Skill Atrophy and AI Productivity Measurement

Tommaso Bondi* Gentry Johnson†

February 2026

Abstract

How should we measure the productivity effects of generative AI? Recent experimental studies document substantial short-run gains. We show that measuring AI's long-run effects introduces two structural biases when adoption affects skill formation. In a dynamic model where workers learn by doing, the effects of AI delegation depend on AI's *pedagogical quality*: the degree to which AI-assisted work contributes to skill formation. When AI substitutes for cognitive effort, two biases arise. First, even within-worker comparisons are distorted: state-conditional productivity gains diverge from path-based comparisons because current skill is endogenous to past AI use (state-path divergence). Second, as adoption spreads, non-users become a degraded counterfactual because mentorship, spillovers, and training environments deteriorate, causing cross-sectional estimates to overstate lifetime effects (spillover bias). We characterize cohort effects and vintage wage premiums, and analyze welfare and policy when decentralized adoption is inefficient. An illustrative calibration anchored to experimental estimates (Bastani et al., 2025; Shen and Tamkin, 2026) suggests the biases could be economically large.

*Cornell University. Email: tbondi@cornell.edu.

†Amazon Web Services. Email: gentry.a.johnson@gmail.com. This work was performed outside of Amazon Web Services and does not relate to the author's role at the company.

We thank Ajay Agrawal, Guy Aridor, Ron Berman, Luis Cabral, Judy Hanwen Shen, Brett Hollenbeck, Vrinda Kadiyali, Jura Liaukonytė, Xueming Luo, Emaad Manzoor, John McHale, Ivan Png, Omid Rafieian, Alex Tamkin, Michael Waldman, and Nathan Yang for helpful comments and suggestions.

1 Introduction

Generative AI has delivered striking short-run productivity gains across knowledge-intensive work. Customer service agents resolve more tickets per hour, consultants complete analyses faster, junior developers ship code more quickly. These gains are especially pronounced for less-skilled workers, compressing the productivity distribution. But the estimates are almost exclusively short-run, measuring output over weeks or months rather than the years across which expertise develops.

This paper shows that short-run productivity estimates can target the wrong object for welfare. The core problem is an estimand mismatch: standard designs measure the value of AI *conditional on the worker's current skill*, but current skill is itself shaped by past AI use. The welfare-relevant comparison is to the skill path that would have obtained absent adoption. When AI affects skill formation, these objects diverge, and the divergence can reverse sign.

We model workers who learn by doing and can delegate tasks to AI. The key parameter is *pedagogical quality*, denoted μ: the degree to which AI-assisted work contributes to skill formation relative to unassisted work. Autocomplete interfaces that minimize user effort correspond to low μ; Socratic tutors that prompt reflection correspond to high μ. The model admits closed-form expressions for the shadow value of human capital and the sensitivity of steady-state skill to adoption intensity, yielding sharp comparative statics despite the fully dynamic structure.

When $\mu \neq 1$, standard empirical designs *condition on an endogenous state* rather than recovering the welfare-relevant counterfactual skill path. When $\mu < 1$, this overstates AI's contribution; when $\mu > 1$, it understates long-run benefits. The measurement problem is general; the direction of bias is parameter-dependent.

Two structural biases emerge. The first, *state-path divergence*, operates at the individual level. A worker who has relied on AI for years possesses lower skill than the counterfactual path would have produced. Measuring AI's value against their current, atrophied skill overstates gains; the *welfare-relevant comparison* is to the skill they would have had absent AI. As skills atrophy, AI appears increasingly indispensable, even holding AI's capabilities fixed, because the outside option has deteriorated. This bias requires only $\mu < 1$; no externalities or cross-agent interactions.

The second bias, *spillover bias*, grows with industry-level AI saturation. As adoption spreads, non-users face degraded learning environments: reduced mentorship from seniors who delegate instructional tasks to AI, weaker peer effects from colleagues who accumulate

less shareable knowledge, and curricula redesigned for AI-assisted workflows. Comparing AI users to these degraded non-users further overstates the benefits of adoption. The bias is zero before adoption affects non-users and grows monotonically with diffusion.

A mechanism distinctive to generative AI interacts with these biases. Unlike calculators or spreadsheets, generative AI learns from human-generated content. When workers delegate tasks, they produce less original content and what they produce reflects less skill, degrading training data for future AI systems. Appendix A develops a full microfoundation for this "skill-data feedback loop," showing that it partially stabilizes human capital but cannot accelerate recovery: degradation is fast but rebuilding expertise takes years. The feedback also generates a novel training data externality that compounds the human capital externality analyzed in Section 5.

When $\mu < 1$, the model generates testable predictions for wages and inequality. Pre-AI cohorts command growing wage premiums as skilled workers retire: scarcity value rises for skills that new workers cannot easily acquire. Wage inequality follows a hump-shaped path over time: rising as skill gaps between pre-AI and post-AI cohorts widen, then falling as pre-AI workers retire and the workforce converges to uniformly lower skills. High-ability workers bear compounding losses: foregone skill development prevents them from reaching their potential, and competition with AI erodes returns to the skills they do acquire.

When human capital generates spillovers beyond its private value, decentralized adoption exceeds the social optimum. Optimal Pigouvian taxes internalize both the skill externality and the training data externality. A distinctive implication is that optimal policy may reduce *measured* productivity while improving welfare, because the metrics themselves are biased. Training mandates (requiring some work be performed without AI, analogous to manual flight hours for pilots or unassisted surgical procedures for residents) offer a practical alternative when monitoring AI use is difficult.

We anchor the model to experimental evidence: Bastani et al. (2025) find GPT-4 access reduces subsequent math performance by 17%, implying $\mu \approx 0.83$; Shen and Tamkin (2026) find a strikingly similar 17% reduction among software developers, with different populations and different tasks yielding the same point estimate. To illustrate magnitudes: at $\mu = 0.5$, steady-state skills fall 20% below the no-adoption counterfactual; measurement overstates AI's welfare contribution by 11% at year 10; vintage premiums for pre-AI workers reach 25% in steady state. At $\mu = 1$, both biases vanish. For $\mu > 1$, they reverse sign: cross-sectional estimates *understate* long-run benefits. This yields a sharp prediction: if $\mu \geq 1$, effect sizes should grow over time rather than shrink.

Early evidence beyond these anchoring estimates favors $\mu < 1$. METR (2025) find experienced developers are slower with AI tools yet believe AI increases their productivity, consistent with skill atrophy impairing self-assessment. Budzyń et al. (2025) document endoscopist deskilling: after three months of AI-assisted colonoscopy, physicians' unassisted adenoma detection rates fell from 28.4% to 22.4%, a 21% relative decline and the first clinical evidence of AI-induced skill atrophy affecting patient outcomes.

del Rio-Chanona et al. (2024) find Stack Overflow activity declined sharply after ChatGPT's release; Burtch et al. (2024) show newer users exited fastest, consistent with failing to build query-formulation skills. Ong and Png (2026) find that deskilling technologies increase labor supply by lowering the skill threshold for entry, a labor market consequence our model does not capture but that reinforces the policy case for training mandates. These patterns are inconsistent with $\mu \geq 1$.[1]

Our contribution. To summarize, we show that common productivity estimands condition on an endogenous state and can therefore have the wrong sign for welfare when AI affects skill formation. We provide an exact decomposition of the bias into a state-gap component and a spillover component, characterize conditions under which measured effects reverse sign, derive welfare and policy implications, and connect the analysis to a training data externality distinctive to generative AI.

Roadmap. Section 1.1 reviews related literature. Section 2 develops the model. Section 3 characterizes equilibrium. Section 4 analyzes mismeasurement, cohort effects, and an illustrative calibration. Section 5 examines welfare and policy. Section 6 concludes.

1.1 Related Literature

For recent surveys, see Acemoglu (2024) and Agrawal et al. (2026). This paper contributes to three literatures. The task-based framework of Acemoglu and Restrepo (2018, 2020) models automation as machines performing tasks previously done by humans, taking human capital as fixed. We introduce a different margin: task frameworks treat skills as a stock determining productivity (Gibbons and Waldman, 2004); we show tasks are also inputs into skill production, so automation can reduce productivity on *all* tasks, not just those directly displaced. Agrawal et al. (2026) develop task-based models where AI augments rather than replaces workers, emphasizing complementarities between AI prediction and

[1] Longitudinal evidence from related technologies corroborates: Dahmani and Bohbot (2020) find GPS use predicts steeper decline in spatial memory over three years ($r = -0.52$ to -0.68); Casner et al. (2014) find cognitive flying skills degrade with heavy autopilot use.

human judgment; our framework complements theirs by showing that even augmentation can degrade the human capital stock when it substitutes for learning – AI may complement the *use* of judgment while substituting for its *development*. Acemoglu (2024) estimates TFP gains of 0.5–0.7% over ten years, assuming no skill atrophy.

A growing empirical literature documents short-run productivity effects: Noy and Zhang (2023) for writing, Peng et al. (2023) for coding, and Dell'Acqua et al. (2023) identifying a "jagged frontier" where AI helps on some tasks but hurts on others. Otis et al. (2023) conduct a field experiment with Kenyan entrepreneurs and find heterogeneous effects: AI mentorship increased performance by 20% for high performers but *decreased* it by 10% for low performers. Gaessler and Piezunka (2023) find chess computers *helped* players improve ($\mu \geq 1$), plausibly because chess provides immediate, objective feedback; but more recent work documents deskilling: endoscopists (Budzyń et al., 2025), robot-assisted workers (Beane, 2019), and knowledge workers (Lee et al., 2025; Dell'Acqua, 2022). Our contribution is to show that productivity and skill formation are jointly determined: measuring one without the other conflates short-run gains with long-run costs.

Our model builds on human capital theory (Becker, 1962), learning-by-doing (Arrow, 1962; Lucas, 1988), and the technology of skill formation (Cunha and Heckman, 2007).[2] Luo et al. (2025) find platforms may optimally restrict AI access to preserve human capital, a market-based analog to the mandates we analyze. We extend Arrow's insight that production generates knowledge as a byproduct to show that AI can sever this link.

Recent work examines how AI threatens training and skill transmission. Garicano and Rayo (2025) show apprenticeships become unviable when AI automates entry-level work: if juniors generate no billable output, the economic foundation of apprenticeship collapses. Ide (2025) develops a growth model where AI reduces opportunities for tacit knowledge acquisition. Beane (2019) provide evidence that robotic surgery made trainees "optional," reducing hands-on practice tenfold. Brynjolfsson et al. (2025b) document early employment effects of AI, finding heterogeneous impacts across occupations. Our contribution is distinct: we study learning *within* jobs rather than access *to* jobs. The mechanisms compound: policies preserving entry-level employment will fail if the resulting work is pedagogically hollow.

[2] Cunha and Heckman (2007) emphasize *self-productivity* (skills acquired early raise the productivity of later investment) and *dynamic complementarity* (early investment raises the return to later investment). Both properties hold in our framework.

2 The Model

2.1 Environment and Primitives

Time is discrete, indexed by $t \in \{0, 1, 2, \ldots\}$. A unit mass of firms, indexed by $i \in [0, 1]$, each employs one worker. Lowercase variables (h, α) denote individual quantities; uppercase (H, A) denote aggregates.

Each period, production requires completing a unit continuum of tasks indexed by $j \in [0, 1]$. Each task can be performed either by the worker or by AI. When the worker performs task j, output from that task is $y_i(j, t) = h_{i,t} \cdot e_{i,t}(j)^\gamma$, where $h_{i,t} \geq 0$ is the worker's human capital, $e_{i,t}(j) \geq 0$ is effort allocated to task j, and $\gamma \in (0, 1)$ governs the returns to effort. When AI performs task j, output is $y_i(j, t) = A_t \cdot g(j)$, where $A_t > 0$ is AI productivity and $g : [0, 1] \to (0, 1]$ is the AI capability function satisfying $g(0) = 1$, $g(1) \in (0, 1)$, $g'(j) < 0$, and $|g'|$ bounded away from zero (tasks differ meaningfully in AI suitability).

The condition $g'(j) < 0$ encodes comparative advantage: AI is more capable at routine, well-defined tasks (low j) than at complex, judgment-intensive tasks (high j). This ordering is without loss of generality given the continuum structure; we are simply labeling tasks by their amenability to AI automation.

Workers face an effort constraint: total effort across all worker-performed tasks is normalized to unity. When a firm adopts AI at intensity $\alpha \in [0, 1]$, it delegates tasks in $[0, \alpha]$ to AI while the worker performs tasks in $(\alpha, 1]$. Standard optimization shows the worker spreads effort uniformly across performed tasks, yielding worker output $h(1 - \alpha)^{1-\gamma}$.[3]

Substituting, period output takes the tractable form

$$Y(h, \alpha; A) = A \cdot G(\alpha) + h \cdot (1 - \alpha)^{1-\gamma} \tag{1}$$

where $G(\alpha) \equiv \int_0^\alpha g(j)\, dj$ is cumulative AI output, with $G'(\alpha) = g(\alpha)$ and $G''(\alpha) = g'(\alpha) < 0$. The first term captures AI's contribution; the second captures the worker's. The exponent $1 - \gamma < 1$ reflects effort concentration: when workers perform fewer tasks, effort is spread less thinly. The function is linear in h, strictly concave in α, and satisfies $\partial Y / \partial \alpha \to -\infty$ as $\alpha \to 1^-$, ensuring interior optima.

[3] With per-task output $he(j)^\gamma$ and effort constraint $\int_\alpha^1 e(j)dj = 1$, uniform effort $e(j) = 1/(1 - \alpha)$ yields total output $\int_\alpha^1 h[1/(1-\alpha)]^\gamma dj = h(1-\alpha)^{1-\gamma}$. With binary adoption $\alpha \in \{0, 1\}$, the results are qualitatively similar but adoption is "lumpy": firms either fully adopt or abstain. The continuum smooths this and allows partial adoption.

2.2 Human Capital Dynamics

Human capital evolves according to

$$h_{t+1} = (1-\delta)h_t + \lambda \cdot L(\alpha_t, h_t; \mu) \tag{2}$$

where $\delta \in (0,1)$ is depreciation, $\lambda > 0$ governs learning intensity, and $L(\alpha, h; \mu)$ is the learning function. AI use at t affects skill through the transition to h_{t+1}; current output Y_t depends on h_t and α_t contemporaneously. The learning function is

$$L(\alpha, h; \mu) = [(1-\alpha) + \mu \cdot \alpha] \cdot \varphi(h) \tag{3}$$

where $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies regularity conditions below, and $\mu \geq 0$ is *pedagogical quality*.

The effective learning rate $\ell(\alpha) \equiv 1 - (1-\mu)\alpha$ is positive for all $\alpha \in [0,1]$ when $\mu \geq 0$. Our main results focus on $\mu \in [0,1)$: when $\mu < 1$, AI substitutes for learning (skill atrophy); when $\mu \geq 1$, AI augments learning (skill enhancement).

Assumption 1 (Learning Capacity)**.** The function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is twice continuously differentiable, strictly positive, bounded above, log-concave, with $\varphi'(h) < 0$ and $\lim_{h\to\infty} \varphi(h) = 0$. The slope φ' is bounded: $\sup_h |\varphi'(h)| < \delta/\lambda$.[4]

The first properties capture diminishing returns: experts face smaller learning gains as most relevant knowledge has been acquired. Log-concavity ($\varphi''/\varphi \leq (\varphi'/\varphi)^2$) ensures that the learning function's curvature is well-behaved: high-ability workers (who scale φ by θ) always forgo more learning from delegation than low-ability workers, ruling out pathological cases where ability and AI are complements at low skill. Standard functional forms including $\varphi(h) = c/(1+h)^k$ for $k \geq 1$ satisfy this condition. The bound on φ' is a mild contraction condition: it requires that the learning function is not so steep that learning feedback overwhelms depreciation. When this fails, the dynamics can diverge – a worker who falls slightly behind learns so much faster that she overshoots, generating oscillations rather than convergence. This is not economically interesting; it is an artifact of steep learning functions and is ruled out in all standard human capital models (Stokey and Lucas, 1989). For our calibrated $\varphi(h) = 0.2/(1+h)$, the bound requires $\delta > 0.03$, which holds for any empirically relevant depreciation rate.

The key property: $\partial L/\partial \alpha = (\mu - 1)\varphi(h)$, which is negative when $\mu < 1$, zero when $\mu = 1$, positive when $\mu > 1$. This derivative governs whether delegation helps or hurts skill

[4] A tractable example is $\varphi(h) = \varphi_0/(1 + h/\xi)$ for $\varphi_0, \xi > 0$.

accumulation.

The parameter μ has clear empirical content. Bastani et al. (2025) show GPT-4 access harms math learning ($\mu < 1$), but pedagogically-designed tutors that require active engagement mitigate this. Dell'Acqua (2022) document reduced cognitive effort with AI, consistent with low effective μ; Shaw and Nave (2026) find participants adopted AI outputs with minimal scrutiny even under strong incentives for correctness. We treat μ as exogenous, though it depends on AI design, workplace norms, and user incentives.[5] Competitive pressure exacerbates low-μ outcomes; Appendix B develops a formal model of endogenous μ under competitive pressure and shows our results are robust.

Settings where $\mu < 1$ is likely to hold include junior professional training, autocomplete-heavy workflows, and time-pressured environments. Settings where $\mu \geq 1$ may apply include AI tutors requiring engagement and tasks where AI feedback accelerates learning.

Remark 1 (Heterogeneous μ). In practice, μ varies across tasks and career stages. Such heterogeneity *strengthens* our results: workers using AI during low-μ phases accumulate less skill than those using it during high-μ phases, introducing additional path dependence.[6]

2.3 The Firm's Dynamic Problem

Firms maximize the present discounted value of output. The discount factor $\beta \in (0, 1)$ governs the weight on future productivity; patient firms (high β) internalize skill costs more heavily. The firm solves

$$V(h_0; A) = \max_{\{\alpha_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t Y(h_t, \alpha_t; A) \tag{4}$$

subject to the human capital law of motion (2). The value function $V(h)$ satisfies the Bellman equation

$$V(h) = \max_{\alpha \in [0,1]} \left\{ Y(h, \alpha; A) + \beta V\big((1-\delta)h + \lambda L(\alpha, h; \mu)\big) \right\}. \tag{5}$$

[5] Low μ can also reflect workers' tendency toward immediate output gains over skill development. Heckman et al. (2006) show "soft" skills like self-regulation predict labor market outcomes and are malleable through intervention, suggesting μ is itself a potential policy target.

[6] The scalar μ can be interpreted as an adoption-weighted average. Lifecycle heterogeneity is particularly relevant: if novices have low μ while experts have high μ, optimal policy may restrict AI for juniors while permitting it for seniors.

Table 1: Notation Guide

Symbol	Definition
h, H	Individual / aggregate human capital
α, $\bar{\alpha}$	Individual / aggregate AI adoption intensity
A	AI productivity level
μ	Pedagogical quality (< 1: substitutes for learning; ≥ 1: augments)
δ, λ	Depreciation rate / learning intensity
β	Discount factor
η	Spillover elasticity
ζ	AI quality adjustment rate
$\psi(H)$	Learning spillover function
$\bar{h}$	No-adoption steady-state skill: $\delta\bar{h} = \lambda\varphi(\bar{h})$
h^*	Steady-state skill under adoption
Δ^{CS}, Δ^{LR}	Cross-sectional / long-run productivity gain
Δ^{SC}, Δ^{PATH}	State-conditional / path-based welfare comparison

Note: In the representative-agent analysis (Sections 3–4), we consider symmetric equilibria where $h_i = h$ and $\alpha_i = \alpha$ for all i, so individual and aggregate variables coincide: $h = H$ and $\alpha = \bar{\alpha}$. The distinction becomes operative in Section 5 when we analyze spillovers across heterogeneous agents.

Standard results ensure V exists, is unique, and is strictly increasing and concave in h.[7] The key trade-off is dynamic: higher adoption today raises current output but, when $\mu < 1$, reduces future human capital.

Assumption 2 (Labor Market Structure)**.** Labor markets are competitive with general human capital (portable across employers). Wages equal marginal products, so workers with lower skills earn lower wages.[8]

The Bellman equation (5) takes output $Y(h, \alpha; A)$ as the firm's flow objective, internalizing skill dynamics by construction. The problem is best interpreted as a representative worker maximizing lifetime income, or equivalently a firm-worker pair bound by an implicit long-term contract.[9] The measurement results (Propositions 1–2) hold regardless of labor market structure because the counterfactual skill path remains endogenous.

[7] Existence and uniqueness follow from Stokey and Lucas (1989); human capital is bounded above by $\bar{h}$, ensuring the problem is well-behaved. Supporting lemmas appear in Appendix C.

