Skill Atrophy and AI Productivity Measurement

Tommaso Bondi* Gentry Johnson†

February 3, 2026

Abstract

How should we measure the productivity effects of generative AI? Recent experimental studies document substantial short-run gains. We show theoretically that measuring the long-run effects of AI introduces two structural sources of bias when adoption affects skill formation over time. In a dynamic model where workers learn by doing, the effects of AI delegation depend on AI's pedagogical quality. When AI delegation slows learning by substituting for cognitive effort, two effects arise. First, as adoption spreads, non-users become a degraded counterfactual because mentorship, spillovers, and training environments deteriorate, causing cross-sectional estimates to overstate lifetime effects (spillover bias). Second, even within-worker comparisons are distorted: state-conditional productivity gains can diverge from path-based comparisons because current skill is endogenous to past AI use, lowering the outside option against which AI is evaluated (state-path divergence). These biases can reverse the sign of estimated productivity effects in high-adoption sectors. We characterize when decentralized adoption is inefficient and discuss identification strategies that recover the welfare-relevant counterfactual.

JEL Codes: O33, J24, D62, L23
Keywords: Generative AI, human capital, learning-by-doing, productivity measurement, path dependence

*Cornell Tech & SC Johnson College 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 Guy Aridor, Ron Berman, Luis Cabral, Sachin Gupta, Brett Hollenbeck, Vrinda Kadiyali, Jura Liaukonytė, Xueming Luo, Emaad Manzoor, Ivan Png, Omid Rafieian, Michael Waldman, and Nathan Yang for helpful comments and suggestions.

1 Introduction

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

This matters because the tasks at which generative AI excels are often those through which humans build skill. Junior developers learn debugging by wrestling with broken code; legal associates develop judgment by drafting arguments from scratch; medical residents acquire diagnostic intuition by working through difficult cases. When AI performs these formative tasks, immediate output rises. But if AI substitutes for the cognitive effort through which expertise develops, skill accumulation slows – and the worker who appears more productive today may become less capable tomorrow. A technology can raise output while degrading the state variable – human capital – that governs future productivity. Short-run productivity gains can coexist with lifetime welfare losses.

This paper develops a framework for understanding when and why short-run productivity estimates diverge from long-run welfare. We model workers who learn by doing and can delegate tasks to AI. The key parameter is *pedagogical quality*, denoted μ: the degree to which AI-assisted work contributes to skill formation relative to unassisted work. Autocomplete interfaces that minimize user effort correspond to low μ; Socratic tutors that prompt reflection correspond to high μ. The central result is that when $\mu \neq 1$, standard empirical designs *condition on an endogenous state* – the worker's current skill – rather than recovering the welfare-relevant counterfactual skill path. When $\mu < 1$, this overstates AI's contribution; when $\mu > 1$, it understates long-run benefits. The measurement problem is general; the direction of bias is parameter-dependent.

The logic extends beyond productivity to verification. A mathematician reviewing a proof builds deeper understanding by attempting each step before reading the solution; passive review misses errors that active reconstruction catches. When AI generates code, *no one* possesses the authorial understanding that comes from writing it – creating vulnerability to subtle bugs that can propagate through codebases and into training data for future AI systems.

Our analysis identifies two structural biases in productivity estimates. The first, *spillover bias*, grows with industry-level AI saturation. As adoption spreads, non-users face degraded learning environments: reduced mentorship from seniors who delegate teaching moments to AI, weaker peer effects as colleagues accumulate less shareable knowledge, and curricula redesigned for AI-assisted workflows. Comparing AI users to these degraded non-users overstates the benefits of adoption.

The second bias, *state-path divergence*, operates at the individual level. A worker who has relied on AI for years has lower skill than they would have developed without it. Measuring AI's value against their current, atrophied skill overstates gains; the *welfare-relevant comparison* is to the skill they would have had absent AI. As skills atrophy, AI appears increasingly indispensable – even holding AI's capabilities fixed – because the outside option

has deteriorated.

A mechanism distinctive to generative AI interacts with these biases. Unlike calculators or spreadsheets, which operate via fixed algorithms, generative AI learns from human-generated content. When workers delegate tasks, they produce less original content, and what they produce reflects less skill. Both effects degrade training data for future AI systems. This creates a feedback loop that partially stabilizes human capital: as skills atrophy, AI quality degrades, which reduces adoption incentives and attenuates further skill loss. But the feedback also creates a novel externality: each firm's adoption degrades AI quality for all users, yet atomistic firms ignore this aggregate effect. Critically, while the feedback loop stabilizes levels, it exhibits a pronounced asymmetry: degradation is fast but recovery is slow. AI systems can be retrained in months, but rebuilding human expertise takes years. A temporary period of overadoption can therefore push the economy into a low-skill regime from which escape is prohibitively slow – making early intervention to prevent skill loss far more effective than later attempts to reverse it.

