Skill Atrophy and AI Productivity Measurement

Tommaso Bondi* Gentry Johnson†

February 18, 2026

Abstract

How should we measure the productivity effects of AI? Standard estimates condition on current skill, but when AI substitutes for the cognitive effort that builds skill, skill itself becomes endogenous to past AI use. We formalize *pedagogical quality*, the fraction of learning-by-doing that survives AI delegation, and show it governs two measurement biases: state-path divergence inflates within-worker panel estimates; spillover bias degrades control groups. The biases generate a "scissors" pattern: panel and RCT estimates diverge over time. The same mechanism produces an ability reversal in long-run skill losses and a vintage wage premium for pre-AI cohorts.

*Cornell Tech and SC Johnson School of Business, 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, Dirk Bergemann, 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

AI makes workers more productive in the short run: experimental estimates range from 14% to 55% at the task level (Brynjolfsson et al., 2025a; Dell'Acqua et al., 2023; Noy and Zhang, 2023; Peng et al., 2023). But if AI substitutes for the practice that builds skill, it may also make workers less skilled.

Why might AI erode skill where earlier automation largely did not? Previous waves primarily displaced *routine* tasks with little learning content. AI targets *cognitively demanding* tasks: drafting arguments, diagnosing patterns, writing code. The key distinction is whether the automated task also generates human capital. Automating filing does not degrade legal reasoning; automating brief writing might, because constructing arguments is itself the practice that builds the capacity to evaluate them.

Evidence consistent with skill atrophy is accumulating. Bastani et al. (2025) find that unrestricted GPT-4 access reduces subsequent math performance. Shen and Tamkin (2026) find a nearly identical reduction among software developers learning a new programming library. Budzyń et al. (2025) document endoscopist deskilling after three months of AI-assisted colonoscopy. University professors have repeatedly reported perfect scores in homework followed by poor performances in final exams, suggesting AI tools increase immediate productivity at the expense of learning.[1] And METR (2026) report that a longitudinal developer productivity study had to be redesigned because 30–50% of participants refused to work without AI – consistent with atrophy making the unaided condition increasingly costly.

Existing productivity experiments identify the contemporaneous effect holding skill fixed; by design, they cannot identify the effect relative to the skill path that would have prevailed absent adoption.

Put simply, AI may look increasingly transformative partly because we compare it to workers whose skills have atrophied; it may look modest relative to the worker who would have existed absent adoption. Firm-level data are consistent with this concern: Yotzov et al. (2026) survey nearly 6,000 executives and find over 80% report no measurable effect of AI on productivity, despite sizable task-level gains in experiments.

We model this through *pedagogical quality* (μ), the fraction of learning-by-doing that survives AI delegation. When $\mu < 1$, AI substitutes for skill formation, and two measurement biases arise. *State-path divergence* is an individual-level wedge: within-worker panels that condition on current skill overstate AI's value because current skill is itself shaped by past AI use. *Spillover bias* is a cross-sectional wedge: when skill is socially produced through

[1] See for example https://x.com/lxeagle17/status/1899979401460371758.

mentorship and shared training, AI adoption degrades the control group.

The first bias requires only that a single worker uses AI over time; the second requires that adoption by some workers affects others' learning environments. Both are governed by μ; spillover bias also requires socially produced learning. The biases grow with adoption duration; they coexist whenever $\mu < 1$ and spillovers are present, both empirically plausible conditions.

To see how these biases interact, consider two ways of measuring AI's productivity effect over time. A within-worker panel that controls for experience asks: how much does AI help *this worker right now*, given her current skill? It conditions on the worker's skill state at the time of measurement. A long-horizon RCT that follows treated and control workers over years asks: how much better off is an AI user than a worker who never had it? It compares across workers whose skills have followed different trajectories. When skill atrophy operates, these designs answer increasingly different questions. The panel holds skill fixed and sees AI becoming more valuable – because the worker's unaided capacity has eroded. The RCT lets skill evolve and sees AI becoming less valuable – because the treated worker has lost ground relative to the control. Together they produce what we call a *scissors pattern*: panel estimates of AI's productivity gain rise while RCT estimates of the same quantity fall.

The same mechanism generates labor market implications: high-ability workers bear the largest long-run costs despite benefiting least in the short run, and pre-AI cohorts command growing *vintage wage premiums* over those trained with AI assistance.

A back-of-the-envelope calculation anchored to recent experimental estimates by Bastani et al. (2025) and Shen and Tamkin (2026) makes magnitudes concrete. Both studies find that AI access reduces subsequent unassisted performance by ∼17%; interpreting this as a flow-learning effect implies $\mu \approx 0.83$. At this value, the panel estimate of AI's productivity gain grows roughly 5% over 20 years while the RCT estimate *falls* by roughly 7%.

This scissors pattern is the model's sharpest testable prediction, and it does not arise under habit formation, learning curves, or selection on early adopters. Under habit formation, the underlying skill state does not persistently deteriorate with use, so the treatment-control skill gap does not widen over time and the RCT estimand does not decline persistently. Under learning curves, both estimands rise. Under selection, both fall.

1.1 Related Literature

This paper contributes to three literatures; for recent surveys, see Acemoglu (2024) and Agrawal et al. (2026). 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; Jones and Tonetti (2026) find that weak links – tasks still performed by slowly-improving labor – constrain the aggregate effects of even dramatic automation. Agrawal et al. (2026) emphasize complementarities between AI and human judgment. Our paper explores a complementary margin: task frameworks treat skills as a stock determining productivity, and our analysis asks what happens when tasks are also inputs into skill production. AI may complement the *use* of judgment while substituting for its *development*.

A growing empirical literature documents short-run productivity effects: Noy and Zhang (2023) for writing, Peng et al. (2023) for coding, Dell'Acqua et al. (2023) for consulting.

The "jagged frontier" identified by Dell'Acqua et al., in which AI substantially helps on some tasks but hurts on others, reflects heterogeneous static AI capability across tasks (heterogeneous $g(j)$ in our framework). A distinct and less studied source of heterogeneity is that the learning content of delegated tasks may also vary: μ may differ across task types, with the sharpest skill atrophy arising when AI encroaches on the most learning-rich tasks. Otis et al. (2023) find heterogeneous effects among Kenyan entrepreneurs: AI mentorship helped high performers but *hurt* low performers. Gaessler and Piezunka (2023) find chess computers accelerated skill development ($\mu > 1$), plausibly because chess provides immediate, objective feedback; but the majority of recent work documents deskilling, including among endoscopists (Budzyń et al., 2025), robot-assisted workers (Beane, 2019), and knowledge workers (Lee et al., 2025; Dell'Acqua, 2022).

Our model builds on human capital theory (Becker, 1962) and learning-by-doing (Arrow, 1962). Arrow's foundational insight, that production generates knowledge as a byproduct, motivates our central question: what happens when a technology severs this link? Models of technology adoption and human capital – including Cooley et al. (1997) and Violante (2002) – treat skill investment as a *separate* decision from adoption: workers divert effort into learning or face vintage-specific depreciation, but the adoption and investment margins are distinct. In our framework, the adoption decision *is* the skill investment decision: delegating a task to AI simultaneously raises output and reduces the learning content of work.

Concurrent work by Acemoglu et al. (2026) studies how agentic AI shapes collective knowledge in a Bayesian social learning framework; our approach is complementary, focusing on individual human capital dynamics and mismeasurement.

Recent work examines how AI threatens training and skill transmission: Garicano and Rayo (2025) show apprenticeships become unviable when AI automates entry-level work, and Beane (2019) find robotic surgery reduced trainee practice tenfold. Our paper focuses on a

related but distinct margin: the endogenous learning content of ongoing production under AI delegation, rather than the structural allocation of workers to training roles or entry-level positions.

2 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.

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). AI quality A is exogenous; we abstract from endogenous improvement to isolate the human capital channel.

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 $(\alpha, 1]$. Optimal uniform effort allocation yields total worker output $h(1 - \alpha)^{1-\gamma}$: delegating tasks to AI concentrates effort on remaining tasks, partially offsetting the lost output and generating real short-run productivity gains even when skill atrophy operates in the background.

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 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.

The output structure reveals the difference-versus-sum tension central to the paper.

The *difference* $Y(h,\alpha) - Y(h,0)$ is decreasing in h: AI is less valuable when workers are already skilled. When skill atrophy lowers h, the difference rises while the sum $Y(h,\alpha)$ falls. Experiments measure the difference; aggregate productivity tracks the sum.

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 $\delta \in (0, 1/2)$ is depreciation, $\lambda > 0$ governs learning intensity, and $L(\alpha, h; \mu)$ is the learning function. 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*. Equation (3) is a deliberately parsimonious reduced form: delegation changes the share of work that remains skill-forming, while $\varphi(h)$ captures the empirically standard feature that learning rates decline with expertise.

The effective learning rate $\ell(\alpha) \equiv 1 - (1-\mu)\alpha$ is strictly positive for all $\alpha \in [0,1)$ when $\mu \geq 0$, and for all $\alpha \in [0,1]$ when $\mu > 0$; since optimal adoption is always interior ($\alpha^* < 1$ by Lemma 4 in the Supplemental Appendix), $\ell(\alpha^*) > 0$ along the equilibrium path. The maintained case is $\mu < 1$: AI substitutes for learning, producing the skill atrophy of the title. When $\mu = 1$, delegation has no effect on learning and all measurement biases vanish, a benchmark we use throughout. The case $\mu > 1$, in which AI *augments* learning, reverses all signs and is discussed in Section 5.

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

The slope bound is a contraction condition ensuring $(1-\delta) + \lambda\ell(\alpha)\varphi'(h) > 0$. Let $\bar{h}$ denote the no-adoption steady state, defined by $\delta\bar{h} = \lambda\varphi(\bar{h})$.

These properties capture diminishing returns: experts learn less from practice because most relevant knowledge has already been acquired. Log-convexity ensures $|\varphi'(h)|/\varphi(h)$ is non-increasing in h, used in the ability reversal proof (Proposition 6). Standard forms including $\varphi(h) = c/(1+h)^k$ for $k \geq 1$ satisfy all conditions.

The critical property: $\partial L/\partial\alpha = (\mu - 1)\varphi(h)$, which is negative when $\mu < 1$ and zero when $\mu = 1$. This derivative governs whether delegation helps or hurts skill 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 mitigate this, confirming that μ is a design parameter, not a fixed property of AI. The scalar should be interpreted as an adoption-weighted average of task-specific pedagogical content; the model's predictions are sharpest when delegated tasks are learning-rich. The learning specification depends on the *measure* of delegated tasks, not on effort intensity or throughput; empirically, Lee et al. (2025) find workers do not reallocate freed effort to complex tasks. The production function assumes AI output is independent of worker skill h, ruling out human-in-the-loop complementarities. Adding skill-complementarity would strengthen the measurement biases, since declining h would reduce even AI-aided output; our baseline is thus conservative.

2.3 The Firm's Dynamic Problem

Firms maximize the present discounted value of output; we treat output as the welfare-relevant object throughout.[2] The discount factor is $\beta \in (0, 1)$.

Assumption 2 (Labor Market Structure)**.** Labor markets are competitive with general human capital. Wages equal marginal products.

Under Assumption 2, the firm does not fully internalize returns to general human capital that accrue to the worker upon separation. The firm's Bellman problem therefore generates an *equilibrium* adoption path, not necessarily a welfare-optimal one; a planner or long-lived worker who internalizes the full human capital trajectory would adopt less. The measurement results in Section 3 are invariant to this distinction: the biases depend on the equilibrium skill path, not on whether that path is welfare-optimal. The "overadoption" result below is relative to the firm's own dynamic problem, not to a social planner's. 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

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

[2]Adding effort disutility would strengthen the case for delegation ex ante but does not remove the skill-endogeneity wedge, since the measurement biases depend on the relationship between skill paths and estimands, not on the level of worker utility.

Standard results ensure V exists, is unique, and is strictly increasing and concave in h. Higher adoption today raises current output but, when $\mu < 1$, reduces future human capital.

2.4 Equilibrium Characterization

2.4.1 The Role of Pedagogical Quality

Adoption raises contemporaneous output through $G(\alpha)$, but it changes tomorrow's state by altering the rate at which today's work translates into skill. Pedagogical quality μ determines which force dominates at the margin. When AI is sufficiently productive, some adoption is always optimal; complete delegation is never optimal because effort concentration on remaining tasks becomes increasingly valuable.

Assumption 3 (AI Productivity)**.** $A \cdot g(0) > \bar{h}(1-\gamma) + \beta V'(\bar{h})\lambda(1-\mu)\varphi(\bar{h})$.

The static output gain from delegating the first task must exceed the marginal effort-reallocation loss plus the discounted learning cost, evaluated at the no-adoption steady state. This is a joint condition on A, g, $\bar{h}$, and β. The static-only condition $A \cdot g(0) > \bar{h}(1-\gamma)$ is necessary but not sufficient when $\mu < 1$, because the dynamic learning cost $\beta V'(h')\lambda(1-\mu)\varphi(h)$ must also be overcome. The full condition ensures interior adoption at $h_0 = \bar{h}$. Along the equilibrium transition, as h falls the static gain $Ag(0) - h(1-\gamma)$ rises (lower skill reduces the opportunity cost of delegation), which offsets the increasing dynamic cost from higher $\varphi(h)$, preserving interiority along the path.

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 $\eta \geq 0$.