[8] With Nash bargaining and worker bargaining power $\theta \in (0, 1)$, workers bear fraction θ of skill atrophy costs. The welfare results are unchanged; only the incidence shifts.

[9] The key inefficiency arises from spillovers across agents, not from a wedge between firm and worker incentives (Acemoglu and Pischke, 1999).

3 Equilibrium Characterization

This section characterizes equilibrium adoption and shows that the parameter μ fundamentally shapes both adoption decisions and long-run outcomes.

3.1 The Role of Pedagogical Quality

The firm's adoption decision balances immediate productivity gains against dynamic skill costs. When AI is sufficiently productive, some adoption is always optimal; complete delegation is never optimal because effort concentration on remaining tasks becomes increasingly valuable.[10]

Assumption 3 (AI Productivity)**.** AI is sufficiently productive that adoption is attractive even accounting for dynamic skill costs:

$$A \cdot g(0) - \bar{h}(1-\gamma) > \beta \bar{V}' \lambda (1-\mu)\varphi(\bar{h})$$

where $\bar{h}$ is the steady-state human capital without AI, and $\bar{V}' \equiv V'(\bar{h})$ is the marginal value of human capital at that steady state. The left side is the static marginal benefit of adoption at $\alpha = 0$; the right side is the discounted marginal learning cost. This ensures interior adoption $\alpha^* > 0$ at the boundary $(h, \alpha) = (\bar{h}, 0)$.[11]

The following lemma characterizes how pedagogical quality shapes adoption:

Lemma 1 (Role of Pedagogical Quality)**.** *Under Assumptions 1–3, the firm's optimal adoption $\alpha^*(h) \in (0,1)$ satisfies:*

(i) When $\mu < 1$, adoption generates a dynamic skill cost: $\partial\alpha^*/\partial\mu > 0$ locally around stable steady states.

(ii) When $\mu = 1$, adoption is determined purely by the static trade-off $\partial Y/\partial\alpha = 0$.

(iii) When $\mu > 1$, adoption generates a dynamic skill benefit: optimal α^* may exceed the static optimum $\arg\max_\alpha Y(h, \alpha; A)$.

[10] Formally, $\partial Y/\partial\alpha \to -\infty$ as $\alpha \to 1^-$ when $h > 0$.

[11] The inequality is stated at $(h, \alpha) = (\bar{h}, 0)$, but interiority extends along the equilibrium path under our maintained assumptions. Since the state space is compact ($h \in [h^*, \bar{h}]$) and $\varphi(h)$ is bounded above by Assumption 1, the dynamic marginal cost $\beta V'(h')\lambda(1-\mu)\varphi(h)$ is uniformly bounded over the relevant range. Assumption 3 effectively requires the static marginal benefit to exceed this uniform bound, ensuring $\alpha^*(h) > 0$ for all h on the equilibrium path.

In the substitution regime ($\mu < 1$), firms face an intertemporal trade-off. The first-order condition makes this precise:

$$\underbrace{A \cdot g(\alpha) - h(1-\gamma)(1-\alpha)^{-\gamma}}_{\text{marginal output gain from delegation}} = \underbrace{\beta V'(h') \cdot \lambda(1-\mu)\varphi(h)}_{\text{shadow cost of foregone learning}} \tag{6}$$

The left side is the static marginal benefit of delegating one more task to AI. The right side is the discounted shadow value of the learning that delegation destroys: $V'(h')$ is the marginal value of human capital in the continuation, $(1-\mu)$ governs how much learning each unit of delegation costs, and $\varphi(h)$ captures diminishing returns. When $\mu = 1$, the right side vanishes and the problem is purely static. When $\mu < 1$, the dynamic cost pushes adoption below the myopic optimum. Unlike prior work on which tasks machines perform (Autor et al., 2003; Acemoglu and Autor, 2011), automation here changes the *supply* of skills by altering how they accumulate.[12]

3.2 Steady-State Equilibria

A steady-state equilibrium is a pair (h^*, α^*) where adoption is optimal given skills, and skills are stationary given adoption. The stationarity condition

$$\delta h^* = \lambda \ell(\alpha^*)\varphi(h^*) \tag{7}$$

balances depreciation against learning, where $\ell(\alpha) \equiv 1 - (1-\mu)\alpha$ is the effective learning rate.

Lemma 2 (Steady-State Human Capital)**.** *For any adoption level α, there exists a unique steady-state skill level $h^*(\alpha)$ on the stable branch of the dynamics. When $\mu < 1$, higher adoption reduces steady-state skill: $\partial h^*/\partial \alpha < 0$. When $\mu \geq 1$, the opposite holds.*

Remark 2 (Stability). With φ strictly decreasing, the stationarity condition admits a unique steady state. Stability is guaranteed when depreciation dominates the learning feedback: $\delta > \lambda \ell(\alpha^*)|\varphi'(h^*)|$. Since $\ell(\alpha^*) \leq 1$ and φ' is bounded (Assumption 1), this follows from Assumption 1.

This yields a sharp characterization:

[12] The comparative static $\partial \alpha^*/\partial \mu > 0$ is local; global results require the contraction and curvature conditions in Assumption 1 and the model primitives. Patient firms are competitively punished in the short run (Proposition 9), and spillovers mean private returns to patience understate social returns.

Lemma 3 (Steady-State Characterization)**.** *Under Assumptions 1–3:*

(i) When $\mu < 1$, steady-state human capital satisfies $h^* < \bar{h}$ for any interior adoption $\alpha^* > 0$.

(ii) When $\mu \geq 1$, steady-state human capital satisfies $h^* \geq \bar{h}$.

Skill atrophy *requires* $\mu < 1$. When $\mu \geq 1$, skills weakly exceed $\bar{h}$ and AI's direct contribution ensures $Y^* > \bar{h}$: adoption is unambiguously beneficial. The lemma thus implies a sharp empirical test: if AI augments learning ($\mu \geq 1$), effect sizes in longitudinal studies should *grow* over time; if AI substitutes for learning ($\mu < 1$), they should shrink. Whether the relevant regime is $\mu < 1$ or $\mu \geq 1$ is first-order for policy, and the distinction cannot be resolved by short-run experiments alone.

The mechanics of atrophy are transparent from the stationarity condition. Implicitly differentiating (7) yields the sensitivity of steady-state skill to adoption:

$$\frac{dh^*}{d\alpha} = \frac{-\lambda(1-\mu)\varphi(h^*)}{\delta - \lambda\ell(\alpha^*)\varphi'(h^*)} \tag{8}$$

The denominator is positive at any stable steady state (a contraction condition). The numerator is negative when $\mu < 1$, so $dh^*/d\alpha < 0$: each unit of additional delegation reduces steady-state skill. The magnitude is governed by the ratio of the learning penalty $\lambda(1-\mu)\varphi(h^*)$ to the restoring force $\delta - \lambda\ell(\alpha^*)\varphi'(h^*)$. When learning is steep ($\varphi(h^*)$ large, as for novices) or the pedagogical penalty is severe ($(1-\mu)$ large), skill is highly sensitive to adoption.

Equilibrium is unique and globally stable:

Lemma 4 (Uniqueness and Global Stability)**.** *Under Assumptions 1–3:*

(i) There exists a steady-state equilibrium (h^*, α^*) with $h^* \in (0, \bar{h})$ and $\alpha^* \in (0, 1)$.

(ii) The steady-state equilibrium is unique.

(iii) For any initial condition $h_0 \in (0, \bar{h}]$, $(h_t, \alpha_t) \to (h^*, \alpha^*)$ as $t \to \infty$.

(iv) When $\mu < 1$ and $h_0 = \bar{h}$, the optimal paths are monotonic: $\{h_t\}$ is strictly decreasing and $\{\alpha_t\}$ is strictly increasing until convergence.

Global stability follows from the contraction condition in Assumption 1 (Appendix C). Part (iv) follows from $d\alpha^*/dh < 0$ when $\mu < 1$: starting from $h_0 = \bar{h} > h^*$, skills decline monotonically and adoption rises monotonically until convergence.

Conditional on $\mu < 1$, how do other parameters shape the equilibrium?

Corollary 1 (Comparative Statics)**.** *At a stable interior steady state with $\mu < 1$:*

(i) $\partial\alpha^*/\partial A > 0$ *and* $\partial h^*/\partial A < 0$*: higher AI productivity raises adoption and lowers skills.*

(ii) $\partial\alpha^*/\partial\beta < 0$ *and* $\partial h^*/\partial\beta > 0$*: more patient firms adopt less and maintain higher skills.*

(iii) $\partial h^*/\partial\lambda > 0$*: faster learners maintain higher skills.*

(iv) $\partial\alpha^*/\partial\mu > 0$*;* $\partial h^*/\partial\mu > 0$ *if and only if* $\alpha^* > (1-\mu)\frac{\partial\alpha^*}{\partial\mu}$.

Result (i) echoes Acemoglu and Restrepo (2018): better automation increases automation, but here it *endogenously degrades* the human capital stock. As AI capabilities improve, the model predicts not just more adoption but compounding skill erosion, a channel absent from standard task-based frameworks. Result (ii) implies short-termism exacerbates skill atrophy, suggesting industries with shorter planning horizons (e.g., startups, consulting) face larger losses than those with longer ones (e.g., medicine, engineering). Result (iii) implies occupations where learning-by-doing is central face the largest stakes; occupations where expertise is codified face smaller ones. Result (iv) reflects offsetting forces: higher μ directly raises h^* but induces more adoption, which lowers it. Appendix B develops a model of firm heterogeneity and shows impatient firms gain market share during the transition, potentially driving out patient firms. Appendix B.3 verifies robustness to alternative functional forms.

4 Mismeasurement of AI Productivity

Standard productivity studies estimate the causal effect of AI on output holding current skill fixed. This section shows that this estimand diverges from welfare-relevant comparisons when skill is endogenous to past AI use, and the divergence can reverse sign. We define the relevant counterfactual objects, characterize each bias, and show conditions under which measured effects reverse sign. All proofs appear in Appendix C.

4.1 State-Path Divergence

The most fundamental measurement problem requires no cross-agent externalities. When AI affects skill formation, *path dependence in human capital* causes state-conditional productivity gains to diverge from welfare-relevant path comparisons at the individual level.

Definition 1 (State-Conditional vs. Path Counterfactuals)**.** The *state-conditional counterfactual* holds human capital fixed: $\Delta_t^{SC} = Y(h_t^U, \alpha_t) - Y(h_t^U, 0)$. The *path counterfactual* compares lifetime output under adoption versus the no-adoption path: $\Delta^{PATH}(\tilde{\beta}) = \sum_{t=0}^{\infty} \tilde{\beta}^t [Y(h_t^U, \alpha_t) - Y(h_t^{NA}, 0)]$, where $\tilde{\beta}$ is the evaluator's discount factor.

Two discount factors appear: β (the firm's, determining adoption) and $\tilde{\beta}$ (the evaluator's, determining welfare). When $\tilde{\beta} = \beta$, revealed preference implies $\Delta^{PATH}(\beta) \geq 0$; under more patient evaluation ($\tilde{\beta} > \beta$), Δ^{PATH} can be negative even though adoption was privately optimal.[13]

The state-conditional gain Δ_t^{SC} overstates AI's welfare contribution because it conditions on current skill h_t^U rather than the counterfactual h_t^{NA}. Most empirical implementations recover Δ^{SC}: the effect of turning AI "on" at a given skill level, whether through explicit controls for experience and tenure or implicitly by comparing the same worker before and after adoption. When the treatment changes the state variable, the welfare-relevant object is the total effect along the counterfactual path.

Proposition 1 (State-Path Divergence)**.** *Suppose $\mu < 1$. Then:*

(i) In steady state, the state-conditional gain is $\Delta_\infty^{SC} = A \cdot G(\alpha^*) - h^*[1 - (1 - \alpha^*)^{1-\gamma}]$, with $\partial \Delta_\infty^{SC} / \partial h^* < 0$: the bias is larger when skill atrophy is more severe. Along the transition, Δ_t^{SC} / h_t^U is eventually increasing and converges to a level strictly above its initial value.

(ii) When steady-state output falls below the no-adoption benchmark ($Y^* < \bar{h}$), we have $\Delta_t^{SC} > 0$ for all t: AI appears indispensable even when it reduces long-run output.

This result requires only $\mu < 1$; no spillovers, no externalities, no cross-agent interaction. As skills atrophy toward $h^* < \bar{h}$, the worker's AI-independent productivity falls, inflating the measured value of AI in state-conditional comparisons. The mechanism is a form of endogeneity familiar from labor economics: controlling for current skill when skill is itself shaped by the treatment biases the estimate upward, analogous to the "bad controls" problem identified by Angrist and Pischke (2009).[14] The bias is proportional to $\bar{h} - h_t^U$, the cumulative human capital deficit, and thus grows with adoption intensity, adoption duration, and the degree to which μ falls below unity.

[13] We maintain $\tilde{\beta} = \beta$ throughout the main analysis to isolate spillover externalities.

[14] The insight that technologies can appear indispensable because they degrade alternatives is familiar from the path dependence literature (David, 1985), but that work concerns technology lock-in, not measurement distortion.

Corollary 2 (Welfare Reversal Under Patient Evaluation)**.** *When steady-state output falls below the no-adoption benchmark ($Y^* < \bar{h}$), there exists $\tilde{\beta}^* > \beta$ such that for all $\tilde{\beta} > \tilde{\beta}^*$, $\Delta^{PATH}(\tilde{\beta}) < 0$: under sufficiently more patient evaluation than the firm's own discount factor, the adoption path is welfare-inferior.*

The corollary exposes a tension between private optimality and social evaluation. Adoption may be privately optimal (revealed preference implies $\Delta^{PATH}(\beta) \geq 0$) yet welfare-reducing under more patient evaluation. This is not a market failure; firms optimize correctly given their discount factor. The welfare loss arises from the divergence between private and social time preferences.

4.2 Spillover Bias

State-path divergence operates at the individual level. A second bias arises at the cross-sectional level when AI adoption degrades the learning environment for non-users.

Definition 2 (Cross-Sectional vs. Long-Run Counterfactuals)**.** The *cross-sectional counterfactual* compares AI users to contemporaneous non-users: $\Delta_t^{CS} = Y(h_t^U, \alpha_t) - Y(h_t^{NU}, 0)$. The *long-run counterfactual* compares to the path where AI was never adopted: $\Delta_t^{LR} = Y(h_t^U, \alpha_t) - Y(h_t^{NA}, 0)$. All paths start from $h_0 = \bar{h}$. In potential-outcomes notation, $h_t^U = h_t(1)$, $h_t^{NA} = h_t(0)$, and $h_t^{NU} = h_t(0 \mid \bar{\alpha} > 0)$: the non-user's skill path in the presence of aggregate adoption.

The cross-sectional counterfactual is the comparison made by most empirical studies, including RCTs that randomize AI access. These diverge from the long-run counterfactual when aggregate AI adoption affects learning opportunities for non-users.

Assumption 4 (Learning Spillovers)**.** Individual learning depends on aggregate human capital: $L_i = [(1 - \alpha_i) + \mu\alpha_i] \cdot \varphi(h_i) \cdot \psi(H)$, where $\psi(H) = (H/\bar{H})^\eta$ and $\eta \geq 0$ governs spillover intensity. Appendix B.2 derives ψ from a matching model where mentor availability depends on the skill distribution.[15]

Under Assumption 4, a non-user in an AI-adopting economy accumulates skill according to

$$h_{t+1}^{NU} = (1 - \delta)h_t^{NU} + \lambda\varphi(h_t^{NU}) \cdot \psi(H_t) \tag{9}$$

[15]Evidence includes peer effects in education (Sacerdote, 2001), coworker effects in firms (Mas and Moretti, 2009), and rising returns to social skills (Deming, 2017). Our results require only $\eta > 0$, not precise estimates. Appendix B.2 develops a matching-based microfoundation for $\psi(H)$.

while the no-adoption counterfactual satisfies $h_{t+1}^{NA} = (1-\delta)h_t^{NA} + \lambda\varphi(h_t^{NA})$ (since $\psi(\bar{H}) = 1$). Define the *spillover skill gap* $s_t \equiv h_t^{NA} - h_t^{NU}$. At $t = 0$, both paths start at $\bar{h}$, so $s_0 = 0$. For $t \geq 1$, subtraction yields the recursive characterization:

$$s_{t+1} = (1-\delta)s_t + \lambda\Big[\varphi(h_t^{NA}) - \varphi(h_t^{NU})\psi(H_t)\Big] \tag{10}$$

The second term is strictly positive when $H_t < \bar{H}$: non-users learn less because $\psi(H_t) < 1$ (degraded environment). Since h_t^{NU} is strictly decreasing from $\bar{h}$ under the contraction condition in Assumption 1 (Appendix C), the gap $s_t = \bar{h} - h_t^{NU}$ is strictly increasing, and the spillover bias $\Delta_t^{CS} - \Delta_t^{LR} = s_t$ inherits this monotonicity.

Proposition 2 (Spillover Bias)**.** *Suppose $\mu < 1$ and Assumption 4 holds with $\eta > 0$. Then $\Delta_t^{CS} > \Delta_t^{LR}$ for all $t > 0$, with the gap strictly increasing in t.*

The bias is zero at $t = 0$ (before adoption affects non-users) and grows as AI diffuses. It is largest in high-adoption sectors with strong mentorship traditions; within-firm studies comparing coworkers are most affected, cross-industry comparisons least affected. When spillovers are absent ($\eta = 0$), cross-sectional estimates correctly measure long-run effects, but state-path divergence (Proposition 1) remains.

Consider a concrete example. A law firm adopts AI for contract drafting. Senior associates who previously mentored juniors now spend less time teaching, since AI handles routine drafting. Non-users face degraded mentorship even though they never use AI themselves. An RCT comparing AI users to these non-users would attribute the full performance gap to AI's direct effect, missing that part of the gap reflects degraded training for the control group. Garicano and Rayo (2025) document a related mechanism: when AI automates entry-level legal work, the economic foundation of apprenticeship collapses because juniors generate insufficient billable output to justify training investment.

The user/non-user dichotomy is itself a simplification. In practice, AI tools offer heterogeneous modes of interaction, from autocomplete interfaces that minimize cognitive engagement to explanation modes that scaffold understanding.[16] Finer-grained data on *how* workers use AI, not just *whether* they use it, would sharpen both measurement and policy design.

Remark 3 (Bias Decomposition). The total bias in cross-sectional estimates admits a useful

[16]The effective μ is partly a choice variable: within the population of "AI users," those who select pedagogical modes accumulate skill faster, introducing selection on μ that cross-sectional comparisons conflate with selection on ability.

decomposition. Adding and subtracting both $Y(h_t^U, 0)$ and $Y(h_t^{NA}, 0)$:

$$\Delta_t^{CS} = \underbrace{Y(h_t^U, \alpha_t) - Y(h_t^U, 0)}_{\Delta_t^{SC} \text{ (state-conditional gain)}} + \underbrace{Y(h_t^U, 0) - Y(h_t^{NA}, 0)}_{\text{state-gap bias } (=h_t^U - h_t^{NA} < 0)} + \underbrace{Y(h_t^{NA}, 0) - Y(h_t^{NU}, 0)}_{\text{spillover bias } (=h_t^{NA} - h_t^{NU} = s_t > 0)} \quad (11)$$

The state-conditional gain Δ_t^{SC} is what panel studies typically estimate. The state-gap component $h_t^U - h_t^{NA} < 0$ reflects the cumulative human capital deficit from past AI use and *reduces* the cross-sectional estimate relative to the state-conditional one. The spillover component $s_t > 0$ *inflates* the estimate by depressing the non-user comparison group. The two biases have distinct remedies: state-path divergence calls for counterfactual-aware research designs, while spillover bias calls for Pigouvian correction.

4.3 When Bias Reverses Sign: The Skill Trap

The mismeasurement biases operate whenever $\mu < 1$. Under additional conditions, they reverse the sign of measured effects: AI appears beneficial in cross-sectional comparisons while reducing long-run output.

Definition 3 (Skill Trap)**.** The economy is in a *skill trap* if the equilibrium path $\{(h_t, \alpha_t)\}_{t=0}^{\infty}$ satisfies:

(T1) **Positive adoption:** $\alpha_t > 0$ for all $t \geq 0$.

(T2) **Level crossing:** There exists $T^* > 0$ such that $Y_t > Y_t^{NA}$ for $t < T^*$ and $Y_t < Y_t^{NA}$ for $t > T^*$, where $Y_t^{NA} = Y(h_t^{NA}, 0) = h_t^{NA}$ is output on the no-adoption path.

(T3) **Individual rationality:** $\alpha_t = \alpha^*(h_t)$ solves the firm's problem (5) at each t.

