3, 4$).

- Delegation duration: $k = 5 - 2 - 1 = 2$ steps.
- Observed change: $\theta(5) - \theta(2) = 0.45 - 0.50 = -0.05$.
- Expected growth (had they not delegated): $\eta\theta(2)(1 - \theta(2))^2 = 0.694 \cdot 0.5 \cdot 0.25 \approx 0.0868$.
- Numerator (Growth - Change): $0.0868 - (-0.05) = 0.1368$.
- Denominator term: $k\eta\theta(2)^2(1 - \theta(2)) = 2 \cdot 0.694 \cdot 0.25 \cdot 0.5 \approx 0.1735$.
- Estimate: $\Delta = 0.1368/0.1735 \approx \mathbf{0.79}$.

Prediction of convergence. With parameters $\theta_a = 0.8, \kappa = 1.41, \Delta = 0.79$, we calculate the precise location of the saddle point and the separatrix. At $t = 5$, the learner's state is $(\theta_0 = 0.45, p_0 = 0.4)$. Using these information, we can also predict the convergence behavior of the learner.

First, we compute the saddle point $(\theta^\dagger, p^\dagger)$:

$$\theta^\dagger = 0.8, \quad p^\dagger = \frac{1 - 0.8}{1 - (1 - 0.79)0.8} = \frac{0.2}{0.832} \approx 0.24.$$

Next, we derive the shape parameters for the stable manifold approximation $\widetilde{\psi}$. The tangent slope at the saddle is $m^\dagger \approx 0.46$. The left-branch exponent is $\beta_l = 1.41(1 - (0.2)^2) \approx 1.35$. Computing the breakpoint θ_l yields $\theta_l = \theta^\dagger = 0.8$, meaning the first branch of the piecewise function (16) applies to the learner's current state $\theta_0 = 0.45$. Substituting these values into $\widetilde{\psi}$, we determine the critical delegation threshold:

$$\widetilde{\psi}(0.45) = p^\dagger \left(\frac{0.45}{\theta^\dagger}\right)^{\beta_l} \approx 0.24 \left(\frac{0.45}{0.8}\right)^{1.35} \approx 0.11.$$

At $t = 5$, the learner's delegation level is $p_0 = 0.40$, which is strictly higher than the critical threshold $p_{crit} \approx 0.11$. Thus, the learner locates inside the low-skill basin. Consequently, we predict that despite having moderate initial skill, the learner is over-reliant on the AI relative to their learning trajectory and will eventually converge to the low-skill equilibrium $(\theta \to 0, p \to 1)$.

9 Conclusions, limitations, and future work

We study the impact of AI on learning through a dynamical system that captures how short-term delegation decisions shape long-run human skill. By coupling delegation and skill updates through a shared objective of instantaneous loss minimization, the framework reveals a structural source of heterogeneous AI effects. Our results characterize when skill converges to a low-skill equilibrium stabilized by persistent delegation, show that this degradation intensifies as AI capability increases, and establish that performance gains from AI assistance are transient rather than cumulative. Together, these findings provide a mechanistic account of *cognitive debt* observed in recent empirical work: AI assistance can improve short-term performance while reducing long-run skill by diminishing practice. These results do not imply that AI is inherently harmful; rather, its long-run effects depend on how delegation evolves relative to opportunities for practice.

From a policy perspective, our findings highlight the importance of sustaining practice and redesigning incentives, such as delaying high-quality AI assistance, rewarding independent problem solving, or valuing learning trajectories, to align short-term performance with long-term skill development as AI capability advances.

Our analysis focuses on a minimal, deterministic dynamical system capturing average-case behavior. This abstraction enables sharp theoretical results but omits several real-world factors. First, we do not model post-delegation verification or corrective behavior, which may partially mitigate skill degradation. Second, the dynamics abstract away stochasticity from noisy AI outputs, learner exploration, or evaluation variance. Third, we consider a single task with a one-dimensional skill representation; in practice, learning spans multiple tasks with transfer and substitution effects, and AI may accelerate some skills while eroding others. Extending the framework to stochastic, multi-task, or partially observed settings that incorporate verification and strategic effort allocation is an important direction for future work.

Acknowledgments

We thank Mariia Eremeeva for useful discussions. This work was funded by NSF Award CCF-2112665.

References

[1] Daron Acemoglu, David Autor, and Simon Johnson. Building pro-worker artificial intelligence. Working Paper 34854, National Bureau of Economic Research, February 2026.

[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] 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.

[5] L. Elisa Celis, Amit Kumar, Anay Mehrotra, and Nisheeth K. Vishnoi. Bias in evaluation processes: An optimization-based model. In *NeurIPS*, 2023.

[6] Fabrizio Dell'Acqua, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine C. Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *SSRN Electronic Journal*, 2023.

[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] 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.

[10] K. Anders Ericsson, Ralf Thomas Krampe, Clemens Tesch-Romer, Catherine Ashworth, Gregory Carey, Robert J. Crutcher, Janet Grassia, Reid Hastie, Stefanie Heizmann, Charles Judd, Ronald Kellogg, Robert Levin, Clayton H. Lewis, William Oliver, Peter G. Poison, Robert Rehder, Kurt Schlesinger, Vivian I. Schneider, and James Wilson. The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100:363–406, 1993.

[11] Thomas L. Griffiths, Charles Kemp, and Joshua B. Tenenbaum. Bayesian models of cognition, 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] 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.

[14] Abhishek Kumar, Prasanna Sattigeri, and Tom Fletcher. Semi-supervised learning with gans: Manifold invariance with improved inference. In *Neural Information Processing Systems*, 2017.

[15] 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.

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

[17] 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.

[18] Dylan R. Nelson and Joey Cheung. Non-linear dynamics and chaos. 2007.

[19] 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.

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

[21] 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.

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

[23] 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.

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

[25] 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.

[26] Salah Rifai, Yann Dauphin, Pascal Vincent, Yoshua Bengio, and Xavier Muller. The manifold tangent classifier. In *Neural Information Processing Systems*, 2011.

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

[28] Hang Shao, Abhishek Kumar, and P. Thomas Fletcher. The riemannian geometry of deep generative models. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 428–4288, 2018.

[29] 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.

[30] Stack Overflow. Stack Overflow developer survey 2025. Online report, 2025.

[31] 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.

[32] 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.

[33] Ruixun Zhang, Thomas J Brennan, and Andrew W Lo. The origin of risk aversion. *Proceedings of the National Academy of Sciences*, 111(50):17777–17782, 2014.