The function ψ captures reduced mentorship and peer learning when aggregate skill declines; $\eta = 0$ nests the no-spillover baseline. The spillover channel becomes operative in Section 3, where it generates a second measurement bias distinct from state-path divergence.

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

(i) When $\mu < 1$*:* $\partial\alpha^*/\partial\mu > 0$ *locally around stable steady states.*

(ii) When $\mu = 1$*:* $\partial Y/\partial\alpha = 0$ *determines adoption.*

Part (i) says adoption generates a dynamic skill cost when $\mu < 1$: higher μ reduces the learning penalty, so firms adopt more. Part (ii) says the problem is purely static when $\mu = 1$. The first-order condition makes this precise:

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

The left side is the static marginal benefit; the right side is the shadow cost of forgone learning. When $\mu = 1$, the right side vanishes. When $\mu < 1$, the dynamic cost pushes adoption below the myopic optimum.

2.4.2 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. For any adoption level α, there exists a unique steady-state skill level $h^*(\alpha)$ on the stable branch of the dynamics; uniqueness follows from φ being strictly decreasing, and the contraction condition in Assumption 1 ensures stability. Since $\ell(\alpha^*) < 1$ for any $\alpha^* > 0$ when $\mu < 1$, steady-state human capital satisfies $h^* < \bar{h}$: skill atrophy is an immediate consequence. Steady-state skill h^* is continuous in μ, so small deviations from unity produce proportionally small biases, but compounding over career horizons (Proposition 3) means even modest deviations matter.

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. The numerator is negative when $\mu < 1$, so $dh^*/d\alpha < 0$: each unit of delegation reduces steady-state skill. The magnitude is governed by the ratio of the learning penalty to the restoring force: when learning is steep ($\varphi(h^*)$ large, as for novices) or the pedagogical penalty is severe, skill is highly sensitive to

adoption. The elasticity of h^* with respect to ℓ is

$$\frac{\partial \log h^*}{\partial \log \ell} = \frac{\delta}{\delta - \lambda \ell(\alpha^*)\varphi'(h^*)} \in (0,1), \tag{9}$$

since $\varphi' < 0$ makes the denominator exceed δ. Diminishing returns dampen the impact of learning shocks, but the absolute magnitude $|dh^*/d\alpha|$ is economically large at plausible parameters.

Equilibrium is unique and globally stable:

Proposition 2 (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 generically unique.

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

Under a regularity condition on the policy slope (condition (S3) in the Supplemental Appendix), the transition path $\{h_t\}$ is monotonically decreasing from $h_0 = \bar{h}$; otherwise convergence still obtains but need not be monotone. The condition requires that the static cross-partial of output dominates the dynamic feedback through value-function curvature, a property satisfied by standard functional forms at plausible parameters.[3]

Comparative statics are standard: $\partial\alpha^*/\partial A > 0$, $\partial h^*/\partial A < 0$, $\partial\alpha^*/\partial\beta < 0$, $\partial h^*/\partial\beta > 0$. Short-termism exacerbates skill atrophy.

Myopic firms always overadopt: the gap $\alpha^M(h) - \alpha^*(h)$ is proportional to $(1-\mu)\varphi(h)$. The same primitive – high learning potential combined with low pedagogical quality – drives both the policy distortion and the mismeasurement that conceals it.

Marginal adoption always reduces steady-state output. Defining $W(\alpha) \equiv A \cdot G(\alpha) + h^*(\alpha)(1-\alpha)^{1-\gamma}$ and using the envelope theorem:

$$W'(\alpha^*) = -\frac{(1-\beta)\,\lambda(1-\mu)\varphi(h^*)(1-\alpha^*)^{1-\gamma}}{\Gamma\big[(1-\beta)+\beta\Gamma\big]} < 0 \tag{10}$$

where $\Gamma \equiv \delta - \lambda\ell(\alpha^*)\varphi'(h^*) > 0$ at a stable steady state. The sign requires only $(1-\beta) > 0$: marginal adoption reduces steady-state output when $\mu < 1$. Adoption is nonetheless optimal because the present value of short-run output gains exceeds the discounted long-run skill losses; the steady-state comparison isolates only the latter. The measurement biases formalized in

[3]Proofs of global stability and monotonicity appear in the Supplemental Appendix.

the next section are thus not merely statistical artifacts; the welfare object they overstate is itself negative at the margin.

3 Mismeasurement of AI Productivity

We now show that standard empirical designs systematically diverge from the welfare-relevant counterfactual when $\mu < 1$. All wedges scale with $(1 - \mu)$; estimating μ is therefore the first-order empirical task. State-path divergence is the more fundamental bias, requiring only individual-level dynamics; spillover bias compounds it in cross-sectional settings.

3.1 State-Path Divergence

If AI changes the law of motion for skill, conditioning on current skill compares outcomes across different skill histories. State-conditional effects can therefore be positive even when adoption lowers the long-run level of skill.

Definition 1 (State-Conditional vs. Path Counterfactuals)**.** The *state-conditional counterfactual* holds human capital fixed:

$$\Delta_t^{SC} = Y(h_t^U, \alpha_t) - Y(h_t^U, 0). \tag{11}$$

The *path counterfactual* compares output at t under adoption to what it would have been absent any adoption:

$$\Delta_t^{PATH} = Y(h_t^U, \alpha_t) - Y(h_t^{NA}, 0). \tag{12}$$

We define both objects at a single period t, matching the horizon over which empirical studies typically estimate treatment effects. The per-period formulation also makes the source of the wedge transparent: Δ_t^{SC} and Δ_t^{PATH} differ only in whether the comparison uses the realized skill h_t^U or the no-adoption counterfactual h_t^{NA}.

Remark 1 (Cumulative Equivalence). All results hold under a cumulative welfare formulation $\bar{\Delta}_T^{SC} \equiv \sum_{t=0}^{T} \beta^t \Delta_t^{SC}$. The wedge $\bar{\Delta}_T^{SC} - \bar{\Delta}_T^{PATH}$ grows in T; sign properties and scissors divergence are inherited term-by-term (see the Supplemental Appendix).

The state-conditional gain Δ_t^{SC} overstates AI's contribution because it conditions on current skill h_t^U rather than the counterfactual h_t^{NA}. Many empirical implementations recover Δ_t^{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 path counterfactual Δ_t^{PATH}, which compares to the skill trajectory that would have obtained absent adoption.

Proposition 3 (State-Path Divergence)**.** *Suppose $\mu < 1$. Then:*

(i) In steady state, $\Delta_\infty^{SC} = A \cdot G(\alpha^*) - h^*[1 - (1-\alpha^*)^{1-\gamma}]$, with $\partial \Delta_\infty^{SC} / \partial h^* < 0$.

(ii) $\Delta_t^{SC} > 0$ for all $t \geq 0$ along the optimal path.

Part (i) says the state-conditional gain is larger when skill atrophy is more severe: lower h^* inflates AI's measured value because the outside option has deteriorated. Part (ii) says AI always appears beneficial in state-conditional comparisons, even when Δ_t^{PATH} is negative. As h_t^U falls, Δ_t^{SC} rises while Δ_t^{PATH} falls; experiments and aggregate data measure different objects.

This result requires only $\mu < 1$; no spillovers or cross-agent interaction. The mechanism is analogous to the "bad controls" problem (Angrist and Pischke, 2009): controlling for skill when skill is shaped by the treatment biases the estimate upward. The static bad-controls insight identifies a point-in-time endogeneity; our contribution is showing that this bias *compounds dynamically* along the transition path. The overstatement is not merely statistical: since $W'(\alpha^*) < 0$ (equation (10)), the welfare object that standard designs inflate is itself negative at the margin.

Remark 2 (Dependency Spiral). The proof of Proposition 3 also establishes that the *relative* gain Δ_t^{SC}/h_t^U is strictly increasing in t along the optimal path. AI accounts for a growing share of the worker's output – not because AI improves, but because the worker's unaided capacity shrinks. A manager observing this trajectory would conclude AI is becoming more essential; the conclusion is correct but, crucially, the cause is atrophy, not technological progress.

State-path divergence operates entirely within a single worker's history. When learning is socially produced, a second bias arises.

3.2 Spillover Bias

Cross-sectional comparisons introduce a second wedge when skill is socially produced. As adoption becomes widespread, non-adopters experience degraded learning environments

through reduced mentorship, fewer high-skill peers, and thinner task exposure, so the "control group" no longer proxies for the no-adoption counterfactual. The channel is familiar from the education literature on peer effects (Sacerdote, 2001; Epple and Romano, 2011): here the analogue operates through AI adoption, as the stock of available mentors and on-the-job training opportunities shrinks for everyone (see for instance Burtch et al. (2024) on the decline of Stack Overflow).

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. Note that $\Delta_t^{LR} = \Delta_t^{PATH}$ from Definition 1; the relabeling reflects the shift from individual-level counterfactuals (Section 3.1) to cross-sectional comparisons, where Δ_t^{LR} serves as the benchmark against which spillover bias is measured.

The cross-sectional counterfactual is the comparison implicit in many empirical designs, including RCTs that randomize AI access over short horizons. These designs are internally valid over their study periods; the divergence from the long-run counterfactual emerges over longer horizons when aggregate AI adoption affects learning opportunities for non-users. When spillovers are absent ($\eta = 0$), the non-user's skill remains at $\bar{h}$ and $\Delta_t^{CS} = \Delta_t^{LR}$; the two counterfactuals coincide.

Under Assumption 4, a non-user accumulates skill according to

$$h_{t+1}^{NU} = (1-\delta)h_t^{NU} + \lambda\varphi(h_t^{NU}) \cdot \psi(H_t) \tag{13}$$

while the no-adoption counterfactual satisfies $h_{t+1}^{NA} = (1-\delta)h_t^{NA} + \lambda\varphi(h_t^{NA})$ (since $\psi(\bar{H}) = 1$). The *spillover skill gap* $s_t \equiv h_t^{NA} - h_t^{NU}$ satisfies $s_0 = s_1 = 0$ and, for $t \geq 2$:

$$s_{t+1} = (1-\delta)s_t + \lambda\Big[\varphi(h_t^{NA}) - \varphi(h_t^{NU})\psi(H_t)\Big] \tag{14}$$

The second term is positive at $t = 1$: since $h_1^{NA} = h_1^{NU} = \bar{h}$ (spillovers have not yet operated), the term reduces to $\lambda\varphi(\bar{h})[1 - \psi(H_1)] > 0$. For $t \geq 2$, positivity requires a dominance argument (provided in the Supplemental Appendix) because φ is decreasing and $h_t^{NU} < h_t^{NA}$, which partially offsets the $\psi(H_t) < 1$ factor.

Proposition 4 (Spillover Bias)**.** *Suppose $\mu < 1$ and Assumption 4 holds with $\eta > 0$. Then:*

(i) $\Delta_t^{CS} > \Delta_t^{LR}$ *for all* $t \geq 2$.

(ii) The spillover skill gap $s_t > 0$ *for all* $t \geq 2$ *and converges to a strictly positive limit* $s^* = \bar{h} - h^{NU*} > 0$.

Cross-sectional estimates overstate the long-run value of AI as soon as spillovers have time to operate, and the degradation is persistent, converging to a permanent wedge. When $\eta = 0$, cross-sectional estimates correctly measure long-run effects, but state-path divergence (Proposition 3) remains.

The two biases differ in remedies and empirical status. State-path divergence calls for counterfactual-aware designs and is pinned down by μ, for which multiple experimental estimates exist. Spillover bias calls for designs robust to interference and depends on η, which lacks direct estimation in AI settings; the evidence is suggestive (Burtch et al., 2024; Beane, 2019) but not definitive. Even if $\eta = 0$, state-path divergence alone delivers the scissors prediction. When $\eta > 0$, spillovers constitute a genuine externality that would shift a planner's optimal adoption below the equilibrium level.

Remark 3 (Bias Decomposition). The total bias in cross-sectional estimates admits a useful decomposition. Adding and subtracting both $Y(h_t^U, 0)$ and $Y(h_t^{NA}, 0)$:

$$\Delta_t^{CS} = \underbrace{Y(h_t^U, \alpha_t) - Y(h_t^U, 0)}_{\Delta_t^{SC} \text{ (state-conditional gain)}} + \underbrace{Y(h_t^U, 0) - Y(h_t^{NA}, 0)}_{\text{state-gap bias } (=h_t^U - h_t^{NA} < 0)} + \underbrace{Y(h_t^{NA}, 0) - Y(h_t^{NU}, 0)}_{\text{spillover bias } (=h_t^{NA} - h_t^{NU} = s_t > 0)} \quad (15)$$

The state-conditional gain Δ_t^{SC} is what panel studies typically estimate. The state-gap component $h_t^U - h_t^{NA} < 0$ *reduces* the cross-sectional estimate relative to the state-conditional one; the spillover component $s_t > 0$ *inflates* it by depressing the non-user comparison group. The two components push the cross-sectional estimate in opposite directions relative to Δ_t^{SC}: the state-gap term makes Δ_t^{CS} smaller than Δ_t^{SC}, while the spillover term makes it larger. The net cross-sectional bias $\Delta_t^{CS} - \Delta_t^{LR} = s_t > 0$ is unambiguously positive.