Condition (T2) concerns productivity *levels*, not growth rates: the trap means AI users eventually produce less than they would have produced without AI. The trap is individually rational: firms optimize at every date, yet the equilibrium path delivers lower long-run output than no adoption.[17]

Remark 4 (Trap Terminology). The model features a unique, globally stable equilibrium; "trap" refers to the practical irreversibility when $\tau_H \gg \tau_A$ – early intervention to prevent skill loss is far more effective than later attempts to reverse it.

Proposition 3 (Existence of the Skill Trap)**.** *Under Assumption 3 with initial condition $h_0 \leq \bar{h}$, the economy is in a skill trap if and only if:*

[17] This relates to the "competency trap" in organizational learning (Levinthal and March, 1993).

(i) $\mu < 1$ *(AI substitutes for learning);*

(ii) $A \cdot G(1) < \bar{h}$ *(human expertise remains economically valuable);*

(iii) $\beta < \bar{\beta}$*, where* $\bar{\beta} \equiv \sup\{\beta : Y^*(\beta) < \bar{h}\}$ *is the patience threshold below which firms over-adopt.*

The trap requires all three conditions. Condition (ii) is the binding constraint: if AI can fully substitute for human expertise, adoption is unambiguously beneficial regardless of skill loss. The trap is thus a transitional phenomenon, applying when AI is productive enough to attract adoption but not yet capable enough to render human expertise irrelevant. As AI improves, condition (ii) is eventually violated and the trap dissolves, though human capital losses incurred during the transition may be irreversible on policy-relevant timescales. The mismeasurement biases (Propositions 1–2) operate under the weaker condition $\mu < 1$ alone; the trap clarifies when bias reverses sign.

The proof (Appendix C) establishes a result of independent interest: at the equilibrium, marginal adoption strictly reduces steady-state output when $\mu < 1$. Defining $W(\alpha) \equiv A \cdot G(\alpha) + h^*(\alpha)(1-\alpha)^{1-\gamma}$ as steady-state output as a function of adoption, the proof derives the closed-form shadow value of skill:

$$V'(h^*) = \frac{(1-\alpha^*)^{1-\gamma}}{1 - \beta\big[(1-\delta) + \lambda\ell(\alpha^*)\varphi'(h^*)\big]} \tag{12}$$

Substituting this and equation (8) into $W'(\alpha^*)$ and simplifying yields

$$W'(\alpha^*) = -\frac{(1-\beta)\,\lambda(1-\mu)\varphi(h^*)(1-\alpha^*)^{1-\gamma}}{\Gamma\big[(1-\beta) + \beta\Gamma\big]} < 0 \tag{13}$$

where $\Gamma \equiv \delta - \lambda\ell(\alpha^*)\varphi'(h^*) > 0$ at a stable steady state. The sign is determined entirely by $(1-\beta) > 0$: marginal adoption always reduces long-run output when $\mu < 1$, regardless of parameter values. The inequality is strict for all $\beta < 1$, so the trap is avoidable only in the limit of infinite patience. Combined with $d\alpha^*/d\beta < 0$ (Corollary 1), this implies $dY^*/d\beta > 0$: patience is unambiguously welfare-improving.

Corollary 3 (Sign Reversal)**.** *When the economy is in a skill trap, the measured effect has the wrong sign:* $\Delta_t^{CS} > 0 > \Delta_t^{LR}$ *for* t *sufficiently large.*

In the skill trap, $\Delta^{LR} < 0$ follows directly from $Y^* < \bar{h}$. What spillovers provide is $\Delta^{CS} > 0$ despite this: learning spillovers degrade non-users' skills so that $h^{NU*} < \bar{h}$, allowing

$Y^* > h^{NU*}$ even when $Y^* < \bar{h}$. Cross-sectional gains can coexist with long-run losses. For empirical researchers, this means a positive and statistically significant treatment effect in an RCT is consistent with the technology reducing aggregate welfare.

4.4 Cohort Effects and Wage Dynamics

The mismeasurement problems above generate predictions for wages and inequality. An important and growing literature documents that AI disproportionately benefits less-skilled workers in the short run (Brynjolfsson et al., 2025a; Noy and Zhang, 2023; Peng et al., 2023), a finding that Autor (2024) interprets as evidence of skill democratization. These results are valuable and may well describe the dominant effect over horizons where skill formation is not at stake. Our framework complements this view by characterizing the longer-run dynamics: when $\mu < 1$, the initial compression reverses as skill atrophy accumulates, and the same data that support democratization in the short run are consistent with growing inequality over careers.

Consider workers who differ in learning ability θ_i, where higher θ implies faster skill accumulation: $\varphi_i(h) = \theta_i \varphi(h)$. Let $h_t^{NA}(\theta)$ and $h_t^U(\theta)$ denote skill paths without and with AI adoption for a worker of ability θ.

Proposition 4 (Ability Reversal and Vintage Premium)**.** *Suppose $\mu < 1$ and let wages equal marginal products.*

(i) The skill loss from AI adoption is increasing in ability: $\partial(h_t^{NA} - h_t^U)/\partial\theta > 0$ for all $t \geq 1$.

(ii) Pre-AI cohorts who never adopt maintain skill $\bar{h}$; post-AI cohorts converge to $h^* < \bar{h}$. The vintage premium $\pi_t = \bar{h}/h_t^{post}$ increases in t until retirement.

Part (i) implies high-ability workers bear the largest long-run costs, precisely those who benefit least from AI in short-run studies. In the short run, AI compresses their productivity advantage by disproportionately helping less-skilled peers; in the long run, AI impedes their skill development, preventing them from reaching their full potential. We call this *ability reversal* because short-run and long-run effects have opposite signs for high-ability workers.

This creates a political economy challenge: early AI adoption generates enthusiasm because those who benefit most visibly (low-ability workers gaining immediate productivity) are not those who bear the largest long-run costs.

Part (ii) implies pre-AI cohorts become increasingly valuable as repositories of expertise that post-AI cohorts cannot easily replicate. The vintage premium is not merely a transitional artifact; it persists until retirement eliminates the pre-AI cohort entirely. Early evidence is consistent: Beane (2019) documents that robotic surgery reduced trainee hands-on experience tenfold, with senior surgeons becoming increasingly valuable for complex cases requiring manual dexterity. Ong and Png (2026) find that deskilling technologies increase labor supply by lowering the skill threshold for entry, which in our framework further depresses wages for post-AI cohorts. As pre-AI workers retire, their expertise is permanently lost, creating a form of organizational knowledge destruction that training programs cannot reverse when the pedagogical environment has itself been degraded.

The cohort dynamics generate predictions for aggregate inequality. Let N_t^{pre} denote the mass of pre-AI workers (declining through retirement) and $\sigma_t^2 = \text{Var}(w_t)$ denote wage variance across all workers at time t.

Corollary 4 (Hump-Shaped Inequality)**.** *Suppose $\mu < 1$ and pre-AI cohorts retire at rate $\nu > 0$. Then wage variance σ_t^2 follows a hump-shaped path: $\sigma_0^2 = 0$ initially; σ_t^2 rises as post-AI workers' skills diverge from pre-AI workers'; and σ_t^2 peaks at some T^{max} then declines as pre-AI cohorts retire.*[18]

The corollary yields a subtle but important implication for empirical measurement of AI's distributional effects. A policymaker observing falling inequality in the first decade after widespread AI adoption might conclude that AI is compressing skill gaps, consistent with the democratization narrative. But this compression is temporary and misleading: it reflects erosion of high-ability workers' skill advantages rather than elevation of low-ability workers' capabilities. The subsequent rise in inequality is driven by scarcity premiums for pre-AI veterans who possess expertise that post-AI training environments can no longer produce.

The timing of the hump depends on the retirement rate ν and the speed of skill atrophy $(1-\mu)\alpha^*$. Under our baseline calibration, $T^{max} \approx 25$ years, roughly one generation. This is long enough that the reversal may be attributed to unrelated structural changes rather than recognized as a consequence of AI-induced skill dynamics. The hump shape also distinguishes our mechanism from standard skill-biased technological change, which predicts monotonic inequality increases (Acemoglu and Autor, 2011).

[18] The vintage premium π_t tracks the bilateral gap between pre-AI and post-AI cohorts (monotonically increasing). The hump shape in σ_t^2 reflects *aggregate* inequality dynamics as cohort composition shifts: variance rises with the skill gap but falls as the high-skill cohort shrinks.

The cohort effects interact with the measurement biases in Section 4.3. Within any cohort, the state-path divergence inflates measured AI benefits. Across cohorts, the vintage premium creates a composition effect: as the workforce shifts toward post-AI cohorts with lower baseline skills, aggregate productivity growth slows even as within-cohort measured AI effects remain positive. A comprehensive accounting of AI's productivity impact must track both margins.

4.5 Illustrative Calibration

We calibrate the model to experimental evidence to illustrate the magnitude of mismeasurement. A central difficulty is that existing experiments estimate μ over short horizons (weeks to months), whereas the biases we characterize operate over careers. Whether short-run estimates of pedagogical quality extrapolate to longer horizons is itself an open empirical question; we treat μ as uncertain and report results across $\mu \in [0.3, 0.9]$, reflecting both this uncertainty and heterogeneity across tasks and AI designs.

Bastani et al. (2025) find GPT-4 access reduces subsequent math test performance by 17%, implying $\mu \approx 0.83$. Shen and Tamkin (2026) find a nearly identical 17% reduction among software developers learning a new Python library – a different population, task domain, and research team, yet the same point estimate.[19] The convergence is suggestive but not dispositive; both experiments study novices over short horizons, and whether μ remains stable over longer exposure or varies across tasks and populations remains an open question.

Other evidence spans a wide range: Dell'Acqua (2022) document reduced cognitive effort with AI ($\mu < 0.5$); Budzyń et al. (2025) find endoscopist deskilling ($\mu \approx 0.6$–0.8); Gaessler and Piezunka (2023) find chess engines accelerated skill development ($\mu > 1$), plausibly because chess provides immediate, unambiguous feedback unlike most knowledge work. We report results for multiple values of μ rather than defending a single estimate.

Baseline parameters. We use $\delta = 0.05$ (5% annual depreciation), $\lambda = 0.15$ (steady-state skill reached in approximately 15 years), $\alpha = 0.5$ (adoption intensity), $A = 1.5$ (AI productivity), $\gamma = 0.3$ (effort concentration), $\varphi(h) = 0.2/(1 + h)$ (diminishing returns to learning), and $\eta = 0.15$ (spillover elasticity). Table 2 reports results across the μ range.

[19] Both experiments study novices over short horizons. Whether μ remains stable over longer exposure is an open question; the convergent estimate nonetheless anchors the parameter where our biases are economically meaningful.

Table 2: Calibration Results by Pedagogical Quality μ

	Pedagogical Quality μ				
Outcome	1.0	0.9	0.7	0.5	0.3
Steady-state skill $h^*/\bar{h}$	1.00	0.96	0.88	0.80	0.71
Bias at year 10 (%)	0.0	2.0	6.2	10.7	15.5
Bias at year 20 (%)	0.0	3.1	9.7	17.0	25.2
Vintage premium at year 10 (%)	0.0	1.9	6.0	10.6	15.6
Vintage premium, steady state (%)	0.0	4.0	13.5	25.3	40.7

Note: Bias defined as $(\Delta_t^{SC} - \Delta_t^{LR})/\Delta_t^{LR}$. Vintage premium is $\bar{h}/h_t^{post} - 1$. AI quality exogenous ($\zeta = 0$). Other parameters: $\delta = 0.05$, $\lambda = 0.15$, $\alpha = 0.5$, $\eta = 0.15$.

Qualitative conclusions are robust: bias is positive and economically meaningful for all $\mu < 1$, ranging from 2% at year 10 when $\mu = 0.9$ to over 15% when $\mu = 0.3$.

The bias is increasing in adoption intensity α, decreasing in μ, and increasing in learning intensity λ; it is relatively insensitive to δ.[20]

Wage and inequality implications. The skill gap translates directly into wage differentials under competitive labor markets. Pre-AI cohorts command growing premiums as their skills become scarce: at $\mu = 0.5$, the vintage premium reaches roughly 25% in steady state. Aggregate inequality follows a hump-shaped path, rising as the skill gap between cohorts widens and then falling as pre-AI workers retire. A policymaker observing falling inequality in the first decade might conclude AI is reducing skill gaps, but this compression is temporary, driven by erosion of high-skill workers' advantages rather than elevation of low-skill workers' capabilities.[21]

4.6 Implications for Empirical Research

Our analysis identifies a precise estimand mismatch at the heart of AI productivity measurement. To make this concrete, define the estimand recovered by a short-run RCT that randomizes AI access at $t = 0$ and measures output at t:

$$\tau^{RCT}(t) = \mathbb{E}[Y_t(1, h_t(1)) - Y_t(0, h_t(0))] \tag{14}$$

[20] A full sensitivity analysis is available in the supplementary appendix. At baseline ($\mu = 0.5$), varying α from 0.3 to 0.7 shifts year-10 bias from 8.9% to 13.6%; varying λ from 0.10 to 0.20 shifts it from 6.9% to 15.1%.

[21] The "democratization" narrative (Autor, 2024) correctly describes short-run compression but misses long-run convergence to uniformly lower skills, punctuated by scarcity premiums for pre-AI veterans.

where $Y_t(d, h)$ is output under treatment $d \in \{0, 1\}$ at skill level h, and $h_t(d)$ is the skill path under treatment d. For small t, $h_t(1) \approx h_t(0)$ and $\tau^{RCT}(t)$ recovers the direct productivity effect. But for large t, skill paths diverge: $h_t(1) < h_t(0)$ when $\mu < 1$. The welfare-relevant object is the discounted path comparison:

$$\Delta^W = \sum_{t=0}^{\infty} \tilde{\beta}^t \, \mathbb{E}[Y_t(1, h_t(1)) - Y_t(0, h_t(0))] \tag{15}$$

In a within-worker panel that controls for experience or tenure, the estimand conditions on current skill, recovering:

$$\tau^{panel}(t) = \mathbb{E}[Y_t(1, h_t) - Y_t(0, h_t) \mid h_t] \tag{16}$$

which equals our state-conditional gain Δ_t^{SC} and *overstates* Δ^W by the state-gap bias (Remark 3). Cross-sectional comparisons of users to non-users recover Δ_t^{CS}, which lies between $\tau^{panel}(t)$ and Δ_t^{LR}: the spillover bias inflates the cross-sectional estimate relative to the long-run counterfactual (Proposition 2), while the state-gap deflates it relative to the panel estimate. Table 3 summarizes which results require which assumptions; Table 4 maps empirical strategies to their bias exposure.

Table 3: Logical Dependence of Main Results

	$\mu < 1$	Spillovers	Feedback
Spillover bias (Prop. 2)	Yes	Yes	No
State-path divergence (Prop. 1)	Yes	No	No
Skill trap (Prop. 3)	Yes	No	No

The choice of research design fundamentally determines exposure to these biases. Within-firm RCTs face maximum spillover bias when coworkers share mentorship networks. Comparing pre-AI to post-AI cohorts approximates the path counterfactual and minimizes both biases. Staggered adoption designs occupy an intermediate position: they control for time-invariant worker heterogeneity but remain vulnerable to spillover effects that operate within industries.

Our analysis predicts that effect sizes should decline in longer panels, with faster decline where learning-by-doing is central. Cross-sectional estimates should systematically exceed within-worker panel estimates from the same setting. These predictions are testable as longitudinal data accumulate.

Data requirements for unbiased long-run estimation are demanding: direct assessments of

Table 4: Empirical Designs and Bias Exposure

Design	Spillover	State-Path	Notes
Novices, learning-intensive	High	High	Maximum bias exposure
Within-firm RCT (long-run)	High	High	Both biases accumulate
Within-firm RCT (short-run)	High	Low	Coworkers share mentors; skills unchanged yet
Staggered adoption DiD	Moderate	Moderate	Within-industry spillovers; timing-dependent
Pre/post AI cohort	Low	Low	Approximates path counterfactual
AI-free training periods	Low	Low	Directly tests skill formation
Expert users, routine tasks	Low	Low	Skill formation not at stake

human capital tracked over time (not just output), longitudinal records of AI usage intensity, measures of mentorship exposure and training environment quality, cohort identifiers relative to AI diffusion, and indicators distinguishing "autocomplete" from "tutor" AI interfaces. No existing dataset satisfies all requirements simultaneously, but the emerging experimental literature provides building blocks: Bastani et al. (2025) measure skill directly, METR (2025) track usage intensity, and Budzyń et al. (2025) observe unassisted performance after AI exposure.

Testable predictions. The framework generates several sharp, falsifiable predictions. First, effect sizes should decline in longer panels when $\mu < 1$, with faster decline where learning-by-doing is central; if $\mu \geq 1$, effect sizes should *grow* over time. Second, cross-sectional estimates should systematically exceed within-worker panel estimates from the same setting. Third, the bias should be larger for novices than for experts (Proposition 4). Fourth, the bias should be larger in low-verifiability domains where training data degradation (Appendix A) is weakest. Fifth, pre-AI cohorts should command growing wage premiums relative to post-AI cohorts. Longitudinal data are accumulating rapidly; these predictions are testable within the next few years.

5 Welfare and Policy

5.1 Sources of Inefficiency

Does mismeasurement reflect an actual welfare loss or merely a measurement artifact? The answer depends on whether decentralized adoption is efficient.

A firm that anticipates skill degradation can optimize intertemporally, and if it bears the full cost of atrophy, the decentralized equilibrium is constrained efficient. Welfare loss requires an externality: human capital must generate social value beyond what the adopting firm captures.

We augment the baseline model to allow learning to depend on aggregate human capital through $\psi(H)$, capturing mentorship and peer effects. Appendix B.2 derives ψ from a matching model where the probability of finding a capable mentor depends on the skill distribution. Let α^D denote decentralized adoption, α^S the social optimum solving

$$\max_{\{\alpha_t\}} \sum_{t=0}^{\infty} \tilde{\beta}^t Y(H_t, \alpha_t; A) \quad \text{s.t.} \quad H_{t+1} = (1-\delta)H_t + \lambda \ell(\alpha_t)\varphi(H_t)\psi(H_t), \tag{17}$$

and let W^D, W^S denote the corresponding welfare levels.

Proposition 5 (Human Capital Externality)**.** *With exogenous AI quality and common discounting ($\beta = \tilde{\beta}$), overadoption ($\alpha^D > \alpha^S$) occurs if and only if $\psi'(H) > 0$.*

The "if and only if" matters. Without spillovers, firms bear the cost of their workers' skill loss through lower future output. Spillovers break this logic: adoption imposes costs on other firms' workers that the adopting firm does not internalize. The overadoption result echoes a classic theme: when training generates positive externalities, decentralized investment is inefficiently low (Becker, 1962; Acemoglu and Pischke, 1999). Our contribution inverts this logic: the externality arises not from underinvestment in training but from overinvestment in a technology that degrades training as a byproduct. The externality is novel because firms do not underinvest in human capital directly; rather, they overinvest in AI, which indirectly reduces the human capital stock through its effect on learning.

A second externality arises from AI's dependence on human-generated training data. When workers delegate to AI, two effects degrade the training signal: AI-generated content is in-distribution, and AI-reliant humans produce lower-quality unassisted output. Appendix A develops the model of endogenous AI quality in detail.

Proposition 6 (Training Data Externality). *With endogenous AI quality* $A_{t+1} = (1-\zeta)A_t + \zeta Q(H_t, \bar{\alpha}_t)$ *where* $\partial Q/\partial H > 0$ *and* $\partial Q/\partial \bar{\alpha} < 0$*: individual adoption degrades future AI quality, generating overadoption. With both externalities present, total welfare loss exceeds the sum of individual effects.*

The decline in Stack Overflow activity after ChatGPT's release (del Rio-Chanona et al., 2024; Burtch et al., 2024) illustrates the training data externality: if future models train on AI-heavy corpora, they inherit the limitations of degraded human input.

The magnitude of overadoption depends on the strength of spillovers. Athey and Scott Morton (2025) analyze welfare effects of AI in competitive markets; our framework adds that competition may exacerbate overadoption when firms cannot internalize skill externalities. When $\psi'(H) = 0$ (no spillovers), decentralized adoption is constrained efficient. As spillovers strengthen, the wedge $\alpha^D - \alpha^S$ widens. At $\eta = 0.15$, overadoption is roughly 8% of the efficient level.

5.2 Policy Responses

Pigouvian taxation is the textbook remedy for externalities, but the same mismeasurement that biases productivity estimates also biases policy evaluation. We focus on quantity restrictions that do not require accurate measurement of shadow values.

Proposition 7 (Training Mandates). *Consider a training mandate* $\rho \in [0, 1]$ *requiring at least fraction* ρ *of work be performed without AI, constraining adoption to* $\alpha \leq 1 - \rho$*:*

(i) A binding mandate $\rho \in (1 - \alpha^D, 1 - \alpha^S]$ *is welfare-improving. The first-best mandate* $\rho^* = 1 - \alpha^S$ *implements the social optimum.*

(ii) Under the optimal mandate, measured productivity may fall while welfare rises.

Training mandates exist where skill maintenance is safety-critical, and our framework provides formal justification for extending such mandates to knowledge work: the FAA recommends pilots manually fly departure and arrival phases (Casner et al., 2014), and medical residency programs mandate unassisted procedure volumes (Beane, 2019). The optimal mandate $\rho^* = 1 - \alpha^S$ is increasing in spillover strength η, decreasing in μ and β, and equals zero when $\mu \geq 1$.[22]

[22] Proofs and comparative statics are in Appendix C.

6 Conclusion

This paper identifies two structural sources of mismeasurement in AI productivity studies. State-path divergence arises because current skill reflects past AI use; standard estimates *condition on an endogenous state*. Spillover bias arises because non-users face degraded learning environments. Both biases hinge on pedagogical quality μ: when $\mu < 1$, estimates overstate long-run benefits and the skill trap becomes possible; when $\mu > 1$, estimates understate benefits.[23] Available evidence favors $\mu < 1$ for most current AI applications: two independent experiments yield $\mu \approx 0.83$ (Bastani et al., 2025; Shen and Tamkin, 2026), corroborated by observational evidence of deskilling (Budzyń et al., 2025; METR, 2025), reduced cognitive effort (Dell'Acqua, 2022), and uncritical adoption (Shaw and Nave, 2026).

Our analysis has limitations. The parameter μ is context-dependent and imprecisely estimated; existing experiments measure μ over weeks, whereas our model concerns career-length horizons, and whether short-run estimates extrapolate is an open question. Workers might reallocate effort freed by AI to complex tasks, but evidence suggests otherwise (Lee et al., 2025). We treat μ as exogenous and abstract from task-level heterogeneity and selection into AI use, though we discuss extensions in Appendix B. Panel data tracking AI usage and skill assessments would permit direct estimation of μ; the emerging experimental literature provides a template.

The two biases require different remedies. State-path divergence is a measurement problem: difference-in-differences and RCTs comparing AI users to non-users will systematically overstate benefits when both groups' skills have been shaped by AI's presence. Cohort comparisons exploiting variation in AI exposure, or cross-country comparisons leveraging differential adoption timing, better approximate the welfare-relevant counterfactual. Spillover degradation is an externality amenable to Pigouvian correction or subsidies for human-generated training data.

A pronounced asymmetry adds urgency. Human capital accumulates slowly – expertise develops over years of deliberate practice – while it can be degraded quickly when AI substitutes for that practice. Degradation is fast because each period of delegation compounds: skill loss reduces the worker's capacity to perform tasks independently, reinforcing reliance on AI. Recovery is slow for a different reason: even though novices learn faster in absolute terms ($\varphi' < 0$), reversing atrophy requires *simultaneously* reducing AI use and rebuilding

[23] When $\mu > 1$, higher adoption increases learning. Non-users benefit from enhanced mentorship, and users' skills exceed the no-adoption counterfactual. Effect sizes in longitudinal studies should grow over time rather than shrink, a testable prediction that distinguishes the regimes.

skill, and the rebuilding timescale is governed by $1/\lambda$, which corresponds to years of practice. Moreover, the worker does not recover to $\bar{h}$ but converges to the atrophy steady state h^* unless adoption is also reduced – and reducing adoption is costly because AI-dependent workers produce less without it. Medical residency programs illustrate: it takes five to seven years to train a surgeon, but surgical skill can atrophy within months when robotic systems perform procedures that residents would otherwise learn by doing (Beane, 2019). Dahmani and Bohbot (2020) find that GPS-induced spatial memory decline was not reversed when GPS use ceased; Casner et al. (2014) find that pilots' cognitive flying skills, once degraded through heavy autopilot use, require extensive retraining to restore. *Preventing* skill degradation is far more effective than *reversing* it, and the training data feedback loop analyzed in Appendix A amplifies this asymmetry by coupling human capital recovery to AI quality recovery. By the time longitudinal data reveal the skill effects of AI adoption, the damage may be substantially harder to reverse than it would have been to prevent.

The framework connects to broader debates about technology and human capital. Automation has historically displaced workers from specific tasks while creating new ones that require different skills. Generative AI may be distinctive in that it targets the *process* of skill acquisition itself, not just the tasks that skills enable. If learning-by-doing is central to expertise development, and if AI substitutes for the cognitive effort that learning requires, then the long-run effects may differ qualitatively from previous waves of automation. A technology can appear *increasingly indispensable* even when it is not improving, because past use has degraded the alternative. The welfare-relevant counterfactual is not the worker's current state without the technology, but *the skill path that would have obtained* absent adoption. Our framework suggests that some of the most consequential effects of transformative technologies may be precisely those that standard productivity measurement is not designed to detect.

References

Acemoglu, D. and J.-S. Pischke (1999). The Structure of Wages and Investment in General Training. *Journal of Political Economy* 107(3), 539–572.

Acemoglu, D. and D. Autor (2011). Skills, Tasks and Technologies: Implications for Employment and Earnings. *Handbook of Labor Economics* 4, 1043–1171.

Acemoglu, D. and P. Restrepo (2018). The Race Between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment. *American Economic Review* 108(6), 1488–1542.

Acemoglu, D. and P. Restrepo (2020). Robots and Jobs: Evidence from US Labor Markets. *Journal of Political Economy* 128(6), 2188–2244.

Acemoglu, D. (2024). The Simple Macroeconomics of AI. *NBER Working Paper* 32487.

Agrawal, A. K., J. McHale, and A. Oettl (2026). Enhancing Worker Productivity Without Automating Tasks: A Different Approach to AI and the Task-Based Model. *NBER Working Paper* 34781.

Alemohammad, S., J. Casco-Rodriguez, L. Luzi, A. I. Humayun, H. Babaei, D. LeJeune, A. Siahkoohi, and R. G. Baraniuk (2024). Self-Consuming Generative Models Go MAD. *International Conference on Learning Representations*.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Arrow, K. J. (1962). The Economic Implications of Learning by Doing. *Review of Economic Studies* 29(3), 155–173.

Athey, S. and F. Scott Morton (2025). Artificial Intelligence, Competition, and Welfare. *NBER Working Paper* 34444.

Autor, D. H., F. Levy, and R. J. Murnane (2003). The Skill Content of Recent Technological Change: An Empirical Exploration. *The Quarterly Journal of Economics* 118(4), 1279–1333.

Autor, D. H. (2024). Applying AI to Rebuild Middle Class Jobs. *NBER Working Paper* 32140.

Bastani, H., O. Bastani, A. Sungu, H. Ge, Ö. Kabakçı, and R. Mariman (2025). Generative AI Without Guardrails Can Harm Learning: Evidence from High School Mathematics. *Proceedings of the National Academy of Sciences* 122(26), e2422633122.

Beane, M. (2019). Shadow Learning: Building Robotic Surgical Skill When Approved Means Fail. *Administrative Science Quarterly* 64(1), 87–123.

Becker, G. S. (1962). Investment in Human Capital: A Theoretical Analysis. *Journal of Political Economy* 70(5), 9–49.

Brynjolfsson, E., D. Li, and L. R. Raymond (2025). Generative AI at Work. *The Quarterly Journal of Economics* 140(2), 889–942.

Brynjolfsson, E., B. Chandar, and R. Chen (2025). Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence. Stanford Digital Economy Lab Working Paper.

Budzyń, K., et al. (2025). Endoscopist Deskilling Risk After Exposure to Artificial Intelligence in Colonoscopy: A Multicentre, Observational Study. *The Lancet Gastroenterology & Hepatology* 10(10), 896–903.

Burtch, G., D. Lee, and Z. Chen (2024). The Consequences of Generative AI for Online Knowledge Communities. *Scientific Reports* 14, 10413.

Casner, S. M., R. W. Geven, M. P. Recker, and J. W. Schooler (2014). The Retention of Manual Flying Skills in the Automated Cockpit. *Human Factors* 56(8), 1506–1516.

Cunha, F. and J. J. Heckman (2007). The Technology of Skill Formation. *American Economic Review* 97(2), 31–47.

Dahmani, L. and V. D. Bohbot (2020). Habitual Use of GPS Negatively Impacts Spatial Memory During Self-Guided Navigation. *Scientific Reports* 10, 6310.

David, P. A. (1985). Clio and the Economics of QWERTY. *American Economic Review* 75(2), 332–337.

Dell'Acqua, F., E. McFowland III, E. R. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. Harvard Business School Working Paper 24-013.

Dell'Acqua, F. (2022). Falling Asleep at the Wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters. Working Paper, Harvard Business School.

Deming, D. J. (2017). The Growing Importance of Social Skills in the Labor Market. *Quarterly Journal of Economics* 132(4), 1593–1640.

Gaessler, F. and H. Piezunka (2023). Training with AI: Evidence from Chess Computers. *Strategic Management Journal* 44(11), 2724–2750.

Garicano, L. and L. Rayo (2025). Training in the Age of AI: A Theory of Apprenticeship Viability. Working Paper.

Gibbons, R. and M. Waldman (2004). Task-Specific Human Capital. *American Economic Review* 94(2), 203–207.

Heckman, J. J., J. Stixrud, and S. Urzua (2006). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics* 24(3), 411–482.

Ide, E. (2025). Automation, AI, and the Intergenerational Transmission of Knowledge. IESE Business School Working Paper.

Kremer, M. (1993). The O-Ring Theory of Economic Development. *Quarterly Journal of Economics* 108(3), 551–575.

Lee, H.-P., et al. (2025). The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects from a Survey of Knowledge Workers. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–22.

Levinthal, D. A. and J. G. March (1993). The Myopia of Learning. *Strategic Management Journal* 14(S2), 95–112.

Lucas, R. E. (1988). On the Mechanics of Economic Development. *Journal of Monetary Economics* 22(1), 3–42.

Luo, L., E. Manzoor, and N. Yang (2025). Platform Design When Creators Train Their AI Substitutes. Working Paper, Cornell University.

Mas, A. and E. Moretti (2009). Peers at Work. *American Economic Review* 99(1), 112–145.

METR (2025). Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. arXiv preprint arXiv:2507.09089.

Noy, S. and W. Zhang (2023). Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. *Science* 381(6654), 187–192.

Ong, P. and I. P. L. Png (2026). Deskilling Technology Affords Work Amenity, Increases Labor Supply. *Strategic Management Journal* 47(1), e70017.

Otis, N. G., R. Clarke, S. Delecourt, D. Holtz, and R. Koning (2023). The Uneven Impact of Generative AI on Entrepreneurial Performance. Harvard Business School Working Paper 24-042.

Peng, S., E. Kalliamvakou, P. Cihon, and M. Demirer (2023). The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. arXiv preprint arXiv:2302.06590.

del Rio-Chanona, R. M., N. Laurentsyeva, and J. Wachs (2024). Large Language Models Reduce Public Knowledge Sharing on Online Q&A Platforms. *PNAS Nexus* 3(9), pgae400.

Sacerdote, B. (2001). Peer Effects with Random Assignment: Results for Dartmouth Roommates. *Quarterly Journal of Economics* 116(2), 681–704.

Shaw, S. D. and G. Nave (2026). Thinking – Fast, Slow, and Artificial: How AI is Reshaping Human Reasoning and the Rise of Cognitive Surrender. Working Paper, The Wharton School, University of Pennsylvania.

Shen, J. H. and A. Tamkin (2026). How AI Assistance Impacts the Formation of Coding Skills. *arXiv preprint arXiv:2601.20245*.

Shumailov, I., Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal (2024). AI Models Collapse When Trained on Recursively Generated Data. *Nature* 631, 755–759.

Spence, M. (1973). Job Market Signaling. *Quarterly Journal of Economics* 87(3), 355–374.

Stokey, N. L. and R. E. Lucas, Jr. (1989). *Recursive Methods in Economic Dynamics*. Harvard University Press.

A The Skill-Data Feedback Loop

This appendix develops a model of endogenous AI quality, formalizing the "skill-data feedback loop" mentioned in the main text (Section 6). The mechanism connects to the computer science literature on model collapse (Shumailov et al., 2024; Alemohammad et al., 2024): platforms like Stack Overflow provide both training data and mentorship networks, so when users exit, they reduce fresh training data *and* degrade peer-learning, compounding the inefficiencies we identify. The key idea is that generative AI, unlike previous automation technologies, learns from human-generated content. When widespread adoption degrades human skills, the quality of training data deteriorates, which in turn degrades AI quality and partially attenuates adoption. We characterize this loop formally, derive its stabilizing properties, and discuss the important caveats.

A.1 Why Generative AI Is Different

Previous automation technologies operate via fixed algorithms invariant to user skill: a calculator, a spreadsheet, and a GPS produce identical output regardless of the operator's expertise. Generative AI differs in two ways. First, it is trained on human-generated data, so the quality of future AI systems depends on the quality of human output today. Second, its training data is *contaminated* by its own outputs: as AI-generated content proliferates, future models increasingly train on machine-generated material. These properties create a feedback loop absent from earlier technologies. When workers delegate tasks, they produce less original content and what they produce reflects diminished expertise, degrading training data for the next generation of AI. Stack Overflow activity declined roughly 25% within six months of ChatGPT's release (del Rio-Chanona et al., 2024), with newer users most likely to exit (Burtch et al., 2024), reducing the pool of expert-authored training examples available to coding assistants trained on that content.

A.2 Formal Model

We model AI productivity as evolving according to

$$A_{t+1} = (1 - \zeta)A_t + \zeta \cdot Q(H_t, \bar{\alpha}_t) \tag{18}$$

where $\zeta \in (0, 1)$ governs how quickly AI quality adjusts, $\bar{\alpha}_t$ is average adoption intensity, and $Q : \mathbb{R}_+ \times [0, 1] \to \mathbb{R}_+$ satisfies $\partial Q/\partial H > 0$ and $\partial Q/\partial \bar{\alpha} < 0$.

Microfoundation. Consider an AI firm that trains its model on a corpus of human-generated content. Worker i produces content of quality $q_i = h_i \cdot (1 - \alpha_i)^\omega$, where $\omega > 0$ governs how AI assistance affects output quality. The term $(1 - \alpha_i)^\omega$ captures that AI-assisted output, while potentially correct, lacks the distinctive features – edge cases, creative solutions, expert judgment – that make training data valuable.

The training corpus has two components: human-generated content and AI-generated content that has "leaked" into the training set. Let π_t denote the fraction of AI-generated content. The effective training signal is:

$$S_t = (1 - \pi_t) \cdot \underbrace{\int q_i \, dF_i}_{\text{human quality}} + \pi_t \cdot \underbrace{A_{t-1}}_{\text{AI quality}} \tag{19}$$

The AI-generated component contributes only A_{t-1} because AI can reproduce what it already knows but cannot generate novel training signal.

Two channels degrade training data. The *quantity channel* operates through π: higher adoption increases the share of AI-generated content in the training corpus. The *quality channel* operates through H: lower human capital reduces the informativeness of human-generated content. Consider the microfoundation

$$Q(H, \bar{\alpha}) = q\Big(\underbrace{(1 - \bar{\alpha})}_{\text{flow share}} \cdot \underbrace{s(H)}_{\text{skill quality}} + \underbrace{\bar{\alpha} \cdot \kappa}_{\text{AI content signal}}\Big) \tag{20}$$

where q is increasing and concave, $s(H)$ is increasing in aggregate skill, and $\kappa \geq 0$ captures the training signal in AI-generated content, with $\kappa < s(\bar{H})$.[24]

For symmetric adoption $\alpha_i = \bar{\alpha}$, average human output quality is $\bar{q}_t \approx H_t \cdot (1 - \bar{\alpha}_t)^\omega$. The full law of motion becomes:

$$A_{t+1} = (1 - \zeta(1 - \pi))A_t + \zeta(1 - \pi)\tilde{Q}(H_t, \bar{\alpha}_t) \tag{21}$$

where $\tilde{Q}(H, \bar{\alpha}) \equiv H \cdot (1 - \bar{\alpha})^\omega$ is the human contribution to training signal quality. The contamination rate π is itself endogenous to adoption: $\pi_t = \pi(\bar{\alpha}_t)$ with $\pi' > 0$, but we suppress this dependence for tractability. For notational simplicity, the main text absorbs these terms into a single function $Q(H_t, \bar{\alpha}_t)$ satisfying $\partial Q/\partial H > 0$ and $\partial Q/\partial \bar{\alpha} < 0$.

[24] The condition $s(H) > \kappa$ ensures $\partial Q/\partial \bar{\alpha} < 0$. When $s(H) < \kappa$ (extreme skill collapse), the loop reverses. The legacy data stock is embedded in $(1 - \zeta)A_t$; when ζ is small, legacy data buffers quality against degradation.

Connection to model collapse. This specification connects to the computer science literature on "model collapse": recursive training on AI-generated content causes distributional tails to disappear (Shumailov et al., 2024; Alemohammad et al., 2024). In our framework, model collapse operates through π: as $\pi \to 1$ (all training data is AI-generated), the AI model increasingly trains on its own outputs, compressing the distribution and losing the tail information that distinguishes expert from routine work. The active debate over "synthetic data" is, in economic terms, a debate over the magnitude of κ – how much training signal AI-generated content retains relative to human-generated content.

A.3 Stabilization, Dynamics, and Slow Recovery

The joint dynamics of (H_t, A_t) form a two-dimensional system:

$$H_{t+1} = (1-\delta)H_t + \lambda \ell(\bar{\alpha}_t)\varphi(H_t)\psi(H_t) \tag{22}$$

$$A_{t+1} = (1-\zeta)A_t + \zeta Q(H_t, \bar{\alpha}_t) \tag{23}$$

where $\bar{\alpha}_t = \alpha^*(H_t, A_t)$ is equilibrium adoption given state (H_t, A_t). The key cross-partials are $\partial H'/\partial A < 0$ (higher AI quality induces more adoption, reducing learning) and $\partial A'/\partial H > 0$ (higher human capital improves training data). This negative feedback is the stabilizing force.

Proposition 8 (Feedback Loop Stability)**.** *The system (H_t, A_t) has a unique stable steady state (H^{**}, A^{**}) satisfying:*

(i) $H^{**} > H^*(A_0)$, where $H^*(A_0)$ is the steady state with AI quality fixed at $A_0 = Q(\bar{H}, 0)$: the feedback loop partially protects human capital relative to the exogenous high-quality AI benchmark.

(ii) $A^{**} < A_0$: equilibrium AI quality is below its potential when humans are fully skilled and no AI is used.

(iii) The steady state (H^{**}, A^{**}) is globally stable when $\mu < 1$ and ζ is sufficiently small.

The proof appears in Appendix C. The intuition follows from Corollary 1: $\partial \alpha^*/\partial A > 0$ and $\partial h^*/\partial A < 0$. When AI quality falls, firms adopt less, preserving more learning-by-doing.

Slow recovery. The stabilization masks a critical asymmetry. Define timescales $\tau_H \equiv 1/\lambda$ (human learning) and $\tau_A \equiv 1/\zeta$ (AI retraining). A temporary shock that increases adoption

for duration T produces a human capital deficit. Recovery requires time $T_R \geq c \cdot \tau_H$ for some constant $c > 0$, *regardless of how fast AI can retrain*. This is because AI quality depends on human capital through the training data channel: even if AI systems can be retrained in months, they cannot recover quality until human capital recovers, and human capital accumulates slowly. When $\tau_H \gg \tau_A$ (the empirically relevant case), $T_R \gg T$: temporary shocks produce persistent effects.[25]

Comparative statics. The stabilization gain $\Delta_S \equiv H^{**} - H^*(A_0)$ is increasing in ζ (faster AI quality adjustment), ω (stronger quality degradation from AI-assisted output), and $|\partial\alpha^*/\partial A|$ (stronger adoption response to AI quality changes).

A.4 Caveats and the Role of Data Curation

The training data degradation channel is subject to several important caveats that qualify its magnitude and relevance.

Task verifiability. The degradation channel operates at full force only for tasks with *unverifiable* outputs. For tasks with objectively verifiable outputs – code that must compile, proofs that can be checked, engineering designs that must satisfy physical constraints – AI-generated content retains training signal regardless of the user's skill. Automated test suites, type checkers, and formal verification provide supervision signals independent of the author's expertise. For unstructured tasks like writing, strategic judgment, or clinical diagnosis, no such filter exists. The feedback stabilization is correspondingly weaker for verifiable tasks: if AI quality does not degrade quickly, the stabilizing brake on skill atrophy is attenuated. Paradoxically, domains where AI output is most reliable may be those where skill atrophy proceeds most aggressively.