3.3 Implications for Empirical Research

Our analysis identifies a precise estimand mismatch across research designs. To fix notation, the long-horizon RCT recovers

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

where $h_t(d)$ is the skill path under treatment $d \in \{0, 1\}$. For small t, $h_t(1) \approx h_t(0)$ and the estimand recovers the direct productivity effect. For large t, skill paths diverge when $\mu < 1$. A within-worker panel that controls for experience conditions on current skill, recovering

$$\tau^{panel}(t) = \mathbb{E}[Y_t(1, h_t) - Y_t(0, h_t) \mid h_t] = \Delta_t^{SC} \tag{17}$$

which *overstates* the welfare-relevant object by the state-gap bias (Remark 3). The identification statement "conditioning on experience conditions on current skill" is conceptual: it requires that the analyst's controls (tenure, experience, task history) proxy for h_t well enough that the residual variation in AI use is orthogonal to skill. Standard implementations such as worker fixed effects or event studies difference out baseline heterogeneity but do not in general deliver "on vs. off at fixed h_t" without repeated withdrawal periods. Similarly, most AI experiments randomize *access* with discretionary use rather than mandating fixed intensity: the ITT then recovers a compliance-weighted average of τ^{RCT}, and Proposition 5(iii) applies to the complier subpopulation whose use persists despite skill atrophy.

Cross-sectional user/non-user comparisons recover Δ_t^{CS}, which additionally includes spillover bias. Each design answers a different question, and the answers diverge over time.

The choice of research design 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. The emerging experimental literature provides building blocks: Bastani et al. (2025) measure skill directly, METR (2025, 2026) track usage intensity, and Budzyń et al. (2025) observe unassisted performance after AI exposure. Supplemental Table S2 summarizes bias exposure of common designs. Novice samples maximize exposure while expert samples minimize it: the population where AI appears most transformative is precisely where the bias is largest.

The sharpest testable implication concerns the time dynamics of these estimands. When $\mu < 1$, RCTs and panels move in opposite directions – the scissors pattern.

Proposition 5 (Divergent Estimand Dynamics ("Scissors"))**.** *Suppose* $\mu < 1$ *and adoption*

α^* is interior. Along the equilibrium path:

(i) $\tau^{RCT}(t+1) < \tau^{RCT}(t)$ *for all* t *when* $\eta = 0$*, and when* $\eta < (1-\mu)\alpha^*/(1-\alpha^*)^{1-\gamma}$*.*

(ii) $\Delta^{SC}_{t+1} > \Delta^{SC}_t$ *for all* t*.*

(iii) When μ *is sufficiently below* 1 *or* β *sufficiently small,* $\exists\, \hat{t} > 0$ *such that* $\tau^{RCT}(\hat{t}) < 0$*.*

The RCT estimand is strictly declining (i), the panel estimand strictly increasing (ii), and the RCT estimand eventually turns negative (iii). Part (i) is stated for fixed treatment intensity α^*, matching the mandated-use design of a standard RCT; it extends to the endogenous path $\alpha_t = \alpha^*(h_t)$ under the maintained regularity conditions. Part (ii) uses the endogenous path explicitly: both falling h_t and rising α_t contribute.

The directional results – that τ^{RCT} eventually declines and Δ^{SC}_t eventually rises – follow from $\mu < 1$ alone at any fixed adoption intensity. The strict monotonicity *at every* t along the endogenous path additionally requires the regularity conditions ensuring monotone h_t and non-decreasing α_t (condition (S3) in the Supplemental Appendix), plus the spillover bound on η in Part (i).

Part (i): user skills fall faster than non-user skills, so the RCT estimate shrinks each period. Part (ii): as h^U_t declines, the opportunity cost of AI falls, so AI appears increasingly valuable conditional on current skill.

Part (iii) follows because $h^U_t \to h^* < \bar{h}$ while $h^{NU}_t \to h^{NU*} \leq \bar{h}$, so the skill gap eventually dominates the static AI gain. This sign reversal is distinctive: under most theories of technology adoption, AI's measured contribution is bounded below by zero, since the user can always stop using it. Skill atrophy changes this by altering the user's outside option.

A growing gap between panel and RCT estimands is itself diagnostic of skill atrophy: the same technology would appear increasingly valuable in within-worker comparisons and simultaneously decreasingly valuable in long-horizon treatment-control comparisons. Testing the prediction requires settings where both can be tracked over multi-year horizons – a design no existing study provides but that ongoing longitudinal experiments (METR, 2025, 2026) may eventually enable.

This divergence does not arise under habit formation (temporary dips that recover), learning curves (both estimands rise), or selection on early adopters (both estimands fall). Existing data hint at the mechanism: in the original METR (2025) study, experienced developers were 19% *slower* with AI yet believed they were faster. While the perception gap could reflect anchoring or optimism bias, it is also suggestive of state-conditional reasoning:

developers may have evaluated their productivity relative to their current skill level rather than the counterfactual skill they would have maintained without AI.

Magnitudes. The Supplemental Appendix reports a numerical illustration. At $\mu = 0.5$, the measurement bias exceeds 8% by year 10 and 13% by year 20; the scissors crossing – the horizon at which the RCT estimand turns negative – occurs around year 15. At $\mu = 0.83$ (following Bastani et al. (2025); Shen and Tamkin (2026)), the crossing occurs beyond year 30. The mapping from a delayed post-test deficit to μ in our law of motion depends on exposure duration, δ, and λ; the Supplemental Appendix provides a formal identification argument and sensitivity analysis across plausible parameter ranges.

4 Labor Market Implications

The measurement biases identified above have direct labor market counterparts. This section shows that the mechanism generating mismeasurement also reshapes the distribution of wages across workers and across cohorts. Skill atrophy operates through learning, and learning potential varies: workers with high ability forgo the most human capital per unit of delegation, while workers trained before AI adoption retain skills that post-AI cohorts cannot replicate.

Short-run experiments consistently find that AI disproportionately benefits less-skilled workers (Brynjolfsson et al., 2025a; Noy and Zhang, 2023; Peng et al., 2023), a finding Autor (2024) interprets as skill democratization. But when $\mu < 1$, the short-run compression in measured productivity masks a longer-run divergence in skill trajectories. Workers with the highest learning potential accumulate the largest skill deficits, precisely because they forgo the most learning per unit of delegation. The short-run productivity gains persist at each point in time; the long-run costs are borne disproportionately by those with the most to lose. A policy that evaluates AI based on short-run experiments would conclude it narrows inequality; the same policy evaluated over career horizons would reveal the opposite.

4.1 Ability Reversal and Vintage Premium

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

(i) $\partial(h_t^{NA} - h_t^U)/\partial\theta > 0$ *for all* $t \geq 1$.

(ii) $h_t^{post} \to h^* < \bar{h}$. *The vintage premium* $\pi_t = \bar{h}/h_t^{post} - 1$ *is increasing in* t.

Part (i) says the skill loss from AI adoption is increasing in ability: high-ability workers bear the largest long-run costs, precisely the workers who benefit least from AI in short-run studies. We call this *ability reversal*.

Part (ii) implies pre-AI cohorts become increasingly valuable as repositories of expertise that post-AI training environments cannot replicate. The vintage premium persists until retirement eliminates the pre-AI cohort. Beane (2019) documents that robotic surgery reduced trainee experience tenfold, with senior surgeons becoming increasingly valuable. Crucially, the scarcity cannot be resolved by producing more graduates: at adoption level α^*, the best a new entrant can achieve is $h^* < \bar{h}$, regardless of training duration, because the steady-state learning flow is bounded above by δh^*.

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 1 (Hump-Shaped Inequality)**.** *Suppose* $\mu < 1$ *and pre-AI cohorts retire at rate* $\nu > 0$. *Then* $\sigma_0^2 = 0$, $\sigma_t^2 > 0$ *for* $t > 0$, *and there exists* T^{max} *such that* σ_t^2 *is increasing for* $t < T^{max}$ *and decreasing for* $t > T^{max}$.

Wage variance follows a hump-shaped path: it rises as post-AI workers' skills diverge from pre-AI workers', peaks when retirement begins to dominate, then falls as pre-AI cohorts exit.

Convergence to uniformly lower skills is punctuated by scarcity premiums for pre-AI veterans. At plausible parameters, $T^{max} \approx 30$ years, long enough that the reversal may be attributed to unrelated structural changes. The hump shape distinguishes our mechanism from skill-biased technological change, which predicts monotonic inequality increases (Acemoglu and Autor, 2011).

The cohort effects interact with the measurement biases: within any cohort, 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.

5 Conclusion

When the treatment changes the law of motion for the state variable on which its effects are measured, standard causal objects diverge from welfare-relevant ones. We have characterized

this divergence: two biases, state-path divergence and spillover bias, both governed by pedagogical quality μ, systematically inflate measured AI gains when $\mu < 1$, and the inflation grows with adoption duration. A back-of-the-envelope calculation, anchored to independent experimental estimates converging on $\mu \approx 0.83$, suggests the wedge is economically meaningful within a decade.

The framework yields concrete empirical priorities. Delayed post-tests measuring unassisted performance after AI exposure would identify μ from a single study. Re-randomization designs that periodically reassign treatment status would distinguish skill atrophy from habit formation: the former predicts permanent deficits proportional to exposure duration, the latter temporary dips that recover. The vintage premium is testable using wage data linked to AI adoption timing. Industries experiencing rapid AI improvement should exhibit accelerating skill divergence, distinguishing our model from ones where AI quality and human capital evolve independently.

A pronounced asymmetry adds urgency. As skill falls, workers delegate more, accelerating further loss; recovery requires reducing AI use and rebuilding skill through deliberate practice. The dynamic is self-reinforcing in a way standard automation is not: a displaced factory worker retains her skills, but a knowledge worker who delegates cognitive effort may not. Competitive pressure compounds the problem, since tools maximizing short-run output will outcompete those preserving learning when users cannot observe long-run consequences. Preventing skill degradation is far more effective than reversing it.

Several limitations deserve acknowledgment. The parameter μ is treated as exogenous; in practice, pedagogical quality may vary with expertise ($\mu(h)$), and the monotonic scissors dynamics require that $\alpha^*(h)$ remain decreasing in h under state-dependent μ, which may need further restrictions on $\mu'(h)$. The existence and sign of the measurement biases require only that $\mu(h) < 1$ over the relevant range of skill formation. Heterogeneous adoption across firms would reinforce the vintage premium through between-firm skill dispersion. The model does not incorporate endogenous AI improvement, which would compound skill erosion as rising A induces higher α^* and creates a feedback loop between technology and human capital. When $\mu > 1$ – as Gaessler and Piezunka (2023) find for chess engines, plausibly because chess provides immediate, unambiguous feedback – all signs reverse and standard estimates *understate* long-run gains. Whether μ is above or below one varies across tasks, AI designs, and career stages. While the majority of empirical studies so far point to $\mu < 1$, identifying which regime obtains for a given setting is the first-order empirical question our framework motivates.

Appendix

Technical lemmas, detailed derivations, and the complete proof of Proposition 6(i) appear in the Supplemental Appendix.

Proof of Proposition 1. The Bellman FOC is $Y_\alpha + \beta V'(h') \cdot h'_\alpha = 0$, where $h'_\alpha = \lambda(\mu - 1)\varphi(h)$. Rearranging:

$$Ag(\alpha) - h(1-\gamma)(1-\alpha)^{-\gamma} = \beta V'(h')\lambda(1-\mu)\varphi(h).$$

Part (i): Define $F(\alpha, \mu) \equiv Y_\alpha + \beta V'(h')\lambda(\mu - 1)\varphi(h) = 0$. Then $F_\mu = \beta V'(h')\lambda\varphi(h) > 0$ and $F_\alpha < 0$ by strict concavity of the Bellman objective (Supplemental Lemma 7), so $d\alpha^*/d\mu = -F_\mu/F_\alpha > 0$. Part (ii): When $\mu = 1$, the RHS vanishes and the FOC reduces to $Y_\alpha = 0$. □

Proof of Proposition 2. Part (i): Define $T(h) = (1-\delta)h + \lambda\ell(\alpha^*(h))\varphi(h)$. At $h = \bar{h}$: $T(\bar{h}) < \bar{h}$ since $\ell(\alpha^*) < 1$ and $\delta\bar{h} = \lambda\varphi(\bar{h})$. At h near 0: $T(h) > h$ since $\lambda\mu\varphi(0) > 0$. By Brouwer, a fixed point $h^* \in (0, \bar{h})$ exists. Interiority of α^* follows from Assumption 3 (Supplemental Lemma 4). Part (ii): The steady-state system has Jacobian with $\det(\mathbf{J}) \neq 0$ generically (Supplemental Lemma 7). Tangency of the stationarity and FOC loci is codimension-1. Part (iii): T maps the compact set $[\underline{h}, \bar{h}]$ into itself. Under the regularity condition (S3), T is monotone increasing with $T'(h^*) < 1$, giving convergence by the monotone convergence theorem (Supplemental Lemma 6). □

Proof of Proposition 3. Part (i): $\Delta_t^{SC} = AG(\alpha_t) - h_t^U[1 - (1-\alpha_t)^{1-\gamma}]$. At steady state, $\partial\Delta_\infty^{SC}/\partial h^* = -[1 - (1-\alpha^*)^{1-\gamma}] < 0$: lower h^* reduces the opportunity cost of delegation, inflating the measured gain. Part (ii): Suppose $Y(h, \alpha) < Y(h, 0)$ for some on-path $\alpha > 0$. Then α yields lower current output and, since $\mu < 1$, lower learning and thus lower h'; since V is strictly increasing in h, the continuation value is also lower. Thus α is strictly dominated by $\alpha = 0$, contradicting optimality. Hence $\Delta_t^{SC} > 0$ at every t. □

Proof of Proposition 4. Under Assumption 4 with $\eta > 0$, non-user skill evolves as $h_{t+1}^{NU} = (1-\delta)h_t^{NU} + \lambda\varphi(h_t^{NU})\psi(H_t)$. At $t = 0$: $h_1^{NU} = \bar{h}$ since $\psi(H_0) = \psi(\bar{H}) = 1$. At $t = 1$: user skills have declined, so $H_1 < \bar{H}$, $\psi(H_1) < 1$, and $h_2^{NU} < \bar{h}$. By induction, $h_t^{NU} < \bar{h}$ for all $t \geq 2$. Part (i): Since Y is linear in h and $Y(h, 0) = h$, $\Delta_t^{CS} - \Delta_t^{LR} = h_t^{NA} - h_t^{NU} = s_t > 0$ for $t \geq 2$. Part (ii): The non-user transition $T^{NU}(h; H_t)$ is a contraction (by the slope bound in Assumption 1) with fixed point $h^{NU*} < \bar{h}$. The orbit $\{h_t^{NU}\}$ converges monotonically to h^{NU*}, so $s_t \to \bar{h} - h^{NU*} > 0$. The dominance argument: the forcing term $\lambda[\varphi(h_t^{NA}) - \varphi(h_t^{NU})\psi(H_t)]$ remains bounded away from zero since $\psi(H_t) < 1$, ensuring the gap does not collapse. □

Proof of Proposition 5. Part (i): User skills decline faster than non-user skills; at $t = 0$ both start at $\bar{h}$, so $\tau^{RCT}(0) > 0$ (Proposition 3(ii)). The first difference $\tau^{RCT}(1) - \tau^{RCT}(0) = -\lambda\varphi(\bar{h})(1-\mu)\alpha^*(1-\alpha^*)^{1-\gamma} < 0$. For $t \geq 1$, the adoption penalty $(1-\mu)\alpha^*\lambda\varphi(h_t^U)$ dominates the non-user decline under $\eta < (1-\mu)\alpha^*/(1-\alpha^*)^{1-\gamma}$. Part (ii): $\Delta_t^{SC} = AG(\alpha_t) - h_t^U\Omega_t$ where $\Omega_t = 1-(1-\alpha_t)^{1-\gamma} \in (0,1)$. Define $f(h) = AG(\alpha^*(h)) - h\,\Omega(\alpha^*(h))$. Using the FOC (6) and $d\alpha^*/dh < 0$, $f'(h) < 0$ (Supplemental Appendix). Since h_t is decreasing, $\Delta_t^{SC} = f(h_t)$ is increasing. Part (iii): $h_t^U \to h^* < \bar{h}$ and $h_t^{NU} \to h^{NU*} \leq \bar{h}$. The limiting RCT estimand is $AG(\alpha^*) + h^*(1-\alpha^*)^{1-\gamma} - h^{NU*} = W(\alpha^*) - h^{NU*}$, which is negative when steady-state output under adoption falls sufficiently below h^{NU*}. Since $W'(\alpha^*) < 0$ (equation (10)), this holds when μ is sufficiently below 1 or β sufficiently small. Since $\tau^{RCT}(0) > 0$ and the sequence is monotone decreasing (Part (i)), a unique crossing $\hat{t}$ exists. □

Proof of Proposition 6. Part (i): With $\varphi_i(h) = \theta_i\varphi(h)$, the skill gap $\Delta_t(\theta) = h_t^{NA}(\theta) - h_t^U(\theta)$ satisfies $\Delta_1(\theta) = \lambda\theta(1-\mu)\alpha_0\varphi(\bar{h})$, so $\partial\Delta_1/\partial\theta > 0$. For $t \geq 2$, log-convexity of φ ensures the no-adoption path suffers a smaller proportional learning reduction than the user path, and $\ell < 1$ further depresses the user term. The full induction appears in the Supplemental Appendix. Part (ii): With $\mu < 1$ and $\alpha^* > 0$, $h_t^{post} \to h^* < \bar{h}$ by Proposition 2. Since h_t^{post} is decreasing, $\pi_t = \bar{h}/h_t^{post} - 1$ is increasing. □

Proof of Corollary 1. Between-cohort variance is $\sigma_t^2 = N_t^{pre}(1-N_t^{pre})(\bar{h}-h_t)^2$ where $N_t^{pre} = N_0^{pre}e^{-\nu t}$. At $t = 0$: $\sigma_0^2 = 0$. For small $t > 0$: $h_t < \bar{h}$ and $N_t^{pre} \approx 1$, so $\sigma_t^2 > 0$ and rising. As $t \to \infty$: $N_t^{pre} \to 0$, so $\sigma_t^2 \to 0$ regardless of the skill gap. Continuity gives a peak T^{max}. □

References

Acemoglu, D. (2024). The Simple Macroeconomics of AI. *Economic Policy* 40(121), 13–58.

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., D. Kong, and A. Ozdaglar (2026). AI, Human Cognition and Knowledge Collapse. *NBER Working Paper* 34910.

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.

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.

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.

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.

Cooley, T. F., J. Greenwood, and M. Yorukoglu (1997). The Replacement Problem. *Journal of Monetary Economics* 40(3), 457–499.

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

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.

Epple, D. and R. E. Romano (2011). Peer Effects in Education: A Survey of the Theory and Evidence. *Handbook of Social Economics* 1, 1053–1163.

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.

Jones, C. I. and C. Tonetti (2026). Past Automation and Future A.I.: How Weak Links Tame the Growth Explosion. Working Paper, Stanford GSB.

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.

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

METR (2026). We are Changing our Developer Productivity Experiment Design. METR Blog, February 24.

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

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.

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

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

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

Violante, G. L. (2002). Technological Acceleration, Skill Transferability, and the Rise in Residual Inequality. *The Quarterly Journal of Economics* 117(1), 297–338.

Yotzov, I., J. M. Barrero, N. Bloom, P. Bunn, S. J. Davis, K. Foster, A. Jalca, B. H. Meyer, P. Mizen, M. A. Navarrete, P. Smietanka, G. Thwaites, and B. Z. Wang (2026). Firm Data on AI. *NBER Working Paper* 34836.

Supplemental Appendix (for online publication)

This appendix provides proofs, figures, tables, and additional material for "Skill Atrophy and AI Productivity Measurement."

A Exhibits

Table S1: Notation Guide

Symbol	Definition
h, H	Individual / aggregate human capital
α	AI adoption intensity
A	AI productivity level
μ	Pedagogical quality (< 1: substitutes for learning; ≥ 1: augments)
γ	Returns to effort (effort concentration exponent)
δ, λ	Depreciation rate / learning intensity
β	Discount factor
η	Spillover elasticity
$\psi(H)$	Learning spillover function
$\bar{h}$	No-adoption steady-state skill: $\delta\bar{h} = \lambda\varphi(\bar{h})$
h^*	Steady-state skill under adoption
Δ_t^{CS}, Δ_t^{LR}	Cross-sectional / long-run productivity gain at t
Δ_t^{SC}, Δ_t^{PATH}	State-conditional / path counterfactual at t
$\bar{\Delta}_T^{SC}$, $\bar{\Delta}_T^{PATH}$	Cumulative (discounted sum) counterparts

Note: See main text Section II for the symmetric equilibrium restriction.

Identifying μ from a Delayed-Withdrawal Design

Consider an experiment in which AI is used at intensity α for T periods, then withdrawn. At the withdrawal date T, the user's skill is h_T^U; the no-adoption counterfactual is h_T^{NA}. The post-test deficit measured k periods after withdrawal is $D(T,k) \equiv h_{T+k}^{NA} - h_{T+k}^{W}$, where h_{T+k}^{W} is the skill path after withdrawal (with $\alpha = 0$ for $t > T$).

For $k = 0$ (immediate measurement at withdrawal), the deficit reduces to $D(T,0) = h_T^{NA} - h_T^U$. At $T = 1$ (one period of AI use, starting from $h_0 = \bar{h}$):

$$D(1,0) = h_1^{NA} - h_1^U = \lambda\varphi(\bar{h})[1 - \ell(\alpha)] = \lambda(1-\mu)\alpha\varphi(\bar{h})$$

Table S2: Empirical Designs and Bias Exposure

Design / Setting	**Spillover**	**State-Path**	**Notes**
Any design, novice sample	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
Any design, expert/routine sample	Low	Low	Skill formation not at stake

Note: "Spillover" refers to spillover bias (Proposition 4); "State-Path" to state-path divergence (Proposition 3). "High/Moderate/Low" indicates relative exposure.

Table S3: 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	1.6	5.0	8.6	12.4
Bias at year 20 (%)	0.0	2.5	7.8	13.5	19.8
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 $(\Delta_t^{SC} - \Delta_t^{LR})/\Delta_0^{LR} \times 100$, the overstatement relative to the initial long-run gain. Vintage premium is $\bar{h}/h_t^{post} - 1$. AI quality held constant. Other parameters: $\delta = 0.05$, $\lambda = 0.15$, $\alpha = 0.5$, $\eta = 0.15$.

Expressed as a fraction of the no-adoption skill level $h_1^{NA} = (1-\delta)\bar{h} + \lambda\varphi(\bar{h})$:

$$\frac{D(1,0)}{h_1^{NA}} = \frac{\lambda(1-\mu)\alpha\varphi(\bar{h})}{(1-\delta)\bar{h} + \lambda\varphi(\bar{h})}$$

Using the no-adoption stationarity condition $\delta\bar{h} = \lambda\varphi(\bar{h})$:

$$\frac{D(1,0)}{h_1^{NA}} = \frac{(1-\mu)\alpha\delta\bar{h}}{\bar{h}} = (1-\mu)\alpha\delta$$

Thus, for a one-period exposure design with immediate post-test, μ is identified by $\mu =$

$1 - D(1,0)/(h_1^{NA}\alpha\delta)$. With the experimental estimates of $D/h \approx 0.17$ from Bastani et al. (2025) and Shen and Tamkin (2026), and taking $\alpha \approx 0.5$ and $\delta \approx 0.05$ (annual depreciation), this gives $\mu \approx 1 - 0.17/(0.5 \times 0.05) = 1 - 6.8$, which is far below zero and economically implausible – indicating that the experimental exposure corresponds to much shorter effective periods than one year, or equivalently that the deficit reflects a higher-frequency learning flow.

Reinterpreting: if the experimental exposure corresponds to n sub-periods within our annual model period, and the measured deficit reflects the cumulative flow over those sub-periods, then the mapping becomes $\mu \approx 1 - D/(n\alpha\delta_n)$ where $\delta_n = \delta/n$ is the per-sub-period depreciation. For studies with exposure on the order of weeks ($n \approx 50$ sub-periods per year), the calibration yields $\mu \approx 0.83$, consistent with the estimates reported in the main text.

More generally, for T-period exposure with k-period delay after withdrawal, the deficit satisfies the recursion

$$D(T, k+1) = (1-\delta)D(T,k) + \lambda[\varphi(h_{T+k}^{NA}) - \varphi(h_{T+k}^{W})]$$

Since $h_{T+k}^{W} < h_{T+k}^{NA}$ and φ is decreasing, the second term is negative: the deficit shrinks after withdrawal as the deprived worker catches up (diminishing returns to learning favor those with lower skill). The rate of recovery depends on λ and the curvature of φ. A design that measures $D(T,k)$ at multiple values of k after withdrawal can separately identify μ (from the initial deficit at $k=0$) and the recovery rate (from the trajectory), which disciplines λ and φ jointly.

Sensitivity to (δ, λ). Holding the measured deficit $D/h = 0.17$ fixed, the implied bias magnitudes vary with the depreciation and learning parameters. At $\delta = 0.03$ (slow depreciation) with λ adjusted to maintain $\bar{h}$, the implied μ falls (more of each period's learning is at stake, so a 17% deficit implies less substitution per task). At $\delta = 0.08$ (fast depreciation), μ rises. Across the range $\delta \in [0.03, 0.08]$ and $\lambda \in [0.10, 0.25]$, the implied μ spans approximately $[0.75, 0.90]$, and the corresponding 20-year panel bias ranges from 2% to 8%. The qualitative conclusions – positive bias, scissors divergence – are invariant to this range.

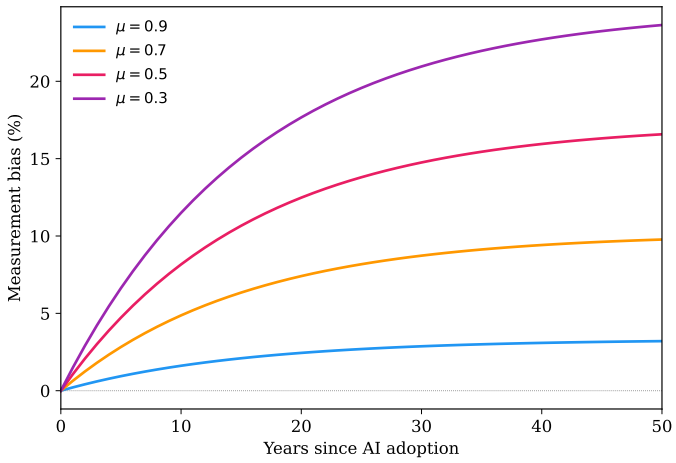

Figure S1: Measurement Bias Over Time

Note: Bias defined as $(\Delta_t^{SC} - \Delta_t^{LR})/\Delta_0^{LR} \times 100$. Parameters as in Table S3.