Synthetic data and data curation. Recent advances in synthetic data generation challenge the assumption that AI quality necessarily depends on human-generated training data. Several considerations qualify the strength of the degradation channel:

First, frontier AI labs invest heavily in data curation pipelines that filter, deduplicate, and quality-score training examples. These pipelines can partially mitigate contamination by

[25] Reinforcement learning systems that learn from self-play (e.g., AlphaZero) are not subject to training data degradation. For such systems, A_t is exogenous to H_t, the stabilizing feedback is absent, and condition (ii) of Proposition 3 is eventually violated as AI improves. RL-capable AI may thus produce deeper skill atrophy but less output loss.

identifying and downweighting AI-generated content. In our framework, effective curation reduces π (the contamination rate), weakening the quantity channel. However, curation cannot address the quality channel: as human skills atrophy, even filtered human-generated content becomes less informative.

Second, reinforcement learning from human feedback (RLHF) and related alignment techniques use human *evaluations* rather than human *outputs* as training signal. If evaluation skill degrades more slowly than production skill – plausible when evaluation requires recognition rather than generation – the feedback loop is weaker than our model implies. Shaw and Nave (2026) provide mixed evidence: participants adopted AI outputs uncritically even under strong incentives for correctness, suggesting that evaluation skill may also be vulnerable.

Third, AI training data quality may depend on the *upper tail* of the human skill distribution rather than the mean. If training data is curated to select the best human exemplars, the relevant quantity is not H_t (average skill) but $\max_i h_{i,t}$ or some percentile of the skill distribution. This is more optimistic for the feedback channel: a small cadre of highly skilled humans could sustain AI quality even as average skills decline. But it is also more fragile: the upper tail shrinks as pre-AI cohorts retire, and the very best human practitioners may be precisely those whose comparative advantage makes them least likely to adopt AI, creating a selection effect that is difficult to sustain indefinitely.

What the feedback loop does and does not require. The formal results require only that $\partial Q/\partial H > 0$ and $\partial Q/\partial \bar{\alpha} < 0$. These conditions can fail if synthetic data fully substitutes for human data, AI achieves recursive self-improvement independent of human input, or curation perfectly separates high-quality from low-quality examples. We view the first as unlikely for domains requiring judgment or tacit knowledge, though it may hold for formal domains; the second remains speculative; the third requires solving the very discrimination problem our framework identifies as difficult. The main text results (Propositions 2–3) do not depend on the feedback channel. The feedback loop is an additional mechanism, distinctive to generative AI, that modifies dynamics but does not drive the core measurement critique.

B Extensions

This appendix develops extensions of the baseline model. Each subsection is self-contained.

B.1 Firm Dynamics and Selection

When firms differ in their discount factors, AI adoption generates selection effects that amplify aggregate skill loss.

Assumption 5 (Heterogeneous Firm Patience)**.** Firms differ in discount factors $\beta_i \sim F_\beta$ distributed on $[\underline{\beta}, \bar{\beta}]$ with $0 < \underline{\beta} < \bar{\beta} < 1$. Firms compete in a product market where market share depends on current productivity.

Proposition 9 (Selection Effects)**.** *Under Assumption 5, during the transition from initial conditions:*

(i) $\frac{d\alpha^*}{d\beta} < 0$*: impatient firms adopt more intensively.*

(ii) Let $s_{i,t}$ *denote firm* i*'s market share. During the transition phase when the static gain from AI dominates skill loss,* $\frac{d}{dt}\mathbb{E}[\beta_i | s_{i,t}] < 0$*: the output-weighted average patience declines.*[26]

(iii) Aggregate human capital $H_t = \int h_{i,t} s_{i,t}\, di$ *satisfies* $H_t^{selection} < H_t^{no\text{-}selection}$ *during the transition: selection amplifies skill atrophy.*

The mechanism operates during the transition: impatient firms adopt AI more intensively, gain short-run productivity advantages (the static AI gain dominates skill loss initially), and capture market share from patient firms. This is a transitional phenomenon – in the very long run, the ranking reverses as patient firms' higher steady-state skills dominate. But the transition can be prolonged, and during this phase the output-weighted average patience declines, accelerating aggregate skill atrophy.

B.2 Microfoundations for Spillovers

This section provides a formal microfoundation for the learning spillover function $\psi(H)$ introduced in Section 5.

[26] This is a *transitional* result. In the long run, Corollary 1 establishes that $Y^*(\beta)$ is increasing in β, so patient firms have higher steady-state output. Eventually, selection may shift market share toward patient firms. The proposition characterizes the economically relevant early phase when impatient firms' higher current output dominates their lower long-run productivity.

Consider a population of workers indexed by $i \in [0, 1]$. Each period, worker i encounters a problem that requires skill level s drawn from distribution $F(s)$. If $h_i \geq s$, worker i solves the problem independently and learns $\varphi(h_i)$. If $h_i < s$, worker i must seek help from a randomly matched colleague j. The match succeeds (colleague can help) if $h_j \geq s$. When a match succeeds, worker i learns $\kappa\varphi(h_i)$ where $\kappa \in (0, 1)$ captures that mentored learning is valuable but less effective than independent problem-solving. When no match succeeds, worker i learns nothing from that problem.

The probability that a random colleague can help with a problem of difficulty s is $\Pr(h_j \geq s) = 1 - G_H(s)$, where G_H is the distribution of human capital in the population. For a worker with skill h_i, expected learning is:

$$\mathbb{E}[L_i] = \int_0^{h_i} \varphi(h_i)\, dF(s) + \int_{h_i}^{\bar{s}} \kappa\varphi(h_i)[1 - G_H(s)]\, dF(s) \tag{24}$$

The first term is learning from problems solved independently; the second is expected learning from mentored problems, weighted by the probability of finding a capable mentor.

Define $\Psi(H) \equiv \int_0^{\bar{s}} [1 - G_H(s)]\, dF(s)$, which measures the "mentorship capacity" of the economy – the average probability that a random worker can help with a random problem. When aggregate human capital H is high, G_H is shifted toward higher values, so $1 - G_H(s)$ is larger for any given s, and $\Psi(H)$ is increasing in H.

Expected learning can be written as:

$$\mathbb{E}[L_i] = \varphi(h_i)\left[F(h_i) + \kappa\Psi(H)[1 - F(h_i)]\right] \tag{25}$$

Normalizing so that $\psi(\bar{H}) = 1$ at the no-adoption steady state, the term in brackets motivates a multiplicative *approximation* $L_i \approx \ell(\alpha_i)\varphi(h_i) \cdot \psi(H)$.[27] The key insight is that aggregate human capital affects individual learning through the availability of mentors: when H falls, the probability of finding a capable mentor declines, reducing learning for all workers – including those who do not adopt AI.

[27] The exact expression $F(h_i) + \kappa\Psi(H)[1 - F(h_i)]$ depends on individual skill via $F(h_i)$ and cannot be literally factored into $g(h_i) \cdot \psi(H)$. The multiplicative form in the main text is a reduced-form approximation that captures the key qualitative feature: aggregate human capital affects individual learning through mentor availability. See Appendix B.3 for robustness to alternative specifications.

B.3 Robustness to Functional Forms

This section verifies that our main results are robust to alternative functional form specifications.

Alternative learning functions. The baseline model assumes a monotonically decreasing learning capacity function $\varphi(h)$. An alternative specification is a hump-shaped function that peaks at intermediate skill levels, capturing that complete novices may lack the framework to learn efficiently. All qualitative results survive under the hump-shaped specification: when $\mu < 1$, higher adoption still reduces steady-state human capital because $\partial L/\partial \alpha = (\mu - 1)\varphi(h) < 0$. The steady-state characterization requires restricting attention to $h^* > \hat{h}$ (above the peak) for stability, but the comparative statics retain their signs.

Alternative AI capability functions. The baseline assumes $g(j)$ is monotonically decreasing in j, so AI is best at routine tasks. Consider instead a U-shaped function where AI is capable at both routine tasks (low j) and highly structured complex tasks (high j), but struggles with intermediate judgment-intensive tasks. The optimal adoption rule becomes more complex (potentially non-convex), but the core mechanism – that delegation reduces learning when $\mu < 1$ – is unchanged. The skill trap can still arise whenever AI handles tasks that would otherwise develop human expertise.

Alternative spillover specifications. Replace the multiplicative specification $L_i = \ell(\alpha_i)\varphi(h_i)\psi(H)$ with an additive form $L_i = \ell(\alpha_i)\varphi(h_i) + \theta_L H$, where $\theta_L > 0$ captures direct knowledge spillovers. The overadoption result (Proposition 5) continues to hold: individual firms ignore their contribution to H, so private adoption exceeds social optima. The quantitative magnitude of the wedge changes, but the qualitative inefficiency result is robust.

Binary adoption (minimal model). The task continuum and effort concentration parameter γ provide a convenient microfoundation for interior adoption, but the bias decomposition does not depend on them. Consider a stripped-down model with binary adoption $\alpha \in \{0, 1\}$, output $Y(\alpha, h) = (1 - \alpha)f(h) + \alpha A$ for some increasing concave f, and learning $h_{t+1} = (1-\delta)h_t + \lambda[(1-\alpha) + \mu\alpha]\varphi(h_t)$. Define the same counterfactual objects as in the main text. Then for any worker who adopts ($\alpha = 1$) with $\mu < 1$: (i) the state-conditional gain $\Delta_t^{SC} = A - f(h_t^U)$ is increasing in t as h_t^U falls; (ii) the path comparison $\Delta_t^{LR} = A - f(h_t^{NA})$ is decreasing in t since $h_t^{NA} = \bar{h} > h_t^U$; (iii) the wedge $\Delta_t^{SC} - \Delta_t^{LR} = f(h_t^{NA}) - f(h_t^U) > 0$ equals the state-gap bias and is strictly increasing. With spillovers, the cross-sectional bias $\Delta_t^{CS} - \Delta_t^{LR} = f(h_t^{NA}) - f(h_t^{NU}) > 0$ adds a second component. The bias decomposition (Remark 3) holds verbatim. The task continuum adds interior adoption and comparative statics in α, but the measurement critique requires only that AI affects skill formation.

Discrete tasks. Replace the continuum of tasks with a finite set $\{1, 2, \ldots, J\}$. Workers choose which tasks to delegate rather than a continuous adoption intensity. The analysis becomes combinatorially more complex, but for large J the continuous approximation is accurate. For small J, the model admits multiple equilibria with different task allocations, but each equilibrium exhibits the same qualitative properties: delegation of learning-intensive tasks reduces skill accumulation when AI substitutes for learning.

Heterogeneous pedagogical quality $\mu(h)$. The baseline model assumes a constant μ, but pedagogical quality plausibly varies with skill level. We analyze two cases:

The learning function becomes $L(\alpha, h) = [1 - (1 - \mu(h))\alpha]\varphi(h)$. Differentiating the steady-state condition $\delta h^* = \lambda[1 - (1 - \mu(h^*))\alpha]\varphi(h^*)$ with respect to α:

$$\frac{dh^*}{d\alpha} = \frac{-(1 - \mu(h^*))\lambda\varphi(h^*)}{\delta - \lambda[1 - (1 - \mu(h^*))\alpha]\varphi'(h^*) - \lambda\alpha\mu'(h^*)\varphi(h^*)}$$

Note the critical minus sign before the $\mu'(h^*)$ term, arising from implicit differentiation of $(1 - \mu(h^*))\alpha$ with respect to h^*.

Case 1: $\mu'(h) > 0$ (AI is more pedagogical for experts). This captures the intuition that novices may lack the framework to learn from AI outputs, while experts can critically evaluate and integrate AI suggestions. When $\mu'(h^*) > 0$, the term $-\lambda\alpha\mu'(h^*)\varphi(h^*)$ is *negative*, making the denominator smaller and $|dh^*/d\alpha|$ larger. Skill atrophy is *amplified*: as skills fall, AI becomes less pedagogical (since μ falls with h), which accelerates further skill loss. This creates a destabilizing force that deepens the trap.

Case 2: $\mu'(h) < 0$ (AI is more pedagogical for novices). This captures the intuition that AI scaffolding is most helpful for beginners, while advanced learners need unassisted struggle. Now the term $-\lambda\alpha\mu'(h^*)\varphi(h^*)$ is *positive*, making the denominator larger and $|dh^*/d\alpha|$ smaller. Skill atrophy is *dampened*: as skills fall, AI becomes more pedagogical, reducing the marginal harm from adoption. This creates a stabilizing force that limits the depth of the trap but does not eliminate it: as long as $\mu(h^*) < 1$ at the equilibrium skill level, the trap can still occur.

The key insight is that allowing $\mu(h)$ to vary introduces a feedback between skill level and the learning effect of adoption, but does not qualitatively change the main results unless $\mu(h) \geq 1$ for all h (which would eliminate skill atrophy entirely). The scalar μ in our baseline model can be interpreted as the value at the relevant equilibrium skill level: $\mu \equiv \mu(h^*)$.

Upper-tail spillover specification. As noted in the main text, the microfoundation in Appendix B.2 implies spillovers that depend on the full skill distribution, not merely the

mean. We verify robustness to an alternative specification where spillovers depend on the upper tail:

$$\tilde{\psi}(G_H) = \psi_0 + \psi_1 \cdot [1 - G_H(h^{threshold})]$$

where $h^{threshold}$ is a fixed mentorship threshold and $1 - G_H(h^{threshold})$ is the fraction of workers above it. This specification captures the spirit of Kremer (1993): when tasks require complementary skills across team members, the marginal worker falling below the threshold imposes disproportionate losses on all others. As AI adoption causes skills to atrophy, more workers fall below the threshold, reducing $\tilde{\psi}$ and impairing learning for all workers. The comparative statics are identical to the mean-based specification: $\partial\tilde{\psi}/\partial\alpha < 0$ when $\mu < 1$, generating overadoption.

C Proofs

This appendix provides formal proofs for all results in the main text. Section C.1 states and proves technical lemmas; Section C.2 proves the main results. The skill trap proof (Proposition 3) appears before the spillover and state-path divergence proofs because Part (ii) of the latter references the trap characterization. Otherwise, proofs follow the order of the main text. Proofs for Appendix A results appear at the end of this section.

C.1 Technical Lemmas

The Firm's Problem. Recall from Section 2 that the firm maximizes (4) subject to the human capital law of motion (2), with the value function satisfying the Bellman equation (5).

Lemma 5 (Optimal Effort Allocation)**.** *Given adoption intensity $\alpha \in [0, 1)$, the worker optimally spreads effort uniformly across worker-performed tasks: $e(j) = 1/(1-\alpha)$ for $j \in (\alpha, 1]$. This yields worker output $h(1-\alpha)^{1-\gamma}$.*

Proof. The worker chooses effort allocation $e(j)$ for $j \in (\alpha, 1]$ to maximize $\int_\alpha^1 h \cdot e(j)^\gamma \, dj$ subject to $\int_\alpha^1 e(j)\, dj = 1$. The FOC implies constant effort $e(j) = 1/(1-\alpha)$. Total output is $\int_\alpha^1 h[1/(1-\alpha)]^\gamma \, dj = (1-\alpha) \cdot h \cdot (1-\alpha)^{-\gamma} = h(1-\alpha)^{1-\gamma}$. □

Lemma 6 (Output and Learning Properties)**.** *The output function $Y(h, \alpha; A) = A \cdot G(\alpha) + h(1-\alpha)^{1-\gamma}$ is linear in h, strictly concave in α for $h > 0$, and satisfies $\partial Y/\partial \alpha \to -\infty$ as $\alpha \to 1^-$. The learning effect satisfies $\partial L/\partial \alpha = (\mu - 1)\varphi(h)$, which is negative iff $\mu < 1$.*

Proof. Concavity of Y: $Y_{\alpha\alpha} = Ag'(\alpha) - h(1-\gamma)\gamma(1-\alpha)^{-\gamma-1} < 0$ since $g'(\alpha) < 0$. As $\alpha \to 1$, $(1-\alpha)^{-\gamma} \to \infty$, so $Y_\alpha \to -\infty$. The learning derivative follows directly from $L(\alpha, h; \mu) = [(1-\alpha) + \mu\alpha]\varphi(h)$. □

Lemma 7 (Value Function Properties)**.** *The value function V exists, is unique, continuous, strictly increasing, concave, and continuously differentiable on $(0, \infty)$.*

Proof. Human capital is bounded above by $\bar{h}$. Existence and uniqueness follow from Theorem 4.6 (Contraction Mapping) of Stokey and Lucas (1989); differentiability from Benveniste-Scheinkman (Theorem 4.11). □

Lemma 8 (Optimal Adoption is Interior)**.** *Under Assumption 3, $\alpha^*(h) \in (0, 1)$ for all $h \in (0, \bar{h}]$.*

Proof. At $\alpha \to 1$: $\partial Y/\partial\alpha \to -\infty$ (Lemma 6), so $\alpha^* < 1$.

At $\alpha = 0$: the full marginal value of adoption in the dynamic problem is

$$\frac{\partial}{\partial\alpha}\{Y(h,\alpha) + \beta V(h')\}\Big|_{\alpha=0} = [A \cdot g(0) - h(1-\gamma)] + \beta V'(h')\lambda(\mu - 1)\varphi(h)$$

The first bracket is the static marginal benefit; the second term is the discounted marginal learning cost (negative when $\mu < 1$). At $h = \bar{h}$ with $\alpha = 0$, we have $h' = \bar{h}$ (steady state), so $V'(h') = \bar{V}'$. Assumption 3 ensures the static benefit exceeds the dynamic cost: $A \cdot g(0) - \bar{h}(1-\gamma) > \beta\bar{V}'\lambda(1-\mu)\varphi(\bar{h})$. For $h < \bar{h}$, the static benefit $A \cdot g(0) - h(1-\gamma)$ is larger (since h is smaller), while the dynamic cost $\beta V'(h')\lambda(1-\mu)\varphi(h)$ is bounded. Thus the total marginal value at $\alpha = 0$ is positive for all $h \in (0, \bar{h}]$, implying $\alpha^* > 0$. □

Lemma 9 (Stability Characterization)**.** *At a steady state h^*, local stability holds when $|T'(h^*)| < 1$, where $T'(h^*) = (1-\delta) + \lambda\ell(\alpha^*)\varphi'(h^*) + \lambda\ell'(\alpha^*)\frac{d\alpha^*}{dh}\varphi(h^*)$. Under Assumption 1, a sufficient condition is $\delta - \lambda\ell(\alpha^*)|\varphi'(h^*)| > 0$: the stability term dominates the policy feedback term, which is bounded under curvature dominance.*

Proof. The transition is $T(h) = (1-\delta)h + \lambda\ell(\alpha^*(h))\varphi(h)$. Differentiating:

$$T'(h) = (1-\delta) + \lambda\ell(\alpha^*(h))\varphi'(h) + \lambda\ell'(\alpha^*(h))\frac{d\alpha^*}{dh}\varphi(h)$$

The first two terms give $(1-\delta) + \lambda\ell(\alpha^*)\varphi'(h^*)$. Since $\varphi'(h^*) < 0$ by Assumption 1, this is less than $(1-\delta) < 1$. The third term – the policy feedback – has magnitude bounded under Assumption 1: the bound on φ' ensures $|d\alpha^*/dh|$ is bounded. Combining, $|T'(h^*)| < 1$ when $\delta - \lambda\ell(\alpha^*)|\varphi'(h^*)| > 0$. □

Lemma 10 (Convergence to Steady State)**.** *Under optimal policy with $\mu < 1$, if $h_0 \in (0, \bar{h}]$, then $h_t \to h^* \in (0, \bar{h})$ as $t \to \infty$.*

Proof. Define the transition map $T(h) = (1-\delta)h + \lambda\ell(\alpha^*(h))\varphi(h)$ where $\alpha^*(h)$ is the optimal policy. A steady state h^* satisfies $T(h^*) = h^*$, i.e., $\delta h^* = \lambda\ell(\alpha^*)\varphi(h^*)$.

Step 1: Existence and location of steady state. By Lemma 2, there exists a unique $h^* > 0$ satisfying the stationarity condition. By Lemma 3, $h^* \in (0, \bar{h})$ when $\mu < 1$.

Step 2: Local stability. The derivative $T'(h^*) = (1-\delta) + \lambda\ell(\alpha^*)\varphi'(h^*) + \lambda\ell'(\alpha^*)\frac{d\alpha^*}{dh}\varphi(h^*)$. Under Assumption 1, $(1-\delta) + \lambda\ell(\alpha^*)\varphi'(h^*) < 1$. The third term is bounded by the same contraction condition. Thus $|T'(h^*)| < 1$, establishing local asymptotic stability.

Step 3: Global convergence from $(0, \bar{h}]$. For $h \in (0, \bar{h}]$, we show $T(h) - h$ has constant sign on each side of h^*. At $h = \bar{h}$: $T(\bar{h}) = (1-\delta)\bar{h} + \lambda\ell(\alpha^*(\bar{h}))\varphi(\bar{h})$. Since $\ell(\alpha) < 1$ when

$\alpha > 0$ and $\mu < 1$, and since $\delta\bar{h} = \lambda\varphi(\bar{h})$ defines $\bar{h}$, we have $T(\bar{h}) < \bar{h}$. At h^*: $T(h^*) = h^*$. By Lemma 2, h^* is the unique steady state on $(0, \bar{h}]$, so T has no other fixed point in this interval. By continuity, $T(h) < h$ for all $h \in (h^*, \bar{h}]$, so the sequence is decreasing and bounded below by h^*. Local stability then implies $h_t \to h^*$. □

Lemma 11 (Jacobian Non-Singularity)**.** *At an interior steady state (h^*, α^*) with $\mu < 1$, the Jacobian of the steady-state system is non-singular with* $\det(\mathbf{J}) \neq 0$.

Proof. The steady-state system comprises the stationarity condition $F^1(h, \alpha) \equiv \delta h - \lambda\ell(\alpha)\varphi(h) = 0$ and the FOC $F^2(h, \alpha) \equiv Y_\alpha + \beta V'(h')\lambda(\mu - 1)\varphi(h) = 0$. The Jacobian is:

$$\mathbf{J} = \begin{pmatrix} \partial F^1/\partial h & \partial F^1/\partial \alpha \\ \partial F^2/\partial h & \partial F^2/\partial \alpha \end{pmatrix} = \begin{pmatrix} D_h & D_{h\alpha} \\ D_{\alpha h} & D_\alpha \end{pmatrix}$$

where:

- $D_h = \delta - \lambda\ell(\alpha)\varphi'(h) > 0$ by Assumption 1
- $D_{h\alpha} = \lambda(1 - \mu)\varphi(h) > 0$ since $\mu < 1$ and $\varphi(h) > 0$
- $D_\alpha = Y_{\alpha\alpha} + \beta V''(h')[\lambda(\mu - 1)\varphi(h)]^2$
- $D_{\alpha h} = -(1 - \gamma)(1 - \alpha)^{-\gamma} + \beta\lambda(\mu - 1)\left[V''(h')\frac{\partial h'}{\partial h}\varphi(h) + V'(h')\varphi'(h)\right]$

Signing D_α: The first term $Y_{\alpha\alpha} = Ag'(\alpha) - h(1-\gamma)\gamma(1-\alpha)^{-\gamma-1} < 0$ by strict concavity of output in α. The second term $\beta V''(h')[\lambda(\mu - 1)\varphi(h)]^2 \leq 0$ by concavity of V. Thus $D_\alpha < 0$ unconditionally – no additional assumption is needed. (Note: since we take a partial derivative with respect to α holding h fixed, the term $\varphi(h)$ does not contribute a $\varphi'(h)$ factor.)

Signing $D_{\alpha h}$: Differentiating $F^2(h, \alpha) = Y_\alpha + \beta V'(h')\lambda(\mu - 1)\varphi(h)$ with respect to h:

$$D_{\alpha h} = -(1 - \gamma)(1 - \alpha)^{-\gamma} + \beta\lambda(\mu - 1)\left[V''(h')\frac{\partial h'}{\partial h}\varphi(h) + V'(h')\varphi'(h)\right]$$

Note that $\frac{\partial h'}{\partial h}$ multiplies only the $V''(h')$ term, not the $V'(h')\varphi'(h)$ term – this follows from the chain rule since $V'(h')$ depends on h through h', while $\varphi'(h)$ depends directly on h. The first term $-(1 - \gamma)(1 - \alpha)^{-\gamma} < 0$. For the bracketed expression when $\mu < 1$: $V''(h') \leq 0$ by concavity, $\frac{\partial h'}{\partial h} > 0$, $\varphi(h) > 0$, so the first bracketed term is non-positive. For the second term: $V'(h') > 0$, $\varphi'(h) < 0$ by Assumption 1, so $V'(h')\varphi'(h) < 0$. Thus the bracket is

non-positive. With $(\mu - 1) < 0$, we have $\beta\lambda(\mu - 1) \cdot (\text{non-positive}) \geq 0$, making the second term non-negative. The sign of $D_{\alpha h}$ depends on which effect dominates.

Non-singularity of J: We have $\det(\mathbf{J}) = D_h D_\alpha - D_{h\alpha} D_{\alpha h}$. The first term $D_h D_\alpha < 0$ since $D_h > 0$ and $D_\alpha < 0$. The second term equals $D_{h\alpha} \cdot D_{\alpha h}$ where $D_{h\alpha} > 0$.

Since $d\alpha^*/dh < 0$ when $\mu < 1$ (shown above), $D_{\alpha h}$ has constant sign on $(0, \bar{h}]$. If $D_{\alpha h} > 0$, then $D_{h\alpha} D_{\alpha h} > 0$ and hence $-D_{h\alpha} D_{\alpha h} < 0$, so both terms in $\det(\mathbf{J}) = D_h D_\alpha - D_{h\alpha} D_{\alpha h}$ are negative and $\det(\mathbf{J}) < 0$ unconditionally. If $D_{\alpha h} \leq 0$, then $D_{h\alpha} D_{\alpha h} \leq 0$ and hence $-D_{h\alpha} D_{\alpha h} \geq 0$, so $\det(\mathbf{J})$ is the sum of a negative term ($D_h D_\alpha < 0$) and a non-negative term ($-D_{h\alpha} D_{\alpha h} \geq 0$); in this case, the sign of $\det(\mathbf{J})$ is ambiguous unless we impose $|D_h D_\alpha| > |D_{h\alpha} D_{\alpha h}|$. This latter condition holds because $|g'(\alpha)|$ is bounded away from zero, which ensures $|Y_{\alpha\alpha}|$ dominates the cross-partial products at any $\beta < 1$.

In either case, $\det(\mathbf{J}) \neq 0$ and the implicit function theorem applies. □

C.2 Proofs of Main Results

Lemma 1 (Role of Pedagogical Quality).

The firm's Bellman equation is $V(h) = \max_\alpha \{Y(h, \alpha; A) + \beta V(h')\}$ where $h' = (1 - \delta)h + \lambda L(\alpha, h; \mu)$. The first-order condition for an interior $\alpha \in (0, 1)$ is:

$$\frac{\partial Y}{\partial \alpha} + \beta V'(h') \cdot \frac{\partial h'}{\partial \alpha} = 0$$

Substituting the derivatives and rearranging:

$$A \cdot g(\alpha) - h(1 - \gamma)(1 - \alpha)^{-\gamma} = \beta V'(h') \cdot \lambda(1 - \mu)\varphi(h)$$

The LHS is the marginal output benefit; the RHS is the marginal learning cost. Since $V'(h') > 0$, $\lambda > 0$, and $\varphi(h) > 0$, the marginal learning cost is positive iff $\mu < 1$. For part (i): when $\mu < 1$, firms face a positive marginal cost through learning. For part (ii): when $\mu = 1$, the RHS is zero. For part (iii): when $\mu > 1$, the RHS is negative. For the comparative static $\partial\alpha^*/\partial\mu > 0$: by the implicit function theorem, $d\alpha^*/d\mu = \beta V'(h')\lambda\varphi(h)/(-Y_{\alpha\alpha} + \cdots) > 0$. □

Lemma 2 (Steady-State Human Capital Function).

(i) Define $\Phi(h; \alpha) \equiv \delta h - \lambda \ell(\alpha)\varphi(h)$ where $\ell(\alpha) = 1 - (1 - \mu)\alpha$. For existence, we require $\ell(\alpha) > 0$, which holds for all $\alpha \in [0, 1)$ when $\mu \geq 0$ since $\ell(\alpha) \geq 1 - \alpha > 0$.

At $h = 0$: $\Phi(0;\alpha) = -\lambda\ell(\alpha)\varphi(0) < 0$ since $\ell(\alpha) > 0$ and $\varphi(0) > 0$. As $h \to \infty$: $\Phi(h;\alpha) \to \infty$ since δh grows without bound while $\lambda\ell(\alpha)\varphi(h) \to 0$ by Assumption 1. By continuity and the intermediate value theorem, at least one solution exists.

For uniqueness, note that $\varphi'(h) < 0$ for all $h > 0$ by Assumption 1, so $\frac{\partial\Phi}{\partial h} = \delta - \lambda\ell(\alpha)\varphi'(h) > \delta > 0$. Thus Φ is strictly increasing for all $h > 0$. Since $\Phi(h) \to -\lambda\ell(\alpha)\varphi(0) < 0$ as $h \to 0^+$ (using $\varphi(0) > 0$) and $\Phi(h) \to \infty$ as $h \to \infty$, by continuity there is exactly one crossing of zero.

(ii) At $\alpha = 0$: $\ell(0) = 1$, so (7) becomes $\delta h = \lambda\varphi(h)$, which defines $\bar{h}$.

(iii)–(iv) Implicitly differentiating (7):

$$\frac{dh^*}{d\alpha} = \frac{\lambda\ell'(\alpha)\varphi(h^*)}{\delta - \lambda\ell(\alpha)\varphi'(h^*)}$$

The denominator is positive at a stable steady state. Since $\ell'(\alpha) = -(1-\mu)$, the numerator has sign opposite to $(1-\mu)$. Thus $\frac{dh^*}{d\alpha} < 0$ when $\mu < 1$ and $\frac{dh^*}{d\alpha} \geq 0$ when $\mu \geq 1$. □

Lemma 3 (Steady-State Characterization).

The characterization follows directly from the properties of the steady-state human capital function $h^*(\alpha)$ established in Lemma 2. □

Lemma 4 (Uniqueness and Global Stability).

Part (i): Existence. Define the equilibrium system as the intersection of two curves in (h, α) space:

- The *stationarity locus* S: pairs (h, α) satisfying $\delta h = \lambda\ell(\alpha)\varphi(h)$.
- The *optimal policy* P: pairs $(h, \alpha^*(h))$ where $\alpha^*(h)$ solves the firm's problem.

For the stationarity locus S: fixing α, there exists a unique $h(\alpha)$ by Lemma 2. As α increases (with $\mu < 1$), $\ell(\alpha) = 1 - (1-\mu)\alpha$ decreases, so stationarity requires lower h. Thus $h_S(\alpha)$ is decreasing with $h_S(0) = \bar{h}$ and $h_S(\alpha) \to 0$ as $\alpha \to 1$.

For the optimal policy P: by Lemma 8, at each $h > 0$ there exists an interior optimal adoption $\alpha^*(h) \in (0, 1)$. The policy function $\alpha^*(h)$ is monotone decreasing in h when $\mu < 1$ (proved below): higher skill reduces the marginal benefit of AI relative to the learning cost.

Both loci are decreasing in (h, α) space. However, they have different boundary behavior that guarantees a unique crossing:

- At h close to 0: The stationarity condition $\delta h = \lambda\ell(\alpha)\varphi(h)$ with $\varphi(0) > 0$ requires α close to $1/(1-\mu) > 1$ for $\mu \in (0, 1)$, which is outside $[0, 1]$. Thus for any $\alpha \in [0, 1]$,

stationarity requires $h > 0$. Meanwhile, the optimal policy has $\alpha^*(h) \to \alpha^{max} < 1$ as $h \to 0$ (AI remains valuable even at low skill).

- At $h = \bar{h}$: Stationarity with $\alpha = 0$ gives $\delta\bar{h} = \lambda\varphi(\bar{h})$, which defines $\bar{h}$. Thus $h_S(0) = \bar{h}$. The optimal policy has $\alpha^*(\bar{h}) > 0$ by Assumption 3.

At $\alpha = 0$: stationarity gives $h = \bar{h}$, while optimal adoption at $\bar{h}$ is $\alpha^*(\bar{h}) > 0$. Thus at this boundary, $\alpha_P > \alpha_S$. As h decreases from $\bar{h}$, both $\alpha_S(h)$ and $\alpha_P(h)$ increase (moving along their respective decreasing curves in the other direction), but at different rates. Since α_S must reach infeasibly high values as $h \to 0$ while α_P remains bounded, and since both are continuous, they must cross exactly once.

Part (ii): Uniqueness. The Jacobian non-singularity established in Lemma 11 implies local uniqueness via the implicit function theorem. For global uniqueness, note that any steady state must lie on both loci, and the boundary analysis above shows there is exactly one such point.

Part (iii): Global Stability. By Lemma 10, for any $h_0 \in (0, \bar{h}]$, the skill path $h_t \to h^*$ as $t \to \infty$. By continuity of the optimal policy $\alpha^*(h)$, the adoption path $\alpha_t = \alpha^*(h_t) \to \alpha^*(h^*) = \alpha^*$.

Part (iv): Monotonicity of Optimal Paths. Suppose $\mu < 1$ and $h_0 = \bar{h}$. We show $\{h_t\}$ is strictly decreasing and $\{\alpha_t\}$ is strictly increasing.

Step 1: The policy function is strictly decreasing. We show $d\alpha^*/dh < 0$ when $\mu < 1$. The FOC for optimal adoption is $Y_\alpha(h, \alpha) + \beta V'(h') \cdot \partial h'/\partial\alpha = 0$. The cross-partial $\partial^2/\partial h\partial\alpha$ of the Bellman objective includes the term $-(1-\gamma)(1-\alpha)^{-\gamma} < 0$ from $Y_{h\alpha}$. Since higher h raises output more when α is lower, and since the dynamic cost $\beta V'(h')\lambda(1-\mu)\varphi(h)$ is positive when $\mu < 1$, the optimal response to higher h is lower α. Thus $d\alpha^*/dh < 0$.

Step 2: Skills are strictly decreasing. At $h_0 = \bar{h}$, the optimal adoption $\alpha_0 = \alpha^*(\bar{h}) > 0$ by Assumption 3. With $\alpha_0 > 0$ and $\mu < 1$, learning is $L_0 = \ell(\alpha_0)\varphi(\bar{h}) < \varphi(\bar{h})$ since $\ell(\alpha) = 1 - (1-\mu)\alpha < 1$. But $\bar{h}$ is defined by $\delta\bar{h} = \lambda\varphi(\bar{h})$, so:

$$h_1 = (1-\delta)\bar{h} + \lambda L_0 < (1-\delta)\bar{h} + \lambda\varphi(\bar{h}) = (1-\delta)\bar{h} + \delta\bar{h} = \bar{h}$$

Thus $h_1 < h_0$. By induction, $h_{t+1} < h_t$ for all t until $h_t = h^*$.

Step 3: Adoption is strictly increasing. Since $\alpha_t = \alpha^*(h_t)$ and $d\alpha^*/dh < 0$, the sequence $\{\alpha_t\}$ inherits the opposite monotonicity from $\{h_t\}$. As h_t decreases, α_t increases. Convergence $h_t \to h^*$ implies $\alpha_t \to \alpha^*$. □

Necessity of Substitution for Skill Atrophy.

When $\mu \geq 1$, the learning function satisfies $\frac{\partial L}{\partial \alpha} = (\mu - 1)\varphi(h) \geq 0$ by Lemma 6. Higher adoption does not reduce learning – it either leaves learning unchanged ($\mu = 1$) or increases it ($\mu > 1$).

Consider the steady-state condition $\delta h^* = \lambda[1 - (1 - \mu)\alpha^*]\varphi(h^*)$. When $\mu \geq 1$, the term $[1 - (1 - \mu)\alpha^*] \geq 1$ for all $\alpha^* \in [0, 1]$. Thus:

$$\delta h^* \geq \lambda\varphi(h^*)$$

with equality only when $\mu = 1$ (for any α^*) or when $\mu > 1$ and $\alpha^* = 0$.

The right side $\lambda\varphi(h)$ intersects δh at the no-adoption steady state $\bar{h}$. Since $\delta h^* \geq \lambda\varphi(h^*)$, the steady-state human capital must satisfy $h^* \geq \bar{h}$. Human capital cannot fall below the no-adoption level regardless of adoption intensity.

By Definition 3, the skill trap requires $Y_t < Y_t^{NA}$ for large t. With $h^* \geq \bar{h}$, long-run human capital under adoption weakly exceeds the no-adoption level. For the trap to be impossible, we need $Y^* \geq Y^{NA} = \bar{h}$.

Now, $Y^* = A \cdot G(\alpha^*) + h^*(1 - \alpha^*)^{1-\gamma}$. Note that $(1 - \alpha^*)^{1-\gamma} < 1$ for $\alpha^* > 0$ since $1 - \gamma \in (0, 1)$. Since $h^* \geq \bar{h}$ and $(1 - \alpha^*)^{1-\gamma} < 1$, we have $h^*(1 - \alpha^*)^{1-\gamma} < h^*$. For $Y^* \geq \bar{h}$, it suffices to show $A \cdot G(\alpha^*) \geq \bar{h} - h^*(1 - \alpha^*)^{1-\gamma}$. Since $h^* \geq \bar{h}$, we have:

$$\bar{h} - h^*(1 - \alpha^*)^{1-\gamma} \leq \bar{h} - \bar{h}(1 - \alpha^*)^{1-\gamma} = \bar{h}[1 - (1 - \alpha^*)^{1-\gamma}]$$

Thus it suffices that $A \cdot G(\alpha^*) \geq \bar{h}[1 - (1 - \alpha^*)^{1-\gamma}]$.

When $\mu \geq 1$, the steady-state FOC implies $Y_\alpha(h^*, \alpha^*) \geq 0$ (the dynamic skill cost is non-positive). At any point where $Y(h, \alpha) < \bar{h}$ with $h \geq \bar{h}$, we have $Y_\alpha(h, \alpha) < 0$ (since the marginal output from adoption must be negative for output to fall below $\bar{h}$ starting from skill at least $\bar{h}$). But $Y_\alpha(h^*, \alpha^*) \geq 0$ by the FOC when $\mu \geq 1$, so we cannot have $Y^* < \bar{h}$. Thus $Y^* \geq \bar{h} = Y^{NA}$, and the trap cannot exist when $\mu \geq 1$. □

Corollary 1 (Comparative Statics).

By the implicit function theorem, $\frac{\partial \mathbf{x}}{\partial \theta_i} = -\mathbf{J}^{-1}\frac{\partial \mathbf{F}}{\partial \theta_i}$ for each parameter θ_i. By Lemma 11, $\det(\mathbf{J}) \neq 0$. Under the conditions established in that lemma's proof, $\det(\mathbf{J}) < 0$.

(i) Effect of A: $\frac{\partial F_1}{\partial A} = 0$ and $\frac{\partial F_2}{\partial A} = g(\alpha^*) > 0$. Computing:

$$\frac{\partial \alpha^*}{\partial A} = \frac{D_h \cdot g(\alpha^*)}{-\det(\mathbf{J})} > 0$$

where $D_h = \delta - \lambda\ell(\alpha^*)\varphi'(h^*) > 0$. From stationarity: $\frac{\partial h^*}{\partial A} = -\frac{D_{h\alpha}}{D_h}\frac{\partial \alpha^*}{\partial A} < 0$.

(ii) Effect of β: $\frac{\partial F_1}{\partial \beta} = 0$ and $\frac{\partial F_2}{\partial \beta} = -V'(h^*)\lambda(1-\mu)\varphi(h^*) < 0$. By analogous calculation, $\frac{\partial \alpha^*}{\partial \beta} < 0$ and $\frac{\partial h^*}{\partial \beta} > 0$. This uses the fact that $V'(h^*) > 0$ (human capital is valuable) and that $V'(h^*)$ is increasing in β – more patient firms place higher marginal value on future human capital. Formally, from the envelope condition $V'(h) = (1-\alpha)^{1-\gamma} + \beta V'(h')[(1-\delta) + \lambda\ell(\alpha)\varphi'(h)]$, higher β raises $V'(h)$ at each h.

(iii) Effect of λ: Both partial derivatives are negative when $\mu < 1$. Cramer's rule gives $\frac{\partial h^*}{\partial \lambda} > 0$.

(iv) Effect of μ: For $\partial\alpha^*/\partial\mu > 0$: higher μ reduces the learning cost term $(1-\mu)\varphi(h)$ in the FOC, so firms adopt more.