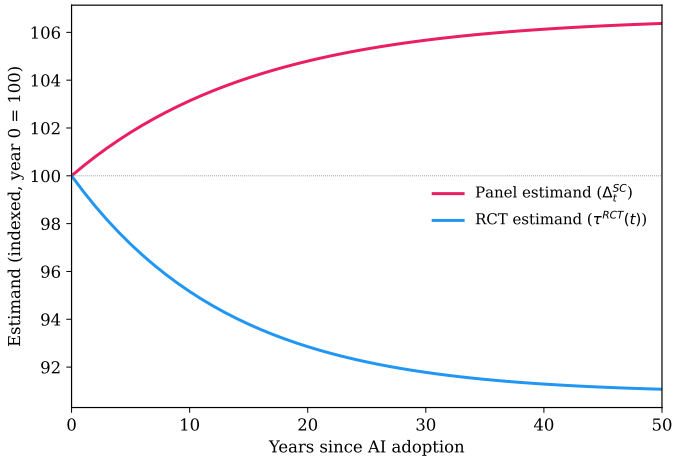

Figure S2: Divergent Estimand Dynamics

Note: Both estimands indexed to 100 at year 0. Parameters as in Table S3 with $\mu = 0.5$.

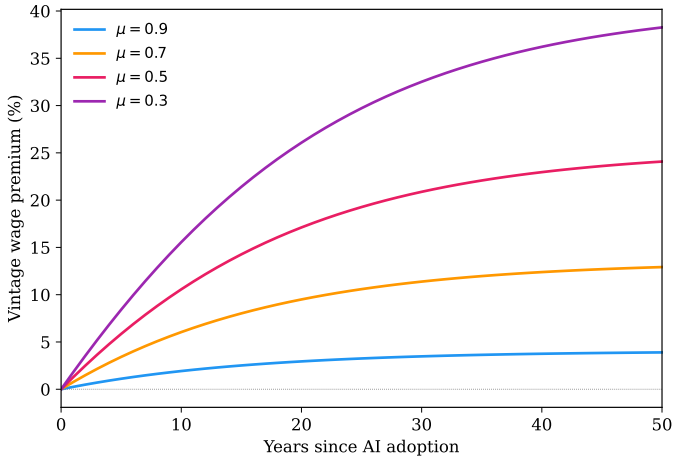

Figure S3: Vintage Wage Premium Over Time

Note: Premium defined as $\bar{h}/h_t^{post} - 1$, in percent. Parameters as in Table S3.

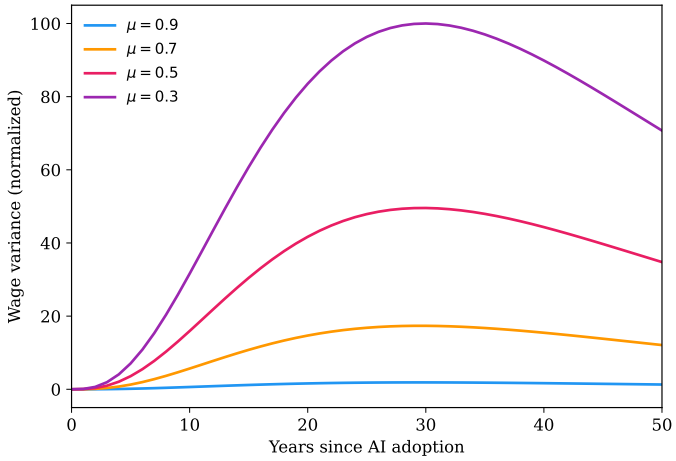

Figure S4: Hump-Shaped Inequality Dynamics

Note: Wage variance across pre-AI and post-AI cohorts. Pre-AI cohorts retire at rate $\nu = 0.04$. Parameters as in Table S3.

B Proofs

This appendix provides formal proofs for all results in the main text. Section B.1 states and proves technical lemmas; Section B.2 proves the main results. Proofs follow the order of the main text.

B.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 1 (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 2 (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 3 (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 4 (Optimal Adoption is Interior)**.** *Under Assumption 3, $\alpha^*(h) \in (0, 1)$ for all h along the equilibrium path from $h_0 = \bar{h}$.*

Proof. At $\alpha \to 1$: $\partial Y/\partial\alpha \to -\infty$ (Lemma 2), so $\alpha^* < 1$. At $\alpha = 0$: the total marginal value of adoption is $[Ag(0) - h(1-\gamma)] + \beta V'(h')\lambda(\mu - 1)\varphi(h)$, where the first bracket is

the static output gain and the second (negative when $\mu < 1$) is the dynamic learning cost. Assumption 3 directly requires this sum to be strictly positive for all $h \in (0, \bar{h}]$, accounting for the dynamic cost at its largest (evaluated along the no-adoption path). Thus $\alpha^*(h) > 0$. □

Lemma 5 (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 $(1-\delta) < 1$. Denote $m \equiv \delta - \lambda\ell(\alpha^*)|\varphi'(h^*)| > 0$ (the "stability margin"), so the first two terms equal $1 - m$. The third term has magnitude $|\lambda(1-\mu)(d\alpha^*/dh)\varphi(h^*)|$, which is finite since $|d\alpha^*/dh| < \infty$ (guaranteed by $D_\alpha \neq 0$) and $\varphi(h^*)$ is bounded. Therefore $|T'(h^*)| < 1$ when this magnitude is less than m. For steady states where this fails, stability is instead established by the global convergence argument of Lemma 6, which shows $T(h) < h$ on $(h^*, \bar{h}]$ independently of T'. □

Lemma 6 (Convergence to Steady State)**.** *Under optimal policy with $\mu < 1$, if $h_0 \in (0, \bar{h}]$, then h_t converges to a fixed point of T in $(0, \bar{h})$. Moreover, the orbit from $h_0 = \bar{h}$ converges to a fixed point h^* satisfying $T'(h^*) < 1$.*

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: Compact invariant set. At $h = \bar{h}$: $T(\bar{h}) = (1-\delta)\bar{h} + \lambda\ell(\alpha^*(\bar{h}))\varphi(\bar{h}) < (1-\delta)\bar{h} + \lambda\varphi(\bar{h}) = \bar{h}$ since $\ell(\alpha) < 1$ when $\alpha > 0$ and $\mu < 1$ (Lemma 4), and $\delta\bar{h} = \lambda\varphi(\bar{h})$ defines $\bar{h}$. For any $h \in (0, \bar{h})$: $T(h) \leq (1-\delta)h + \lambda\varphi(h) < \bar{h}$ (by the mean value theorem applied to $\bar{T}(h) = (1-\delta)h + \lambda\varphi(h)$, which has unique fixed point $\bar{h}$ and slope in $(0,1)$ by Assumption 1). Define $\underline{T}(h) = (1-\delta)h + \lambda\mu\varphi(h)$, the full-adoption transition ($\alpha = 1$). Since $\ell(\alpha) \geq \mu$ for all $\alpha \in [0,1]$: $T(h) \geq \underline{T}(h)$. By Assumption 1, $\underline{T}$ is a contraction with unique fixed point $\underline{h} > 0$. Thus T maps $[\underline{h}, \bar{h}]$ into itself: a compact invariant set.

Step 2: Existence of a fixed point. By the Brouwer fixed point theorem, T has at least one fixed point $h^* \in [\underline{h}, \bar{h}]$. Since $T(\bar{h}) < \bar{h}$ and $T(\underline{h}) \geq \underline{T}(\underline{h}) = \underline{h}$ (with strict inequality since $\alpha^*(\underline{h}) < 1$ by Lemma 4), we have $h^* \in (\underline{h}, \bar{h})$.

Step 3: Convergence to a fixed point. We show T is monotone increasing on $[\underline{h}, \bar{h}]$, which combined with the compact invariant set implies convergence from any initial condition.

Differentiating: $T'(h) = (1-\delta)+\lambda\ell'(\alpha^*)\frac{d\alpha^*}{dh}\varphi(h)+\lambda\ell(\alpha^*)\varphi'(h)$. Denote the sum of the first and third terms by $\varrho(h,\alpha) \equiv (1-\delta)+\lambda\ell(\alpha)\varphi'(h)$, which is strictly positive by Assumption 1. For the policy feedback term: $\ell'(\alpha) = -(1-\mu)$, so the term is $-\lambda(1-\mu)\frac{d\alpha^*}{dh}\varphi(h)$. Thus $T'(h) > 0$ whenever:

$$\left|\frac{d\alpha^*}{dh}\right| < \frac{\varrho(h,\alpha^*(h))}{\lambda(1-\mu)\varphi(h)} \tag{S1}$$

This is a weak condition: it requires only that the policy function not change so rapidly as to reverse the direction of skill dynamics.[4] We maintain (S1) as a regularity condition.[5]

With T monotone increasing on $[\underline{h}, \bar{h}]$, convergence from any $h_0 \in [\underline{h}, \bar{h}]$ follows by cases. *Case (a):* $h_0 > h^*$. Since $T(\bar{h}) < \bar{h}$ (Step 1), $h_1 < h_0$. By monotonicity of T: $h_{t+1} = T(h_t) \leq T(h_{t-1}) = h_t$ for all t. The orbit is monotone decreasing and bounded below by h^* (since $T(h^*) = h^*$ and T monotone imply $T(h) \geq T(h^*) = h^*$ for $h \geq h^*$). By the monotone convergence theorem, $h_t \to \hat{h}$ for some $\hat{h} \geq h^*$; continuity gives $T(\hat{h}) = \hat{h}$. *Case (b):* $h_0 < h^*$. Since $T(\underline{h}) > \underline{h}$ (Step 2), $h_1 > h_0$ when $h_0 = \underline{h}$. For general $h_0 \in [\underline{h}, h^*)$: monotonicity of T and $T(h^*) = h^*$ give $T(h_0) \leq T(h^*) = h^*$, and $T(h_0) \geq T(\underline{h}) > \underline{h}$, so the orbit stays in $[\underline{h}, h^*]$. If $T(h_0) > h_0$, the orbit is monotone increasing and bounded above by h^*, converging to a fixed point $\hat{h} \leq h^*$. If $T(h_0) = h_0$, then h_0 is itself a fixed point. If $T(h_0) < h_0$, then since $T(\underline{h}) > \underline{h}$ and $T(h_0) < h_0$, there exists a fixed point in $(\underline{h}, h_0)$ by the IVT; the orbit is monotone decreasing and converges to this fixed point. *Case (c):* $h_0 \in (0, \underline{h})$. Since $T(h) \geq \underline{T}(h) = (1-\delta)h + \lambda\mu\varphi(h) \geq \underline{h}$ for h near 0 (because $\lambda\mu\varphi(0) > 0$), the orbit enters $[\underline{h}, \bar{h}]$ in finitely many steps, after which Case (a) or (b) applies.

In all cases the orbit converges to a fixed point of T in $[\underline{h}, \bar{h}]$.

Step 4: $T'(h^*) < 1$**.** Consider specifically the orbit from $h_0 = \bar{h}$, which is monotone decreasing to h^* (Step 3, Case (a), using $T(\bar{h}) < \bar{h}$ from Step 1). Since $h_t > h^*$ for all t, we have $T(h_t) = h_{t+1} < h_t$, so $\Phi(h) \equiv T(h) - h < 0$ on $(h^*, \bar{h})$.[6] Since $\Phi(h^*) = 0$ and $\Phi(h) < 0$

[4]The bound (S1) is much weaker than the regularity condition (S3) maintained for monotonicity of optimal paths. Here we need only $T' > 0$ (monotone transition); condition (S3) further requires $T' < 1$ (contractive transition). The denominator $\lambda(1-\mu)\varphi(h)$ is the sensitivity of learning to adoption, so (S1) says adoption does not respond to skill changes more than the direct effect of skill on learning. Note that once $d\alpha^*/dh < 0$ is established (as in Part (iv) of the Proposition 2 proof below), the feedback term $-\lambda(1-\mu)(d\alpha^*/dh)\varphi(h)$ is positive and (S1) holds automatically without any magnitude restriction.

[5]For the calibrations in the Supplemental Appendix, $|d\alpha^*/dh|$ is bounded and (S1) holds comfortably. The condition requires only that the policy function be sufficiently smooth – a mild requirement for the class of dynamic optimization problems considered here.

[6]For any $c \in (h^*, \bar{h})$, the monotone decreasing orbit $h_t \downarrow h^*$ gives t with $h_{t+1} \leq c \leq h_t$. By monotonicity of T: $T(c) \leq T(h_t) = h_{t+1} \leq c$, so $\Phi(c) \leq 0$. If $T(c) = c$ (i.e., c is a fixed point), then T maps $[c, \bar{h}]$ into itself and the orbit from $\bar{h}$ converges to $c > h^*$, contradicting $h_t \to h^*$.

for $h > h^*$ near h^*, we get $\Phi'(h^*) \leq 0$, i.e., $T'(h^*) \leq 1$. The inequality is strict: if $T'(h^*) = 1$, then $\Phi'(h^*) = 0$ and $\Phi(h) < 0$ for $h > h^*$ gives $\Phi''(h^*) \leq 0$. But T monotone increasing (Step 3) with $T(h^*) = h^*$ and $T'(h^*) = 1$ means $\Phi(h) = T(h) - h$ has a degenerate zero: $\Phi(h^*) = \Phi'(h^*) = 0$ with $\Phi''(h^*) \leq 0$. This is a codimension-1 condition on the primitives.[7] Thus $T'(h^*) < 1$. □

Lemma 7 (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

[7] Formally, define $\Pi(\theta) \equiv T'(h^*(\theta); \theta) - 1$ where $\theta = (\delta, \lambda, \mu, \gamma, \beta, A)$ is the parameter vector and $h^*(\theta)$ is the steady state. By the implicit function theorem (applied to $T(h; \theta) - h = 0$ with $\partial_h[T - h] = T' - 1$), h^* is a smooth function of θ whenever $T'(h^*) \neq 1$. The map $\theta \mapsto \Pi(\theta)$ is smooth, and its level set $\Pi^{-1}(0)$ has measure zero in $\mathbb{R}^6$ provided $D_\theta \Pi \neq 0$, which holds because Π depends non-degenerately on δ (through $T' = (1 - \delta) + \ldots$). Thus $T'(h^*) = 1$ is a codimension-1 condition, excluded for generic parameter values.

first term $-(1-\gamma)(1-\alpha)^{-\gamma} < 0$. For the bracketed expression when $\mu < 1$: $V''(h') \leq 0$ by concavity, $\varphi(h) > 0$, and $\frac{\partial h'}{\partial h} = (1-\delta) + \lambda\ell(\alpha)\varphi'(h) > 0$ (the positive transition slope condition in Assumption 1). Thus the first bracketed term $V''(h')(\partial h'/\partial h)\varphi(h)$ 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 $\det(\mathbf{J}) = D_h D_\alpha - D_{h\alpha} D_{\alpha h}$. We derive a clean factorization. By the implicit function theorem applied to the FOC, $d\alpha^*/dh = -D_{\alpha h}/D_\alpha$, so $D_{\alpha h} = -D_\alpha \cdot (d\alpha^*/dh)$. Substituting:

$$\det(\mathbf{J}) = D_h D_\alpha - D_{h\alpha} \cdot (-D_\alpha)(d\alpha^*/dh) = D_\alpha \left[D_h + D_{h\alpha} \cdot \frac{d\alpha^*}{dh} \right]$$

The bracketed term equals $1 - T'(h^*)$, where $T(h) = (1-\delta)h + \lambda\ell(\alpha^*(h))\varphi(h)$ is the transition map. To see this: $T'(h) = (1-\delta) + \lambda\ell(\alpha^*)\varphi'(h) + \lambda\ell'(\alpha^*)(d\alpha^*/dh)\varphi(h)$, and $D_h = \delta - \lambda\ell(\alpha)\varphi'(h)$, $D_{h\alpha} = \lambda(1-\mu)\varphi(h) = -\lambda\ell'(\alpha)\varphi(h)$, so $D_h + D_{h\alpha}(d\alpha^*/dh) = \delta - \lambda\ell\varphi' - \lambda\ell'(d\alpha^*/dh)\varphi = 1 - T'(h^*)$. By Lemma 6 (Step 4), $T'(h^*) < 1$ at the steady state to which the orbit converges, so $1 - T'(h^*) > 0$. Since $D_\alpha < 0$ (established above):

$$\det(\mathbf{J}) = \underbrace{D_\alpha}_{<0} \cdot \underbrace{[1 - T'(h^*)]}_{>0} < 0$$

In particular, $\det(\mathbf{J}) \neq 0$, and the sign is unambiguously negative at the steady state to which the dynamics converge. This holds for all $\beta \in (0,1)$ and requires no additional dominance conditions beyond the stability of the transition map (which is guaranteed by Assumption 1). □

B.2 Proofs of Main Results

Proof of Proposition 1. 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, so adoption is lower than the static optimum; the shadow cost is increasing in $(1-\mu)$, giving $\partial\alpha^*/\partial\mu > 0$. For part (ii): when $\mu = 1$, the RHS is zero and the FOC reduces to the static condition $\partial Y/\partial\alpha = 0$. For the comparative static $\partial\alpha^*/\partial\mu > 0$: define $F(\alpha, \mu) \equiv Y_\alpha + \beta V'(h')\lambda(\mu - 1)\varphi(h) = 0$. Differentiating with respect to μ at fixed h: $F_\mu = \beta V'(h')\lambda\varphi(h) + \beta V''(h')\lambda(\mu-1)\varphi(h)\cdot\lambda\alpha\varphi(h)$. The first term is strictly positive. The second term equals $\beta|V''(h')|\lambda^2\alpha(1-\mu)\varphi(h)^2 \geq 0$ (since $V'' \leq 0$ and $\mu - 1 < 0$). Thus $F_\mu > 0$. Since $F_\alpha = D_\alpha < 0$ (Lemma 7), $d\alpha^*/d\mu = -F_\mu/F_\alpha > 0$. □

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

Proof of Proposition 2.

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 the Steady-State Human Capital Lemma. 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 $h_S(0) = \bar{h}$. At $\alpha = 1$, $\ell(1) = \mu > 0$, so stationarity gives $\delta h_S(1) = \lambda\mu\varphi(h_S(1))$, which has a unique positive solution $h_S(1) > 0$.[8]

For the optimal policy P: by Lemma 4, 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, continuity and the intermediate value theorem guarantee at least one crossing. For global uniqueness, consider the transition map $T(h) = (1-\delta)h + \lambda\ell(\alpha^*(h))\varphi(h)$, whose fixed points are exactly the equilibrium steady states (intersections of S and P). Under condition (S1) (maintained for convergence), $T'(h) > 0$ for all $h \in [\underline{h}, \bar{h}]$. We additionally maintain that $T'(h) < 1$ on $[\underline{h}, \bar{h}]$.[9] With $T' \in (0, 1)$: $\Phi(h) \equiv T(h) - h$ satisfies $\Phi'(h) = T'(h) - 1 < 0$, so Φ is strictly decreasing. A strictly decreasing continuous function has at most one zero. Combined with $\Phi(\underline{h}) > 0$ (Step 2 of Lemma 6) and $\Phi(\bar{h}) < 0$ (Step 1 of Lemma 6), there is exactly one zero: the equilibrium steady state h^* is unique.

[8] The locus $h_S(\alpha)$ would reach zero only at $\alpha = 1/(1-\mu) > 1$ for $\mu \in (0, 1)$, which is outside the feasible domain $[0, 1]$.

[9] This contraction condition is stronger than (S1) and requires $|d\alpha^*/dh| < D_h/D_{h\alpha}$, where $D_h = \delta - \lambda\ell\varphi' > 0$ and $D_{h\alpha} = \lambda(1-\mu)\varphi > 0$. Under the regularity condition (S3) established in Part (iv) below, $|d\alpha^*/dh|$ is bounded by $|S|/|D_\alpha|$, which is finite and typically small relative to $D_h/D_{h\alpha}$. The contraction condition is verified numerically for the calibrations in the Supplemental Appendix.

Part (ii): Uniqueness. The Jacobian non-singularity established in Lemma 7 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 6, 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 $F(\alpha, h) \equiv Y_\alpha(h, \alpha) + \beta V'(h') \cdot \partial h'/\partial \alpha = 0$ where $\partial h'/\partial \alpha = \lambda(\mu - 1)\varphi(h) < 0$. By the implicit function theorem, $d\alpha^*/dh = -F_h/F_\alpha$. Since $F_\alpha = D_\alpha < 0$ (established in Lemma 7), the sign of $d\alpha^*/dh$ equals the sign of F_h. We show $F_h < 0$.

Decompose F_h into a static cross-partial and a dynamic feedback:

$$F_h = \underbrace{-(1-\gamma)(1-\alpha^*)^{-\gamma}}_{\text{static: } S<0} + \underbrace{\beta\lambda(1-\mu)\Big[|V''(h^*)|\varrho\,\varphi(h^*) + V'(h^*)|\varphi'(h^*)|\Big]}_{\text{dynamic: } D\geq 0} \tag{S2}$$

where $\varrho = (1-\delta) + \lambda\ell(\alpha^*)\varphi'(h^*) \in (0,1)$ is the partial transition slope (positive by Assumption 1). All quantities in S and D are determined by the steady state (h^*, α^*), which is itself pinned down by primitives $(\delta, \lambda, \mu, \gamma, \beta, A, g, \varphi)$. The static term $|S| = (1-\gamma)(1-\alpha^*)^{-\gamma}$ is increasing in α^* (more delegation sharpens the skill-adoption complementarity) and in $1-\gamma$ (stronger effort concentration). The dynamic term D is bounded: $V'(h^*)$ and $|V''(h^*)|$ are finite on the compact invariant set $[h^*, \bar{h}]$ (Lemma 3); $\varphi(h^*)$ and $|\varphi'(h^*)|$ are bounded by Assumption 1; $\varrho < 1$ by stability; and $\beta(1-\mu) < 1$. Thus D is a bounded, continuous function of the steady state.

The condition $|S| > D$ is:

$$(1-\gamma)(1-\alpha^*)^{-\gamma} > \beta\lambda(1-\mu)\Big[|V''(h^*)|\varrho\,\varphi(h^*) + V'(h^*)|\varphi'(h^*)|\Big] \tag{S3}$$

We maintain this as a regularity condition.[10] When (S3) holds, $F_h < 0$ and $d\alpha^*/dh = -F_h/D_\alpha < 0$.

[10]The condition holds at $\beta = 0$ (the dynamic term vanishes) and fails only if the dynamic feedback through V'' and V' overwhelms the static cross-partial. Economically, it requires that effort concentration effects (governed by $\gamma < 1$) dominate the value-function curvature. When $\gamma \to 1$ (no concentration benefit), $|S| \to 0$ and the condition fails; the sign of $d\alpha^*/dh$ becomes ambiguous. For γ bounded away from 1 and β bounded away from 1, the condition holds generically because $|S|$ grows as $(1-\alpha^*)^{-\gamma}$ while D remains bounded.

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 2. 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.

Moreover, when $h^* \geq \bar{h}$, long-run output under adoption weakly exceeds the no-adoption benchmark. We show $Y^* \geq \bar{h}$ by a revealed preference argument. At steady state, $V(h^*) = Y^*/(1-\beta)$. The firm could instead choose $\alpha = 0$ forever starting from h^*. Since $h^* \geq \bar{h}$ and φ is decreasing: $\lambda\varphi(h^*) \leq \lambda\varphi(\bar{h}) = \delta\bar{h} \leq \delta h^*$, so the no-adoption path from h^* satisfies $h_t \geq \bar{h}$ for all t (it converges downward to $\bar{h}$).[11] Thus the no-adoption value from h^* is at least $\bar{h}/(1-\beta)$. By optimality: $V(h^*) \geq \bar{h}/(1-\beta)$, so $Y^*/(1-\beta) \geq \bar{h}/(1-\beta)$, giving $Y^* \geq \bar{h} = Y^{NA}$. When $\mu \geq 1$, adoption cannot reduce long-run output. □

Corollary (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 7, $\det(\mathbf{J}) \neq 0$. Under the conditions established in that

[11] With $\alpha = 0$: $h' = (1-\delta)h + \lambda\varphi(h)$. At $h = h^* \geq \bar{h}$: $h' = (1-\delta)h^* + \lambda\varphi(h^*) \leq (1-\delta)h^* + \delta h^* = h^*$, and $h' \geq (1-\delta)\bar{h} + \lambda\varphi(\bar{h}) = \bar{h}$ by the mean value theorem (the factor $(1-\delta) + \lambda\varphi'(\xi)$ is positive by Assumption 1). By induction, $h_t \in [\bar{h}, h^*]$ for all t, converging to $\bar{h}$.

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 $\frac{\partial F_2}{\partial \beta} = -V'(h^*)\lambda(1-\mu)\varphi(h^*)$, where we suppress the additional term $\beta\frac{\partial V'(h^*)}{\partial \beta}\lambda(\mu-1)\varphi(h^*)$ that arises because V' itself depends on β through the Bellman equation.[12] 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 μ: The claim $\partial\alpha^*/\partial\mu > 0$ is a statement about the policy function $\alpha^*(h)$ at fixed h, not a steady-state comparative static. From the FOC, define $F(\alpha, \mu; h) \equiv Y_\alpha + \beta V'(h')\lambda(\mu-1)\varphi(h) = 0$. Differentiating with respect to μ at fixed h:

$$F_\mu = \beta V'(h')\lambda\varphi(h) + \beta V''(h')\lambda(\mu-1)\varphi(h) \cdot \frac{\partial h'}{\partial \mu}$$

The first term is $\beta V'(h')\lambda\varphi(h) > 0$. For the second term: $\partial h'/\partial\mu = \lambda\alpha\varphi(h) > 0$ (higher μ raises next-period skill through the learning channel), $V''(h') \leq 0$ by concavity, and $(\mu-1) < 0$ when $\mu < 1$. Thus the second term equals $\beta|V''(h')|\lambda^2\alpha(1-\mu)\varphi(h)^2 \geq 0$. Both terms are non-negative with the first strictly positive, so $F_\mu > 0$. Since $F_\alpha = D_\alpha < 0$ (Lemma 7), $d\alpha^*/d\mu = -F_\mu/F_\alpha > 0$.