For $\partial h^*/\partial\mu$: implicitly differentiate the stationarity condition $\delta h^* = \lambda[1-(1-\mu)\alpha^*]\varphi(h^*)$:

$$\delta\frac{\partial h^*}{\partial \mu} = \lambda\alpha^*\varphi(h^*) + \lambda[1-(1-\mu)\alpha^*]\varphi'(h^*)\frac{\partial h^*}{\partial \mu} - \lambda(1-\mu)\varphi(h^*)\frac{\partial \alpha^*}{\partial \mu}$$

Solving:

$$\frac{\partial h^*}{\partial \mu} = \frac{\lambda\alpha^*\varphi(h^*) - \lambda(1-\mu)\varphi(h^*)\frac{\partial \alpha^*}{\partial \mu}}{\delta - \lambda\ell(\alpha^*)\varphi'(h^*)}$$

The denominator is positive by Assumption 1. The numerator has two terms:

- Direct effect: $\lambda\alpha^*\varphi(h^*) > 0$. Higher μ means more learning per unit of AI-assisted work.
- Indirect effect: $-\lambda(1-\mu)\varphi(h^*)\frac{\partial \alpha^*}{\partial \mu} < 0$. Higher μ induces more adoption ($\partial\alpha^*/\partial\mu > 0$), which reduces learning.

The sign of $\partial h^*/\partial\mu$ is thus ambiguous in general. However, $\partial h^*/\partial\mu > 0$ when the direct effect dominates:

$$\alpha^* > (1-\mu)\frac{\partial \alpha^*}{\partial \mu}$$

This holds when adoption responses to μ are moderate. In our calibrations with $\mu \in [0.3, 0.5]$ and $\alpha^* \approx 0.5$, this condition is satisfied and $\partial h^*/\partial\mu > 0$. Intuitively, when μ is substantially below 1, the direct benefit of better learning quality outweighs the indirect cost of induced adoption. □

Proposition 3 (Existence of Skill Trap).

We verify each condition of Definition 3 and establish uniqueness of $\bar{\beta}$.

Step 1: Condition (T1) holds. By Assumption 3, $A > \bar{h}(1-\gamma)$. By Lemma 8,

$\alpha^*(h) > 0$ for all $h \in (0, \bar{h}]$. Since $h_0 \leq \bar{h}$ and human capital remains bounded in $(0, \bar{h}]$ along any equilibrium path (Lemma 7), we have $\alpha_t > 0$ for all t.

Step 2: Short-run gain. At $t = 0$, consider the adoption decision. No-adoption output is $Y_0^{NA} = h_0$. With adoption $\alpha_0 > 0$:

$$Y_0 = A \cdot G(\alpha_0) + h_0(1 - \alpha_0)^{1-\gamma}$$

Differentiating at $\alpha_0 = 0$: $\partial Y_0/\partial \alpha|_{\alpha=0} = A \cdot g(0) - h_0(1-\gamma) = A - h_0(1-\gamma) > 0$ by Assumption 3. Since $Y(h_0, \cdot\,; A)$ is strictly concave in α (Lemma 6) with $Y_\alpha(h_0, 0) > 0$, we have $Y(h_0, \alpha, A) > Y(h_0, 0, A)$ for all $\alpha \in (0, \alpha^{peak})$ where $\alpha^{peak} = \arg\max_\alpha Y(h_0, \alpha, A)$ is the static maximizer. When $\mu < 1$, the dynamic skill cost further restricts adoption: $\alpha_0^* \leq \alpha^{peak}$. Thus $Y_0 > Y_0^{NA}$.

Step 3: Monotonicity of steady-state output in β. Define $W(\alpha) \equiv A \cdot G(\alpha) + h^*(\alpha)(1-\alpha)^{1-\gamma}$ as steady-state output as a function of adoption. We show $W'(\alpha^*) < 0$. Throughout, we restrict attention to interior steady states where the policy correspondence $\alpha^*(h)$ is single-valued and continuously differentiable; this is guaranteed under Assumptions 1–3 by Lemma 8 and the implicit function theorem.

From the stationarity condition $\delta h^* = \lambda \ell(\alpha)\varphi(h^*)$, implicit differentiation yields:

$$\frac{dh^*}{d\alpha} = \frac{-\lambda(1-\mu)\varphi(h^*)}{\delta - \lambda \ell(\alpha)\varphi'(h^*)} \tag{26}$$

The denominator is positive at a stable steady state (Lemma 9). When $\mu < 1$, the numerator is negative, so $dh^*/d\alpha < 0$.

Differentiating W:

$$W'(\alpha) = Ag(\alpha) + \frac{dh^*}{d\alpha}(1-\alpha)^{1-\gamma} - h^*(1-\gamma)(1-\alpha)^{-\gamma} \tag{27}$$

From the steady-state FOC: $Ag(\alpha^*) = h^*(1-\gamma)(1-\alpha^*)^{-\gamma} + \beta V'(h^*)\lambda(1-\mu)\varphi(h^*)$. Substituting:

$$\begin{aligned} W'(\alpha^*) &= h^*(1-\gamma)(1-\alpha^*)^{-\gamma} + \beta V'(h^*)\lambda(1-\mu)\varphi(h^*) \\ &\quad + \frac{dh^*}{d\alpha}(1-\alpha^*)^{1-\gamma} - h^*(1-\gamma)(1-\alpha^*)^{-\gamma} \\ &= \underbrace{\beta V'(h^*)\lambda(1-\mu)\varphi(h^*)}_{>0} + \underbrace{\frac{dh^*}{d\alpha}(1-\alpha^*)^{1-\gamma}}_{<0} \end{aligned}$$

The first term is positive ($V'(h^*) > 0$ by Lemma 7, and all other factors positive when $\mu < 1$); the second is negative since $dh^*/d\alpha < 0$. The sign of $W'(\alpha^*)$ is thus ambiguous in general. To resolve this ambiguity, we derive $V'(h^*)$ explicitly.

Derivation of $V'(h^*)$. At steady state, the envelope theorem applied to the Bellman equation (5) yields:

$$V'(h) = \frac{\partial Y}{\partial h} + \beta V'(h') \cdot \frac{\partial h'}{\partial h}$$

where $\partial Y/\partial h = (1-\alpha)^{1-\gamma}$ and $\partial h'/\partial h = (1-\delta) + \lambda \ell(\alpha)\varphi'(h)$. At steady state $h' = h^*$, so:

$$V'(h^*) = (1-\alpha^*)^{1-\gamma} + \beta V'(h^*)\left[(1-\delta) + \lambda \ell(\alpha^*)\varphi'(h^*)\right]$$

Solving for $V'(h^*)$:

$$V'(h^*) = \frac{(1-\alpha^*)^{1-\gamma}}{1 - \beta(1-\delta) - \beta\lambda\ell(\alpha^*)\varphi'(h^*)} \tag{28}$$

The denominator can be rewritten as $(1-\beta) + \beta[\delta - \lambda\ell(\alpha^*)\varphi'(h^*)]$. Since $\varphi'(h^*) < 0$ by Assumption 1, the term $\delta - \lambda\ell(\alpha^*)\varphi'(h^*) > \delta > 0$, so the denominator is strictly positive.

Substituting into $W'(\alpha^*)$. Recall from (26):

$$\frac{dh^*}{d\alpha} = \frac{-\lambda(1-\mu)\varphi(h^*)}{\delta - \lambda\ell(\alpha^*)\varphi'(h^*)}$$

Substituting (28) and this expression into $W'(\alpha^*)$:

$$W'(\alpha^*) = \frac{\beta(1-\alpha^*)^{1-\gamma}\lambda(1-\mu)\varphi(h^*)}{(1-\beta) + \beta[\delta - \lambda\ell(\alpha^*)\varphi'(h^*)]} - \frac{\lambda(1-\mu)\varphi(h^*)(1-\alpha^*)^{1-\gamma}}{\delta - \lambda\ell(\alpha^*)\varphi'(h^*)}$$

Factoring out $\lambda(1-\mu)\varphi(h^*)(1-\alpha^*)^{1-\gamma} > 0$:

$$W'(\alpha^*) = \lambda(1-\mu)\varphi(h^*)(1-\alpha^*)^{1-\gamma}\left[\frac{\beta}{(1-\beta)+\beta\Gamma} - \frac{1}{\Gamma}\right]$$

where $\Gamma \equiv \delta - \lambda\ell(\alpha^*)\varphi'(h^*) > 0$. The term in brackets equals:

$$\frac{\beta\Gamma - (1-\beta) - \beta\Gamma}{\Gamma[(1-\beta)+\beta\Gamma]} = \frac{-(1-\beta)}{\Gamma[(1-\beta)+\beta\Gamma]} < 0$$

since all terms in the denominator are positive.

Therefore $W'(\alpha^*) < 0$ at any stable interior steady state with $\mu < 1$: marginal adoption reduces steady-state output at the equilibrium. The sign does not require any additional assumption beyond those already imposed (Assumptions 1–3).

Remark 5. The result $W'(\alpha^*) < 0$ is a statement about the marginal effect of adoption *at the equilibrium*, not a claim that $W(\alpha)$ is globally decreasing. The proof uses the FOC to eliminate A and g, so the sign depends only on $(1-\beta) > 0$. This suffices for the trap proof: combined with $d\alpha^*/d\beta < 0$ (Corollary 1), it yields $dY^*/d\beta > 0$, which is the comparative static needed.

Remark 6 (Impatience Condition). The following condition, while not required for $W'(\alpha^*) < 0$, ensures well-behaved comparative statics: $\delta - \lambda\ell(\alpha^*)\varphi'(h^*) < (1-\beta)/\beta$. This holds when firms are sufficiently impatient relative to depreciation.

By Corollary 1(ii), $d\alpha^*/d\beta < 0$. Combined with $W'(\alpha^*) < 0$:

$$\frac{dY^*}{d\beta} = W'(\alpha^*) \cdot \frac{d\alpha^*}{d\beta} = (\text{negative}) \times (\text{negative}) > 0$$

Steady-state output is strictly increasing in firm patience.

Step 4: Existence and uniqueness of $\bar{\beta}$. Define $\Psi(\beta) \equiv Y^*(\beta) - \bar{h}$. From Step 3, Ψ is strictly increasing.

Limit as $\beta \to 1$: From Step 3, $Y^*(\beta)$ is strictly increasing in β. Since Y^* is bounded above by $\max_\alpha Y(\bar{h}, \alpha) < \infty$, the limit $Y^*(1^-) = \lim_{\beta\to 1} Y^*(\beta)$ exists. Two cases arise:

Case (i): If $Y^*(1^-) \geq \bar{h}$, then by monotonicity and the intermediate value theorem, there exists unique $\bar{\beta} \in (0,1)$ with $\Psi(\bar{\beta}) = 0$.

Case (ii): If $Y^*(1^-) < \bar{h}$, then $Y^*(\beta) < \bar{h}$ for all $\beta < 1$, so $\bar{\beta} = 1$.

Limit as $\beta \to 0$: Myopic firms maximize current output. As $\beta \to 0$, $\alpha^*(\beta) \to \alpha^{myopic}$ where $\alpha^{myopic} = \arg\max_\alpha Y(h, \alpha)$. Since $Y_\alpha \to -\infty$ as $\alpha \to 1$ (Lemma 6), $\alpha^{myopic} \in (0,1)$. From Step 3, $W(\alpha) \equiv Y^*(\alpha)$ is strictly decreasing in α when $\mu < 1$. Therefore $Y^*(\beta \to 0) = W(\alpha^{myopic}) < W(0) = \bar{h}$. Thus $\Psi(0^+) < 0$.

By continuity and strict monotonicity, the intermediate value theorem yields unique $\bar{\beta} \in (0,1)$ with $\Psi(\bar{\beta}) = 0$.

Step 5: Long-run loss when $\beta < \bar{\beta}$. By Step 4, $\Psi(\beta) < 0$ for $\beta < \bar{\beta}$, i.e., $Y^* < \bar{h} = Y^{NA*}$. Combined with Step 2, there exists unique $T^* > 0$ with $Y_t > Y_t^{NA}$ for $t < T^*$ and $Y_t < Y_t^{NA}$ for $t > T^*$.

Step 6: Individual rationality. Condition (T3) holds by construction: $\alpha_t = \alpha^*(h_t)$ solves the Bellman equation at each t.

Step 7: Necessity. (a) If $\mu \geq 1$: as shown above (Necessity of Substitution), $h^* \geq \bar{h}$. For the trap to fail, we need $Y^* \geq \bar{h}$. We have $Y^* = AG(\alpha^*) + h^*(1-\alpha^*)^{1-\gamma}$. Since $h^* \geq \bar{h}$

and $AG(\alpha^*) > 0$ for $\alpha^* > 0$, a sufficient condition for $Y^* \geq \bar{h}$ is:

$$AG(\alpha^*) \geq \bar{h}[1 - (1 - \alpha^*)^{1-\gamma}]$$

Under Assumption 3, $Ag(0) > \bar{h}(1-\gamma)$. Since $g(\alpha) > 0$ for all α and $[1-(1-\alpha)^{1-\gamma}] \leq (1-\gamma)\alpha$ for α small (by convexity), Assumption 3 implies $AG(\alpha^*) > \bar{h}[1 - (1 - \alpha^*)^{1-\gamma}]$ for α^* in a neighborhood of zero. For larger α^*, condition (ii) ($AG(1) < \bar{h}$) may bind. However, when $\mu \geq 1$, the equilibrium α^* is bounded away from 1 by the static shape of $Y(h, \alpha)$: since $Y_\alpha \to -\infty$ as $\alpha \to 1$ and the dynamic learning term is non-negative when $\mu \geq 1$, the FOC $Y_\alpha + \beta V'(h')(\mu - 1)\lambda\varphi(h) = 0$ requires $Y_\alpha \leq 0$, which bounds α^* strictly below 1. Thus condition (T2) fails when $\mu \geq 1$. (b) If $A \cdot G(1) \geq \bar{h}$: even with $h^* = 0$ and $\alpha^* = 1$, we have $Y^* \geq \bar{h}$. The trap cannot occur. (c) If $\beta \geq \bar{\beta}$: by definition of $\bar{\beta}$, $Y^* \geq \bar{h}$. □

Lemma 12 (Learning Spillover Properties)**.** *If $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is weakly increasing with $\psi(\bar{H}) = 1$, then along any path where $H_t < \bar{H}$, we have $\psi(H_t) < 1$.*

Proof. Since ψ is weakly increasing and $H_t < \bar{H}$, we have $\psi(H_t) \leq \psi(\bar{H}) = 1$. If ψ is strictly increasing on some neighborhood of $\bar{H}$, the inequality is strict. If ψ is constant on $[H_t, \bar{H}]$, then $\psi(H_t) = 1$, but this contradicts the assumption that spillovers affect learning (i.e., $\psi'(H) > 0$ for some H). Under the maintained assumption that learning spillovers are operative, $\psi(H_t) < 1$ when $H_t < \bar{H}$. □

Proposition 2 (Spillover Bias).

Let h_t^U, h_t^{NU}, and h_t^{NA} denote human capital at time t for users, non-users in an AI-adopting economy, and the no-adoption counterfactual, respectively.

With learning spillovers $\psi(H)$, non-users' skill accumulation depends on aggregate human capital: $h_{t+1}^{NU} = (1 - \delta)h_t^{NU} + \lambda\varphi(h_t^{NU}) \cdot \psi(H_t)$. By Lemma 12, $\psi(H_t) < \psi(\bar{H}) = 1$ when $H_t < \bar{H}$, so non-users accumulate skills more slowly than in the no-adoption counterfactual. By induction, $h_t^{NU} < h_t^{NA} = \bar{h}$ for all $t > 0$.

We show h_t^{NU} is strictly decreasing from $\bar{h}$ toward h^{NU*}. Define the non-user transition map $T^{NU}(h; H_t) = (1-\delta)h + \lambda\varphi(h)\psi(H_t)$. At $h = \bar{h}$: $T^{NU}(\bar{h}; H_t) = (1-\delta)\bar{h} + \lambda\varphi(\bar{h})\psi(H_t) < (1 - \delta)\bar{h} + \lambda\varphi(\bar{h}) = \bar{h}$, since $\psi(H_t) < 1$. The non-user steady state h^{NU*} satisfies $\delta h^{NU*} = \lambda\varphi(h^{NU*})\psi(H^*)$, giving $h^{NU*} < \bar{h}$. By Assumption 1, the contraction condition ensures $|(T^{NU})'(h)| = |(1 - \delta) + \lambda\varphi'(h)\psi(H_t)| < 1$ for all h, so T^{NU} is a contraction mapping. Since $T^{NU}(h; H_t) < h$ for all $h \in (h^{NU*}, \bar{h}]$ (the map is below the 45-degree line above the fixed

point) and $h_0^{NU} = \bar{h}$, the sequence $\{h_t^{NU}\}$ is strictly decreasing. Therefore $s_t = \bar{h} - h_t^{NU}$ is strictly increasing.

The cross-sectional counterfactual is:

$$\Delta_t^{CS} = Y(h_t^U, \alpha_t) - Y(h_t^{NU}, 0) = A \cdot G(\alpha_t) + h_t^U(1-\alpha_t)^{1-\gamma} - h_t^{NU}$$

The long-run counterfactual is:

$$\Delta_t^{LR} = Y(h_t^U, \alpha_t) - Y(h_t^{NA}, 0) = A \cdot G(\alpha_t) + h_t^U(1-\alpha_t)^{1-\gamma} - \bar{h}$$

The difference is:

$$\Delta_t^{CS} - \Delta_t^{LR} = \bar{h} - h_t^{NU} > 0$$

since $h_t^{NU} < \bar{h}$ for $t > 0$. The gap is zero at $t = 0$ and strictly increasing in t since $\{h_t^{NU}\}$ is strictly decreasing (shown above). □

Lemma 13 (On-Path Positivity)**.** *Under optimal policy with $\mu < 1$, $\Delta_t^{SC} = Y(h_t^U, \alpha_t) - Y(h_t^U, 0) > 0$ for all $t \geq 0$.*

Proof. Suppose $Y(h, \alpha) < Y(h, 0)$ for some $h > 0$ and $\alpha > 0$. Then adopting α is worse than $\alpha = 0$ for current output. Since $\mu < 1$, $\alpha > 0$ also reduces learning: $L(\alpha, h; \mu) < L(0, h; \mu)$. Thus α is dominated – it yields lower output today *and* lower human capital tomorrow – so it cannot be optimal. Contrapositive: along the optimal path, $\alpha_t > 0$ implies $Y(h_t, \alpha_t) \geq Y(h_t, 0)$, with strict inequality since $\alpha_t \in (0, 1)$ and Y is strictly concave in α. □

Proposition 1 (State-Path Divergence).

Part (i): We establish two claims about Δ_t^{SC}.

Claim 1: Bounded absolute gain, growing relative gain. By Lemma 2 and Lemma 4, $h_t^U \to h^* < \bar{h}$ as $t \to \infty$ when $\mu < 1$. The state-conditional gain is $\Delta_t^{SC} = Y(h_t^U, \alpha_t) - Y(h_t^U, 0) = A \cdot G(\alpha_t) + h_t^U(1-\alpha_t)^{1-\gamma} - h_t^U$. Rewriting:

$$\Delta_t^{SC} = A \cdot G(\alpha_t) - h_t^U \underbrace{[1 - (1-\alpha_t)^{1-\gamma}]}_{>0 \text{ for } \alpha_t > 0}$$

As $h_t^U \to h^*$, the absolute gain $\Delta_t^{SC} \to A \cdot G(\alpha^*) - h^*[1-(1-\alpha^*)^{1-\gamma}]$, which is bounded. The comparative static $\partial \Delta_\infty^{SC} / \partial h^* = -[1-(1-\alpha^*)^{1-\gamma}] < 0$ since $\alpha^* > 0$ and $1-\gamma \in (0,1)$: more

severe skill atrophy (lower h^*) produces larger bias. The *relative* gain Δ_t^{SC}/h_t^U satisfies:

$$\frac{\Delta_t^{SC}}{h_t^U} = \frac{A \cdot G(\alpha_t)}{h_t^U} - [1 - (1 - \alpha_t)^{1-\gamma}]$$

For parameterizations where h^* is small relative to $\bar{h}$ (i.e., when skill atrophy is severe), this ratio can become large. In the limit as $h^* \to 0$ across parameter sequences, the relative gain diverges.