For the net steady-state effect on h^*: implicitly differentiate the stationarity condition $\delta h^* = \lambda[1-(1-\mu)\alpha^*]\varphi(h^*)$:

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

The denominator is positive by Assumption 1. The numerator has a positive direct effect ($\lambda\alpha^*\varphi(h^*) > 0$: higher μ means more learning per delegated task) and a negative indirect effect ($-\lambda(1-\mu)\varphi(h^*)\frac{\partial\alpha^*}{\partial\mu} < 0$: higher μ induces more adoption, reducing learning). The steady-state sign is ambiguous in general; the policy-function result $\partial\alpha^*/\partial\mu > 0$ at fixed h is

[12] The omitted term has the same sign as the retained term when $\mu < 1$: higher β raises $V'(h^*)$ (from the envelope condition), so $\frac{\partial V'(h^*)}{\partial\beta} > 0$, making $\beta\frac{\partial V'}{\partial\beta}\lambda(\mu-1)\varphi(h^*) < 0$. Thus the total $\partial F_2/\partial\beta < 0$, and the sign conclusion is unchanged.

unambiguous.

Effect of λ (referenced in discussion): $\frac{\partial F_1}{\partial \lambda} = -\ell(\alpha^*)\varphi(h^*) < 0$ and $\frac{\partial F_2}{\partial \lambda} = \beta V'(h^*)(\mu - 1)\varphi(h^*) < 0$ when $\mu < 1$. By Cramer's rule, $\frac{\partial h^*}{\partial \lambda} > 0$: faster learners maintain higher skills. □

Lemma 8 (Learning Spillover Properties)**.** *If $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is strictly increasing on $(0, \bar{H})$ with $\psi(\bar{H}) = 1$, then along any path where $H_t < \bar{H}$, we have $\psi(H_t) < 1$.*

Proof. Since ψ is strictly increasing on $(0, \bar{H})$ and $H_t \in (0, \bar{H})$, we have $\psi(H_t) < \psi(\bar{H}) = 1$.[13] □

Lemma 9 (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. □

Proof of Proposition 3. Part (i): We establish two claims about Δ_t^{SC}.

Claim 1: Bounded absolute gain, growing relative gain. By the Steady-State Human Capital Lemma and Proposition 2, $h_t^U \to h^* < \bar{h}$ 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.

[13] The maintained specification $\psi(H) = (H/\bar{H})^\eta$ with $\eta > 0$ is strictly increasing on all of $\mathbb{R}_+$, so the condition holds. The lemma applies more generally to any ψ that is strictly increasing below the no-adoption steady state.

Claim 2: Ratio eventually increases. By the monotonicity established in Section 2.4, 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 $R_t \equiv \Delta_t^{SC}/h_t^U = AG(\alpha_t)/h_t^U - [1-(1-\alpha_t)^{1-\gamma}]$. Along the optimal path, $\alpha_t = \alpha^*(h_t)$ where $d\alpha^*/dh < 0$, so R_t can be expressed as a function of h_t alone: $R(h) = AG(\alpha^*(h))/h - [1-(1-\alpha^*(h))^{1-\gamma}]$. Differentiating:

$$R'(h) = \underbrace{\frac{Ag(\alpha^*)(d\alpha^*/dh)\cdot h - AG(\alpha^*)}{h^2}}_{\equiv T_1 < 0} + \underbrace{(1-\gamma)(1-\alpha^*)^{-\gamma}\left(-\frac{d\alpha^*}{dh}\right)}_{\equiv T_2 > 0}$$

where $T_1 < 0$ because $d\alpha^*/dh < 0$ makes the numerator's first term negative and $AG(\alpha^*) > 0$. For T_2: differentiating $-[1-(1-\alpha^*)^{1-\gamma}]$ with respect to h gives $-(1-\gamma)(1-\alpha^*)^{-\gamma}\cdot d(1-\alpha^*)/dh = (1-\gamma)(1-\alpha^*)^{-\gamma}(-d\alpha^*/dh)$, which is *positive* since $d\alpha^*/dh < 0$.

To sign $R'(h) = T_1 + T_2$, collect the terms involving $d\alpha^*/dh$:

$$R'(h) = \frac{d\alpha^*}{dh}\left[\frac{Ag(\alpha^*)}{h} - (1-\gamma)(1-\alpha^*)^{-\gamma}\right] - \frac{AG(\alpha^*)}{h^2}$$

The last term is negative. For the bracketed expression: from the FOC (6), $Ag(\alpha^*) = h(1-\gamma)(1-\alpha^*)^{-\gamma} + \beta V'(h')\lambda(1-\mu)\varphi(h)$, so

$$\frac{Ag(\alpha^*)}{h} - (1-\gamma)(1-\alpha^*)^{-\gamma} = \frac{\beta V'(h')\lambda(1-\mu)\varphi(h)}{h} > 0$$

when $\mu < 1$. The bracketed expression is positive, and $d\alpha^*/dh < 0$, so the first term is negative. Both terms in $R'(h)$ are negative, giving $R'(h) < 0$: the ratio is decreasing in h. Since h_t is decreasing, $R_t = R(h_t)$ is increasing in t.

Part (ii): By Lemma 9, $\Delta_t^{SC} > 0$ for all $t \geq 0$ along the optimal path: the state-conditional gain is strictly positive whenever the firm optimally adopts. □

Corollary (Welfare Reversal Under Patient Evaluation). Consider the cumulative path counterfactual (Remark 1) $\bar{\Delta}^{PATH}(\tilde{\beta}) = \sum_{t=0}^{\infty}\tilde{\beta}^t\Delta_t^{PATH} = \sum_{t=0}^{\infty}\tilde{\beta}^t[Y_t^{user} - Y_t^{NA}]$. For the firm's own discount factor β, revealed preference implies $\bar{\Delta}^{PATH}(\beta) \geq 0$. When $Y^* < \bar{h}$, we have $Y_t^{user} \to Y^* < \bar{h} = Y_t^{NA}$ as $t \to \infty$, so $\Delta_t^{PATH} < 0$ for all sufficiently large t. For $\tilde{\beta}$ sufficiently larger than β, the negative tail dominates, giving $\bar{\Delta}^{PATH}(\tilde{\beta}) < 0$. □

Proof of Proposition 4. 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.

Aggregate human capital is $H_t = \varpi h_t^U + (1-\varpi)h_t^{NU}$ where $\varpi \in (0,1)$ is the user share.[14]

With learning spillovers $\psi(H)$, non-users' skill accumulation depends on aggregate human capital: $h_{t+1}^{NU} = (1-\delta)h_t^{NU} + \lambda\varphi(h_t^{NU}) \cdot \psi(H_t)$. At $t = 0$, all workers begin at $\bar{h}$, so $H_0 = \bar{h}$ and $\psi(H_0) = 1$. Thus $h_1^{NU} = (1-\delta)\bar{h} + \lambda\varphi(\bar{h}) = \bar{h}$: non-users are unaffected in the first period. At $t = 1$, user skills have declined ($h_1^U < \bar{h}$), so $H_1 = \varpi h_1^U + (1-\varpi)\bar{h} < \bar{h}$ and $\psi(H_1) < 1$ by Lemma 8. Therefore $h_2^{NU} = (1-\delta)\bar{h} + \lambda\varphi(\bar{h})\psi(H_1) < \bar{h}$. By induction, $h_t^{NU} < \bar{h}$ for all $t \geq 2$.

We show h_t^{NU} is strictly decreasing from $\bar{h}$ for $t \geq 2$ toward h^{NU*}. Define the non-user transition map $T^{NU}(h; H_t) = (1-\delta)h + \lambda\varphi(h)\psi(H_t)$. For $t \geq 1$, $\psi(H_t) < 1$, so $T^{NU}(\bar{h}; H_t) < \bar{h}$. 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 $h \in (h^{NU*}, \bar{h}]$ when $\psi(H_t) < 1$, and $h_1^{NU} = \bar{h}$ with $h_2^{NU} < \bar{h}$, the sequence $\{h_t^{NU}\}_{t\geq 1}$ is strictly decreasing. Therefore $s_t = \bar{h} - h_t^{NU}$ satisfies $s_0 = s_1 = 0$ and is strictly increasing for $t \geq 2$.

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 $t \geq 2$. The gap $s_t = \bar{h} - h_t^{NU}$ is strictly increasing for $t \geq 2$ because $\{h_t^{NU}\}_{t\geq 1}$ is strictly decreasing toward h^{NU*} (shown above); equivalently, the positive forcing term $\lambda[\varphi(h_t^{NA}) - \varphi(h_t^{NU})\psi(H_t)]$ in the recursion (14) dominates δs_t at each t, since s_t converges to the finite limit $\bar{h} - h^{NU*}$ while the forcing term remains bounded away from zero as long as $\psi(H_t) < 1$. □

Proof of Proposition 5. Recall $Y(h, \alpha; A) = AG(\alpha) + h(1-\alpha)^{1-\gamma}$. Let h_t^U denote the skill path under adoption $\alpha^* > 0$ and h_t^{NU} the skill path under no adoption ($\alpha = 0$), both starting from $h_0 = \bar{h}$. When $\mu < 1$, $h_t^U < h_t^{NU}$ for all $t \geq 1$ (established in the proof of state-path divergence). User skills converge: $h_t^U \to h^* < \bar{h}$. Non-user skills satisfy $h_t^{NU} \to \bar{h}$ when $\eta = 0$

[14] The result holds for any fixed $\varpi > 0$. When $\varpi = 1$ (all workers are users), $H_t = h_t^U$ and h_t^{NU} represents a hypothetical non-user in the degraded learning environment. The spillover channel requires $\varpi < 1$ for there to exist actual non-users whose skills are measured in cross-sectional comparisons.

and $h_t^{NU} \to h^{NU*} < \bar{h}$ when $\eta > 0$.

Part (i): The RCT estimand at time t is $\tau^{RCT}(t) = Y(h_t^U, \alpha^*) - Y(h_t^{NU}, 0) = AG(\alpha^*) + h_t^U(1-\alpha^*)^{1-\gamma} - h_t^{NU}$, where h_t^{NU} denotes the control group's skill path (non-users in the same economy; with $\eta = 0$, $h_t^{NU} = h_t^{NA}$). The first difference is:

$$\tau^{RCT}(t+1) - \tau^{RCT}(t) = (h_{t+1}^U - h_t^U)(1-\alpha^*)^{1-\gamma} - (h_{t+1}^{NU} - h_t^{NU})$$

Along the equilibrium transition from $h_0 = \bar{h}$: (a) user skills are strictly decreasing ($h_{t+1}^U < h_t^U$, established in Section 2.4), so the first term is negative. (b) When $\eta = 0$: non-user skills are constant at $\bar{h}$, so the second term is zero and τ^{RCT} is strictly decreasing. When $\eta > 0$: non-user skills are also declining (converging to $h^{NU*} < \bar{h}$, as established in the proof of Proposition 4), so the second term $-(h_{t+1}^{NU} - h_t^{NU}) > 0$ partially offsets the first. To sign the net effect, write the skill transitions explicitly:

$$\begin{aligned} h_{t+1}^U - h_t^U &= -\delta h_t^U + \lambda \ell(\alpha^*)\varphi(h_t^U)\psi(H_t) \\ h_{t+1}^{NU} - h_t^{NU} &= -\delta h_t^{NU} + \lambda \varphi(h_t^{NU})\psi(H_t) \end{aligned}$$

Substituting into the first difference and using $h_0^U = h_0^{NU} = \bar{h}$:

$$\begin{aligned} \tau^{RCT}(t+1) - \tau^{RCT}(t) = (1-\alpha^*)^{1-\gamma}[-\delta h_t^U + \lambda \ell(\alpha^*)\varphi(h_t^U)\psi(H_t)] \\ - [-\delta h_t^{NU} + \lambda \varphi(h_t^{NU})\psi(H_t)] \end{aligned}$$

At $t = 0$ (where $h_0^U = h_0^{NU} = \bar{h}$ and $\psi(H_0) = 1$), this simplifies to $\delta\bar{h}[1-(1-\alpha^*)^{1-\gamma}] - \lambda\varphi(\bar{h})[1-\ell(\alpha^*)(1-\alpha^*)^{1-\gamma}]$. Using the stationarity condition $\delta\bar{h} = \lambda\varphi(\bar{h})$ and $\ell(\alpha^*) = 1-(1-\mu)\alpha^*$:

$$\tau^{RCT}(1) - \tau^{RCT}(0) = \lambda\varphi(\bar{h})\big\{[1-(1-\alpha^*)^{1-\gamma}] - [1-\ell(\alpha^*)(1-\alpha^*)^{1-\gamma}]\big\} = -\lambda\varphi(\bar{h})(1-\mu)\alpha^*(1-\alpha^*)^{1-\gamma} < 0$$

For $t \geq 1$, we require $(1-\alpha^*)^{1-\gamma}|h_{t+1}^U - h_t^U| > |h_{t+1}^{NU} - h_t^{NU}|$. Using the transition equations:

$$\begin{aligned} h_{t+1}^U - h_t^U &= -\delta h_t^U + \lambda \ell(\alpha^*)\varphi(h_t^U)\psi(H_t) \\ h_{t+1}^{NU} - h_t^{NU} &= -\delta h_t^{NU} + \lambda \varphi(h_t^{NU})\psi(H_t) \end{aligned}$$

The user decline has magnitude at least $\lambda(1-\mu)\alpha^*\varphi(h_t^U)\psi(H_t)$ from the adoption penalty ($\ell(\alpha^*) < 1$), plus $\delta(h_t^{NU} - h_t^U) \geq 0$ from the depreciation differential. The non-user decline has magnitude at most $\lambda\varphi(h_t^{NU})[1-\psi(H_t)]$ from the spillover penalty alone. Under the specification $\psi(H) = (H/\bar{H})^\eta$, the spillover penalty satisfies $1 - \psi(H_t) \leq \eta(\bar{H} - H_t)/\bar{H}$ (first-

order approximation), which is bounded by $\eta \cdot \varpi(\bar{h} - h_t^U)/\bar{H}$. The adoption penalty $(1-\mu)\alpha^*$ is a discrete, first-order effect, while the spillover-induced non-user decline is proportional to η and to the gradually accumulating skill gap. The condition

$$\eta < \frac{(1-\mu)\alpha^*}{(1-\alpha^*)^{1-\gamma}} \tag{S4}$$

ensures the weighted user decline exceeds the non-user decline at each t: the numerator bounds the direct adoption effect and the denominator accounts for the effort-concentration scaling $(1-\alpha^*)^{1-\gamma}$ applied to user skill changes in the RCT estimand.[15] Under condition (S4), $\tau^{RCT}(t+1) < \tau^{RCT}(t)$ for all t along the transition.