Claim 2: Ratio eventually increases. By Lemma 4(iv), h_t^U is strictly decreasing and α_t is strictly increasing along the optimal path from $h_0 = \bar{h}$. Write the ratio as $R_t \equiv \Delta_t^{SC}/h_t^U = AG(\alpha_t)/h_t^U - [1 - (1 - \alpha_t)^{1-\gamma}]$. Since $\alpha_t \to \alpha^*$ and $h_t \to h^*$, there exists T such that for all $t \geq T$, $|\alpha_t - \alpha^*| < \epsilon$ for any $\epsilon > 0$. For $t \geq T$, the second term $[1 - (1 - \alpha_t)^{1-\gamma}]$ changes by at most $O(\epsilon)$ per period, while the first term $AG(\alpha_t)/h_t^U$ is strictly increasing because $G(\alpha_t)$ is non-decreasing and h_t^U is strictly decreasing. Thus for T large enough, $R_{t+1} - R_t > 0$ for all $t \geq T$. The steady-state ratio $R_\infty = AG(\alpha^*)/h^* - [1 - (1 - \alpha^*)^{1-\gamma}]$ strictly exceeds $R_0 = AG(\alpha_0)/\bar{h} - [1 - (1 - \alpha_0)^{1-\gamma}]$ because $h^* < \bar{h}$ raises the first term and $\alpha^* > \alpha_0$ raises G.

Part (ii): By Lemma 13, $\Delta_t^{SC} > 0$ for all $t \geq 0$ along the optimal path. When the economy is in a skill trap, steady-state output satisfies $Y^* < \bar{h} = Y^{NA}$ (Proposition 3). Thus $\Delta_t^{SC} > 0$ while $Y^* < \bar{h}$: AI appears to raise output in state-conditional comparisons even when it reduces long-run output. □

Corollary 2 (Welfare Reversal Under Patient Evaluation).

Consider the path counterfactual $\Delta^{PATH}(\tilde{\beta}) = \sum_{t=0}^{\infty} \tilde{\beta}^t [Y_t^{user} - Y_t^{NA}]$. For the firm's own discount factor β, revealed preference implies $\Delta^{PATH}(\beta) \geq 0$. When $Y^* < \bar{h}$ and $\tilde{\beta} > \beta$, more weight is placed on long-run outcomes where $Y_t^{user} < Y_t^{NA}$ (for t large). Since the tail of the sum is negative, for $\tilde{\beta}$ sufficiently larger than β, $\Delta^{PATH}(\tilde{\beta}) < 0$. □

Corollary 3 (Sign Reversal).

In the skill trap, $Y^* < \bar{h}$ by Proposition 3, so $\Delta^{LR} = Y^* - \bar{h} < 0$. For $\Delta^{CS} > 0$, we need $Y^* > h^{NU*}$. Learning spillovers ensure $h^{NU*} < \bar{h}$: non-users' steady-state skill satisfies $\delta h^{NU*} = \lambda \varphi(h^{NU*}) \psi(H^*)$ with $\psi(H^*) < 1$, implying $h^{NU*} < \bar{h}$. When $Y^* > h^{NU*}$ (AI users outperform degraded non-users) but $Y^* < \bar{h}$ (AI users underperform the no-adoption benchmark), we have $\Delta^{CS} > 0 > \Delta^{LR}$. □

Corollary (Inequality Dynamics).

Wage variance is $\sigma_t^2 = \mathbb{E}[w_t^2] - (\mathbb{E}[w_t])^2$. With two groups, this simplifies to:

$$\sigma_t^2 = \frac{N_t^{pre}}{N}(w^{pre})^2 + \frac{N_t^{AI}}{N}(w_t^{AI})^2 - \left(\frac{N_t^{pre}}{N}w^{pre} + \frac{N_t^{AI}}{N}w_t^{AI}\right)^2$$

Short run: AI compresses wages by raising w_t^{AI} for low-skill workers. With w^{pre} fixed and w_t^{AI} rising, the gap shrinks and σ_t^2 falls.

Long run: As $h_t^{AI} \to h^* < h^{pre}$, the wage gap $w^{pre} - w_t^{AI}$ widens. Combined with $N_t^{pre} \to 0$, variance eventually rises as the small pre-AI cohort commands large premiums.

The turning point T^* occurs when compression effects are overtaken by scarcity. Faster atrophy (higher $(1-\mu)\alpha^*$) accelerates this transition. □

Proposition 4 (Ability Reversal and Vintage Premium).

Part (i): Consider workers with ability θ_i, so $\varphi_i(h) = \theta_i\varphi(h)$. The skill dynamics are $h_{t+1} = (1-\delta)h_t + \lambda\theta_i\ell(\alpha_t)\varphi(h_t)$. Define the skill gap $\Delta_t(\theta) \equiv h_t^{NA}(\theta) - h_t^U(\theta)$, where h_t^{NA} is the no-adoption path ($\alpha = 0$) and h_t^U is the user path ($\alpha > 0$). Both paths start from $h_0 = \bar{h}$.

At $t = 1$: $h_1^{NA}(\theta) = (1-\delta)\bar{h} + \lambda\theta\varphi(\bar{h})$ and $h_1^U(\theta) = (1-\delta)\bar{h} + \lambda\theta\ell(\alpha_0)\varphi(\bar{h})$. Thus:

$$\Delta_1(\theta) = \lambda\theta[1 - \ell(\alpha_0)]\varphi(\bar{h}) = \lambda\theta(1-\mu)\alpha_0\varphi(\bar{h})$$

Since $(1-\mu) > 0$ when $\mu < 1$, we have $\partial\Delta_1/\partial\theta = \lambda(1-\mu)\alpha_0\varphi(\bar{h}) > 0$.

For the induction step, note that the skill gap evolves as:

$$\Delta_{t+1} = (1-\delta)\Delta_t + \lambda\theta[\varphi(h_t^{NA}) - \ell(\alpha_t)\varphi(h_t^U)]$$

Differentiating with respect to θ:

$$\frac{\partial\Delta_{t+1}}{\partial\theta} = (1-\delta)\frac{\partial\Delta_t}{\partial\theta} + \lambda[\varphi(h_t^{NA}) - \ell(\alpha_t)\varphi(h_t^U)] + \lambda\theta\left[\varphi'(h_t^{NA})\frac{\partial h_t^{NA}}{\partial\theta} - \ell(\alpha_t)\varphi'(h_t^U)\frac{\partial h_t^U}{\partial\theta}\right]$$

The second term is positive since $\varphi(h_t^{NA}) > \ell(\alpha_t)\varphi(h_t^U)$ (the no-adoption path has higher effective learning). For the third term: $\varphi' < 0$ by Assumption 1, $\partial h_t^{NA}/\partial\theta > 0$, and $\partial h_t^U/\partial\theta > 0$. Since $h_t^{NA} > h_t^U$ (the no-adoption path yields higher skill), the sign of the third term depends on the curvature of φ. By the log-concavity of φ (Assumption 1), the "direct scaling effect" (higher θ scales up the learning differential) dominates the "convergence effect"

(higher θ pushes both paths into regions where φ is lower). Formally, log-concavity implies $\varphi'(h)/\varphi(h)$ is decreasing, so for $h^{NA} > h^U$: $|\varphi'(h^{NA})|/\varphi(h^{NA}) \leq |\varphi'(h^U)|/\varphi(h^U)$. This ensures the curvature terms cannot overturn the positive second term. Thus $\partial\Delta_{t+1}/\partial\theta > 0$, completing the induction. The intuition: ability scales learning, so high-ability workers forgo more learning when AI substitutes for practice.

Part (ii): Let $\bar{h}$ denote pre-AI cohort skill (constant, as they trained without AI) and h_t^{post} denote post-AI cohort skill at time t. With $\mu < 1$ and positive adoption, $h_t^{post} \to h^* < \bar{h}$ by Lemma 3. The vintage premium is $\pi_t = \bar{h}/h_t^{post}$. Since h_t^{post} is decreasing toward $h^* < \bar{h}$ (Lemma 4(iv)), π_t is increasing in t until pre-AI cohorts retire. □

Proposition 4 (Hump-Shaped Inequality).

Let N_t^{pre} denote the mass of pre-AI workers at time t, with $N_t^{pre} = N_0^{pre}e^{-\nu t}$ for retirement rate $\nu > 0$, and $N_t^{post} = 1 - N_t^{pre}$ the mass of post-AI workers.

Part (i): At $t = 0$, all workers are in the pre-AI steady state with skill $\bar{h}$, so the wage distribution is degenerate: $\sigma_0^2 = 0$.

Part (ii): For $t > 0$, post-AI workers have skill $h_t < \bar{h}$ (by Lemma 2), while pre-AI workers maintain $\bar{h}$. The variance for a two-group population with masses N_t^{pre} and N_t^{post} and wages $w^{pre} = \bar{h}$ and $w_t^{post} = h_t$ is:

$$\sigma_t^2 = N_t^{pre}(1 - N_t^{pre})(\bar{h} - h_t)^2$$

At $t = 0$, $N_0^{pre} = 1$ and $h_0 = \bar{h}$, so $\sigma_0^2 = 0$. For small $t > 0$, $N_t^{pre} \approx 1 - \nu t$ and $h_t < \bar{h}$, so $\sigma_t^2 > 0$ and increasing.

Part (iii): As $t \to \infty$, $N_t^{pre} \to 0$, so $\sigma_t^2 \to 0$ regardless of the wage gap. The variance is maximized at some finite T^{max} where the effects of the widening wage gap and shrinking pre-AI cohort exactly offset. Differentiating:

$$\frac{d\sigma_t^2}{dt} = (1 - 2N_t^{pre})(-\nu N_t^{pre})(\bar{h} - h_t)^2 + N_t^{pre}(1 - N_t^{pre}) \cdot 2(\bar{h} - h_t) \cdot \left(-\frac{dh_t}{dt}\right)$$

The first term is negative when $N_t^{pre} < 1/2$ (retirement effect); the second is positive when $dh_t/dt < 0$ (skill gap widening). The peak occurs when these balance. Under baseline parameters with $\nu = 0.05$, the peak is around $t \approx 25$. □

Proposition 5 (Human Capital Externality).

The social planner maximizes $\sum_t \beta^t[Y(H_t, \alpha_t; A) + \theta H_t^\eta]$ subject to $H_{t+1} = (1 - \delta)H_t + \lambda L(\alpha_t, H_t; \mu) \cdot \psi(H_t)$, where $\psi(H)$ captures learning spillovers.

The FOC with respect to α includes the term $\beta \frac{\partial W}{\partial H'} \cdot \frac{\partial L}{\partial \alpha} \cdot \psi(H) = \beta \frac{\partial W}{\partial H'} \lambda(1-\mu)\varphi(H)\psi(H)$ from human capital dynamics. The social value of human capital $\frac{\partial W}{\partial H'}$ includes the spillover term $\theta\eta(H')^{\eta-1}$ from the output spillover and additional terms from the learning spillover $\psi'(H)$, which are absent from the private value $V'(h')$.

When $\theta > 0$ or $\psi'(H) > 0$, social valuation of human capital exceeds private valuation, so the social marginal cost of adoption exceeds the private marginal cost. The social optimum therefore involves lower adoption: $\alpha^S < \alpha^D$.

When $\theta = 0$ and $\psi(H) \equiv 1$, social and private valuations coincide, the FOCs are identical, and the decentralized equilibrium is efficient. □

Proposition 6 (Training Data Externality).

Part (i): With endogenous AI quality, $A_{t+1} = (1-\zeta)A_t + \zeta Q(H_t, \bar{\alpha}_t)$ with $\partial Q/\partial \bar{\alpha} < 0$. Each atomistic firm i chooses α_i taking $\bar{\alpha}$ as given. The private FOC is:

$$\frac{\partial Y}{\partial \alpha_i} = \beta V'(h')\lambda(1-\mu)\varphi(h)$$

which ignores the effect of α_i on $\bar{\alpha}$ (since firm i is measure zero) and hence on future AI quality. The social planner internalizes that aggregate adoption affects AI quality, adding the term $\beta(\partial W/\partial A') \cdot \zeta(\partial Q/\partial \bar{\alpha}) < 0$ to the FOC. This additional cost implies $\alpha^S < \alpha^D$.

Part (ii): Define $\Delta W^{HC} \equiv W(\bar{H}, A_0) - W(H^*, A_0)$ as the welfare loss from human capital externalities alone (holding A fixed at A_0), and $\Delta W^{data} \equiv W(\bar{H}, A_0) - W(\bar{H}, A^{data})$ as the loss from training data externalities alone (holding H fixed at $\bar{H}$). The total loss is $\Delta W^{total} \equiv W(\bar{H}, A_0) - W(H^{**}, A^{**})$, where (H^{**}, A^{**}) is the joint equilibrium. Since $H^{**} < \bar{H}$ worsens data quality ($\partial Q/\partial H > 0$) and $A^{**} < A_0$ affects adoption incentives, we have $\Delta W^{total} > \Delta W^{HC} + \Delta W^{data}$: the externalities reinforce each other in general equilibrium. □

Proposition 7 (Training Mandates).

Without policy, the decentralized equilibrium features adoption $\alpha^D > \alpha^S$ (by Proposition 5). A mandate ρ constrains $\alpha \leq 1 - \rho$.

If $\rho < 1 - \alpha^D$, the mandate is not binding and has no effect. If $\rho > 1 - \alpha^S$, the mandate forces $\alpha < \alpha^S$, which is below the social optimum – welfare falls.

For $\rho \in [1 - \alpha^D, 1 - \alpha^S]$, the mandate binds and reduces adoption toward the social optimum. Welfare rises as ρ increases (adoption falls) until $\alpha = \alpha^S$.

The optimal mandate $\rho^* = 1 - \alpha^S$ exactly implements the social optimum: firms choose

$\alpha = 1 - \rho^* = \alpha^S$ since the constraint binds.

Productivity effect: Current output is $Y(H, \alpha) = A \cdot G(\alpha) + H(1-\alpha)^{1-\gamma}$. At $\alpha^D > \alpha^S$, unregulated output exceeds mandated output in the short run (since $Y_\alpha > 0$ locally when firms are adopting). But welfare includes the present value of human capital:

$$W = \sum_t \beta^t [Y_t + \theta H_t^\eta]$$

The mandate sacrifices current Y to raise future H, improving W when externalities are present. □

Proposition 9 (Selection Effects).

Part (i): The FOC for firm i's adoption choice is:

$$A \cdot g(\alpha_i) - h_i(1-\gamma)(1-\alpha_i)^{-\gamma} = \beta_i V'(h_i')\lambda(1-\mu)\varphi(h_i)$$

With β_i heterogeneous, patient firms (high β_i) have higher RHS, implying lower α_i^*. Selection on patience: impatient firms adopt more, gaining short-run competitive advantage but losing long-run human capital.

Part (ii): Let $s_{i,t}$ be firm i's market share. With $s_{i,t} \propto Y_{i,t}$, firms with high α_i have high $s_{i,t}$ in the short run. Survivor bias: cross-sectional samples overweight high-α firms because they have larger market shares, overstating measured AI benefits.

Part (iii): Define output-weighted aggregate human capital as $H_t = \int_0^1 h_{i,t}\, s_{i,t}\, di$, where $s_{i,t}$ denotes firm i's market share with $\int_0^1 s_{i,t}\, di = 1$. Decompose using the identity $H_t = \bar{h}_t + \text{Cov}(h_{i,t}, s_{i,t})$, where $\bar{h}_t \equiv \int_0^1 h_{i,t}\, di$ is unweighted mean skill:

$$H_t = \bar{h}_t + \int_0^1 (h_{i,t} - \bar{h}_t)(s_{i,t} - 1)\, di.$$

Under selection, market share satisfies $s_{i,t} \propto Y_{i,t} = A \cdot G(\alpha_i) + h_{i,t}(1-\alpha_i)^{1-\gamma}$. From Part (i), impatient firms choose higher α_i, so α_i and β_i are negatively correlated. From Lemma 2, higher α_i implies lower $h_{i,t}$ when $\mu < 1$. Meanwhile, higher α_i raises short-run output $Y_{i,t}$ (the static gain from AI dominates the skill loss initially), so $s_{i,t}$ is increasing in α_i. Combining: $h_{i,t}$ and $s_{i,t}$ are negatively correlated, i.e.,

$$\text{Cov}(h_{i,t}, s_{i,t}) = \int_0^1 (h_{i,t} - \bar{h}_t)(s_{i,t} - 1)\, di < 0.$$

It follows that $H_t^{\text{selection}} = \bar{h}_t + \text{Cov}(h_{i,t}, s_{i,t}) < \bar{h}_t = H_t^{\text{no-selection}}$: market selection amplifies aggregate skill atrophy beyond the unweighted level because the firms commanding the largest market shares are precisely those whose workers have experienced the most skill degradation.

Scope and timing: Parts (ii) and (iii) are *transitional* results that hold when firms start from common initial skills and the static gain from AI dominates skill loss. In the very long run, Corollary 1 establishes that $Y^*(\beta)$ is increasing in β, so patient firms eventually have higher steady-state output. The model does not characterize the steady-state distribution of market shares across β-types – only that during the economically relevant early phase, selection favors impatient, high-adoption firms. □

Proposition 8 (Feedback Loop Stability).

Part (i): By Corollary 1(i), $\partial\alpha^*/\partial A > 0$ and $\partial h^*/\partial A < 0$. With endogenous A, skill atrophy causes AI quality to fall: $A^{**} = Q(H^{**}, \alpha^{**}) < Q(\bar{H}, 0) = A_0$ since $\partial Q/\partial H > 0$, $\partial Q/\partial \alpha < 0$, and $H^{**} < \bar{H}$ with $\alpha^{**} > 0$. Lower AI quality reduces adoption and raises steady-state skills: $H^{**} = h^*(A^{**}) > h^*(A_0) = H^*(A_0)$ since $\partial h^*/\partial A < 0$. The feedback loop partially protects human capital.

Part (ii): For uniqueness, note that the joint steady-state conditions define a continuous map. The H steady-state locus is downward-sloping in (H, A) space (higher A induces more adoption, which lowers steady-state H), while the A steady-state locus $A = Q(H, \alpha^*(H, A))$ is upward-sloping (higher H improves training data quality). The single crossing implies a unique intersection. For local stability, let (H^*, A^*) be a steady state. Consider perturbation $(H^* + \epsilon, A^* + \delta)$. The dynamics are:

$$H_{t+1} - H^* \approx J_{11}(H_t - H^*) + J_{12}(A_t - A^*)$$
$$A_{t+1} - A^* \approx J_{21}(H_t - H^*) + J_{22}(A_t - A^*)$$

where the Jacobian entries are:

$$J_{11} = (1-\delta) + \lambda\ell(\alpha^*)\varphi'(H^*)\psi(H^*) + \lambda\ell(\alpha^*)\varphi(H^*)\psi'(H^*) < 1$$
$$J_{12} = \lambda\ell'(\alpha^*)\frac{\partial\alpha^*}{\partial A}\varphi(H^*)\psi(H^*) < 0 \quad (\text{since } \ell' < 0 \text{ when } \mu < 1,\ \partial\alpha^*/\partial A > 0)$$
$$J_{21} = \zeta\left(\frac{\partial Q}{\partial H} + \frac{\partial Q}{\partial \alpha}\frac{\partial \alpha^*}{\partial H}\right) > 0 \quad (\text{since } \partial Q/\partial H > 0, \text{ and } \partial Q/\partial\alpha < 0,\ \partial\alpha^*/\partial H < 0 \text{ when } \mu < 1)$$
$$J_{22} = (1-\zeta) + \zeta\frac{\partial Q}{\partial \alpha}\frac{\partial \alpha^*}{\partial A} < 1 \quad (\text{since } \partial Q/\partial\alpha < 0,\ \partial\alpha^*/\partial A > 0)$$

The characteristic polynomial is $\lambda^2 - (J_{11} + J_{22})\lambda + (J_{11}J_{22} - J_{12}J_{21}) = 0$. For stability, we verify the Schur conditions: (i) $|\det J| < 1$; (ii) $|\mathrm{tr} J| < 1 + \det J$. Under Assumption 1 (local contraction), $J_{11} < 1$. Since $J_{12} < 0$ and $J_{21} > 0$, the off-diagonal product $-J_{12}J_{21} > 0$ raises the determinant; this relaxes the trace inequality (ii) but tightens requirement (i). Stability ultimately follows from ζ being small: when ζ is small, $J_{22} \approx 1 - \zeta$, $J_{21} \approx 0$, and the eigenvalues are approximately J_{11} and $1 - \zeta$, both with modulus less than 1. The slow AI adjustment rate ensures the feedback loop does not destabilize the system.

Part (iii): With ζ small (slow AI adjustment), $J_{21} \approx \zeta \cdot \partial Q/\partial H$ and $J_{22} \approx 1 - \zeta$. In the limit $\zeta \to 0$, the eigenvalues are $\lambda_1 = J_{11}$ (the H-only eigenvalue, stable by Lemma 9) and $\lambda_2 = 1$. For small $\zeta > 0$, $\lambda_2 = 1 - \zeta + O(\zeta^2) < 1$, so the system is locally stable. Slow AI adjustment ensures the A dynamics do not destabilize the system. □