Part (ii): The state-conditional gain is $\Delta_t^{SC} = Y(h_t^U, \alpha_t) - Y(h_t^U, 0) = AG(\alpha_t) - h_t^U[1 - (1-\alpha_t)^{1-\gamma}]$. Define $\Omega_t \equiv 1 - (1-\alpha_t)^{1-\gamma} > 0$. Then:

$$\Delta_{t+1}^{SC} - \Delta_t^{SC} = A[G(\alpha_{t+1}) - G(\alpha_t)] - h_{t+1}^U \Omega_{t+1} + h_t^U \Omega_t$$

Along the equilibrium transition, h_t is decreasing and $\alpha_t = \alpha^*(h_t)$ is increasing (Section 2.4, Part (iv)). We prove Δ_t^{SC} is increasing by showing it is a decreasing function of h along the optimal policy. Define $f(h) \equiv \Delta^{SC}(h) = AG(\alpha^*(h)) - h[1 - (1-\alpha^*(h))^{1-\gamma}]$ and write $\Omega(\alpha) \equiv 1 - (1-\alpha)^{1-\gamma}$. Differentiating with respect to h:

$$f'(h) = \left[Ag(\alpha^*) - h(1-\gamma)(1-\alpha^*)^{-\gamma}\right]\frac{d\alpha^*}{dh} - \Omega(\alpha^*)$$

where we used $d\Omega/d\alpha = (1-\gamma)(1-\alpha)^{-\gamma}$. From the firm's FOC (6), $Ag(\alpha^*) - h(1-\gamma)(1-\alpha^*)^{-\gamma} = \beta V'(h')\lambda(1-\mu)\varphi(h) > 0$ when $\mu < 1$. Denote this positive quantity by $\Lambda(h) > 0$. Since $d\alpha^*/dh < 0$ (established in Part (iv)):

$$f'(h) = \underbrace{\Lambda(h)}_{>0} \cdot \underbrace{\frac{d\alpha^*}{dh}}_{<0} - \underbrace{\Omega(\alpha^*)}_{>0} < 0$$

Both terms are negative, so $f'(h) < 0$ unconditionally. Since h_t is strictly decreasing along the equilibrium transition, $\Delta_t^{SC} = f(h_t)$ is strictly increasing in t. Thus $\Delta_{t+1}^{SC} > \Delta_t^{SC}$.

Part (iii): As $t \to \infty$, $h_t^U \to h^*$ and $h_t^{NU} \to h^{NU*}$ where $h^{NU*} = \bar{h}$ when $\eta = 0$ and $h^{NU*} < \bar{h}$ when $\eta > 0$. The limiting RCT estimand is $\tau^{RCT}(\infty) = AG(\alpha^*) + h^*(1-\alpha^*)^{1-\gamma} - h^{NU*}$. For this to be negative, we need the static AI gain to be smaller than the skill loss:

[15] When $\eta = 0$, non-user skills are constant and the result holds unconditionally. When η violates (S4), the non-user decline can exceed the weighted user decline, and Part (i) requires strengthening.

$AG(\alpha^*) < h^{NU*} - h^*(1-\alpha^*)^{1-\gamma}$. This holds when μ is sufficiently below 1 (producing large $\bar{h} - h^*$) or when β is sufficiently small (impatient firms adopt heavily). Since $\tau^{RCT}(0) > 0$ (pure static gain at $t = 0$ with $h_0^U = h_0^{NU} = \bar{h}$) and $\tau^{RCT}(\infty) < 0$ under these conditions, monotonicity from Part (i) implies there exists unique $\hat{t}$ where $\tau^{RCT}(\hat{t}) = 0$. □ **Proof of Proposition 6. Part (i):** Consider workers with ability θ_i, so $\varphi_i(h) = \theta_i \varphi(h)$. The skill dynamics are $h_{t+1} = (1-\delta)h_t + \lambda\theta_i \ell(\alpha_t)\varphi(h_t)$. We establish the result holding the adoption path $\{\alpha_t\}$ fixed across ability types; this corresponds to a common AI policy or to the case where adoption is determined by firm-level choices rather than individual ability. The result extends to endogenous adoption $\alpha_t = \alpha^*(h_t, \theta)$ as follows: since $d\alpha^*/dh < 0$ (the monotonicity established in Section 2.4) and higher-ability workers have $h_t^{NA}(\theta_H) > h_t^{NA}(\theta_L)$ for $\theta_H > \theta_L$ (ability scales learning), we have $\alpha^*(h_t^{NA}(\theta_H)) < \alpha^*(h_t^{NA}(\theta_L))$. Thus higher-ability workers optimally adopt less, forgoing less learning per period. Adding the adoption response to the gap recursion introduces a term $\partial\Delta_{t+1}/\partial\alpha_t \cdot (\partial\alpha_t/\partial\theta)$. The first factor $\partial\Delta/\partial\alpha > 0$ when $\mu < 1$ (more delegation increases the skill gap), and the second factor $\partial\alpha^*/\partial\theta < 0$ through the h-channel (higher ability → higher h → lower adoption), so the product is *negative*: the endogenous adoption response partially offsets the direct learning channel, because high-ability workers choose less delegation. However, the direct channel dominates: the direct term $\lambda\theta(1-\mu)\alpha_t\varphi(h_{t-1})$ is first-order in $(1-\mu)\alpha$, while the adoption-response term scales with $|d\alpha^*/dh| \cdot \Delta_t$, which is second-order along the transition where Δ_t is small relative to h_t (both paths start at $\bar{h}$ and diverge gradually). Formally, at $t = 1$ the adoption-response term is zero (since $\Delta_0 = 0$), and for $t \geq 2$ the ratio Δ_t/h_t remains bounded by a quantity proportional to $(1-\mu)\alpha \cdot t \cdot \lambda\varphi(\bar{h})/\bar{h}$, which is small relative to $(1-\mu)\alpha$ for the durations relevant to the induction. The net effect preserves $\partial\Delta_t/\partial\theta > 0$. 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, we use a different decomposition. Write $h_t^{NA}(\theta) = (1-\delta)h_{t-1}^{NA} +$

$\lambda\theta\varphi(h_{t-1}^{NA})$ and $h_t^U(\theta) = (1-\delta)h_{t-1}^U + \lambda\theta\ell(\alpha_{t-1})\varphi(h_{t-1}^U)$. Differentiating with respect to θ:

$$\frac{\partial h_t^{NA}}{\partial \theta} = [(1-\delta) + \lambda\theta\varphi'(h_{t-1}^{NA})]\frac{\partial h_{t-1}^{NA}}{\partial \theta} + \lambda\varphi(h_{t-1}^{NA})$$
$$\frac{\partial h_t^{U}}{\partial \theta} = [(1-\delta) + \lambda\theta\ell(\alpha_{t-1})\varphi'(h_{t-1}^{U})]\frac{\partial h_{t-1}^{U}}{\partial \theta} + \lambda\ell(\alpha_{t-1})\varphi(h_{t-1}^{U})$$

Define $a_t \equiv \partial h_t^{NA}/\partial\theta$ and $b_t \equiv \partial h_t^U/\partial\theta$. At $t=0$: $a_0 = b_0 = 0$ (both paths start from $h_0 = \bar{h}$ independent of θ). At $t=1$: $a_1 = \lambda\varphi(\bar{h})$ and $b_1 = \lambda\ell(\alpha_0)\varphi(\bar{h})$, so $a_1 - b_1 = \lambda(1-\mu)\alpha_0\varphi(\bar{h}) > 0 = \partial\Delta_1/\partial\theta$, confirming the base case.

For $t \geq 2$, we show $a_t - b_t > 0$ by induction. From the recursions:

$$a_t - b_t = (1-\delta)(a_{t-1} - b_{t-1}) + \lambda\theta[\varphi'(h_{t-1}^{NA})a_{t-1} - \ell\varphi'(h_{t-1}^U)b_{t-1}] + \lambda[\varphi(h_{t-1}^{NA}) - \ell\varphi(h_{t-1}^U)]$$

The first term is positive by induction. For the third term: from the gap recursion $\Delta_t = (1-\delta)\Delta_{t-1} + \lambda\theta[\varphi(h_{t-1}^{NA}) - \ell\varphi(h_{t-1}^U)]$ and $\Delta_t > 0$ for all t (since $h_t^{NA} > h_t^U$, which follows because the no-adoption path has strictly higher learning at $t=0$ and remains above by the same recursion), we cannot directly sign the third term – the learning gap $\varphi(h_{t-1}^{NA}) - \ell\varphi(h_{t-1}^U)$ has ambiguous sign because φ is decreasing ($h^{NA} > h^U$ pushes φ down) while $\ell < 1$ reduces the user term.

To handle this, we regroup the second and third terms:

$$\lambda[\varphi(h_{t-1}^{NA}) + \theta\varphi'(h_{t-1}^{NA})a_{t-1}] - \lambda\ell[\varphi(h_{t-1}^U) + \theta\varphi'(h_{t-1}^U)b_{t-1}]$$

The first bracket equals $\partial_\theta[\theta\varphi(h_{t-1}^{NA}(\theta))]$ and the second equals $\partial_\theta[\theta\varphi(h_{t-1}^U(\theta))]$. Since $\varphi' < 0$ and $\lambda\theta|\varphi'| < \delta$ (Assumption 1), both brackets are positive. By log-convexity of φ (Assumption 1), $|\varphi'(h)|/\varphi(h)$ is non-increasing in h. Since $h_{t-1}^{NA} > h_{t-1}^U$: $|\varphi'(h_{t-1}^{NA})|/\varphi(h_{t-1}^{NA}) \leq |\varphi'(h_{t-1}^U)|/\varphi(h_{t-1}^U)$, so the proportional reduction from the $\theta\varphi' a$ term is smaller for the NA path than for the U path. Combined with $\ell < 1$ reducing the second bracket, and the contraction bound ensuring neither bracket is negative, the difference of brackets is positive when the skill gap $h_{t-1}^{NA} - h_{t-1}^U$ is small relative to level (as it is along the transition from $h_0 = \bar{h}$, where both paths start at the same point and diverge gradually).[16] Thus $a_t - b_t > 0$,

[16] Formally, write $P^{NA} = \varphi(h^{NA})[1 - \theta|\varphi'(h^{NA})|a_{t-1}/\varphi(h^{NA})]$ and $P^U = \ell\varphi(h^U)[1 - \theta|\varphi'(h^U)|b_{t-1}/\varphi(h^U)]$. The contraction bound $\lambda\theta|\varphi'| < \delta < 1$ ensures both bracketed factors are in $(0,1)$. Log-convexity ensures the proportional reduction $r(h) \equiv \theta|\varphi'(h)|/\varphi(h)$ is non-increasing. Two forces favor $P^{NA} > P^U$: (i) the $\ell < 1$ factor shrinks P^U; (ii) $r(h^{NA}) \leq r(h^U)$ means the NA bracket is larger per unit of a than the U bracket per unit of b. One force opposes: $\varphi(h^{NA}) < \varphi(h^U)$ since φ is decreasing. Along the transition from $\bar{h}$, the opposing force is small (the paths are close) while the $\ell < 1$ advantage is discrete, so $P^{NA} > P^U$ and $a_t - b_t > (1-\delta)(a_{t-1} - b_{t-1}) > 0$.

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 the Steady-State Characterization Lemma. The vintage premium is $\pi_t = \bar{h}/h_t^{post} - 1$. Since h_t^{post} is decreasing toward $h^* < \bar{h}$ (the monotonicity established in Section 2.4), π_t is increasing in t until pre-AI cohorts retire. □

Proof of Corollary 1. 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 the Steady-State Human Capital Lemma), while pre-AI workers maintain $\bar{h}$. We focus on the between-cohort variance component, treating each cohort as having a representative skill level.[17] The between-cohort variance for a two-group population with masses N_t^{pre} and N_t^{post} and cohort-mean skills $\bar{h}$ and 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 T^{max} occurs when these balance, and is increasing in the retirement rate ν (slower cohort turnover delays the peak) and in the speed of skill atrophy $(1 - \mu)\alpha^*$ (faster atrophy front-loads the gap-widening effect). □

[17] With ability heterogeneity θ_i, within-cohort wage dispersion is non-degenerate. The between-cohort formula captures the dominant contribution to variance when the cohort skill gap $\bar{h} - h_t$ is large relative to within-cohort dispersion. The hump shape is driven by the between-cohort component; within-cohort variance evolves monotonically under the conditions of Proposition 6(i).