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

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

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

Early evidence beyond the calibration sample favors $\mu < 1$. METR (2025) find experienced developers are slower with AI tools yet believe AI increases their productivity – consistent with skill atrophy impairing self-assessment. Budzyń et al. (2025) document endoscopist deskilling: after just three months of AI-assisted colonoscopy, physicians' unassisted adenoma detection rates fell from 28.4% to 22.4% – a 21% relative decline – providing the first clinical evidence of AI-induced skill atrophy affecting patient outcomes. del Rio-Chanona et al. (2024) find Stack Overflow activity declined sharply after ChatGPT's release; Burtch et al. (2024) show newer users exited fastest – consistent with failing to build query-formulation skills, harder to explain by substitution alone. These patterns are inconsistent

with $\mu \geq 1$.

Longitudinal evidence from related technologies suggests these effects compound over time. Dahmani and Bohbot (2020) track drivers over three years and find GPS use predicts steeper decline in hippocampal-dependent spatial memory, with moderate-to-large effect sizes ($r = -0.52$ to -0.68 for cognitive mapping and landmark encoding). Crucially, those who used GPS more did *not* report worse sense of direction at baseline – ruling out reverse causality and supporting a causal interpretation that technology use degrades the skill it replaces. Casner et al. (2014) find similar patterns for pilots: motor skills remain intact with automation, but cognitive skills – precisely those needed when automation fails – degrade with heavy autopilot use. If labor-saving technology produces measurable skill atrophy for spatial navigation and manual flying, the mechanism should operate even more powerfully for generative AI, which targets higher-order cognitive tasks central to professional expertise and operates across virtually all knowledge work simultaneously.

More broadly, our analysis highlights a general limitation of performance measurement when current actions reshape the state variables that determine future productivity. A technology can appear *increasingly indispensable* even when it is not improving, because past use has degraded the alternative against which it is evaluated. The welfare-relevant counterfactual is not the worker's current state without the technology, but the *skill path that would have obtained* absent adoption. Standard productivity measurement conflates these objects; when technology affects skill formation, the conflation can reverse the sign of measured effects. Generative AI is the leading contemporary example, but the theoretical structure applies wherever learning-by-doing meets labor-saving technology.

To sum up, we make three contributions. First and foremost, we show that when AI affects skill formation ($\mu \neq 1$), standard productivity estimates *condition on an endogenous state* – overstating AI's value when $\mu < 1$ and understating it when $\mu > 1$. This measurement result is general and does not depend on whether AI ultimately raises or lowers long-run productivity. Second, we derive predictions for wages and inequality that emerge when $\mu < 1$: pre-AI cohorts command growing premiums, high-ability workers lose most in the long run, and inequality follows a hump-shaped path. Third, we characterize optimal policy when human capital generates spillovers, showing that the welfare-maximizing adoption level can differ from the privately optimal level even when AI unambiguously raises productivity.

The paper proceeds as follows. The remainder of this section reviews related literature. Section 2 develops the model. Section 3 characterizes equilibrium. Section 4 analyzes mismeasurement, cohort effects, and quantification. Section 5 examines welfare and policy. Section 6 concludes.

1.1 Related Literature

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

AI augments rather than replaces workers, emphasizing how human capital – including AI expertise and higher-order skills – mediates the productivity and distributional effects of AI; our framework complements theirs by showing that even augmentation can degrade the human capital stock when it substitutes for learning. Eloundou et al. (2024) estimate 80% of U.S. workers could have at least 10% of tasks affected by LLMs; Acemoglu (2024) estimates TFP gains of 0.5–0.7% over ten years – both assuming no skill atrophy. Agrawal et al. (2018, 2019) emphasize complementarities between AI prediction and human judgment; our framework identifies a tension – 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, and Dell'Acqua et al. (2023) identifying a "jagged frontier" where AI helps on some tasks but hurts on others. Otis et al. (2023) conduct a field experiment with Kenyan entrepreneurs and find heterogeneous effects: AI mentorship increased performance by 20% for high performers but *decreased* it by 10% for low performers, with the divergence driven by task selection – low performers sought help on more challenging problems. Handa et al. (2025) analyze millions of AI conversations to measure usage patterns across occupations, finding 57% of usage suggests augmentation while 43% suggests automation – but these classifications treat skill as fixed. Gaessler and Piezunka (2023) find chess computers *helped* players improve ($\mu \geq 1$) – plausibly because chess engines provide immediate, objective feedback and players actively analyze engine suggestions rather than passively accepting them – but more recent work documents deskilling: endoscopists (Budzyń et al., 2025), navigators (Ying et al., 2024), robot-assisted workers (Cho, 2024), and knowledge workers (Lee et al., 2025; Dell'Acqua, 2022). Chen et al. (2025) identify a "mediocrity trap" whereby GenAI reduces effort investment in creative tasks. Our contribution is to show that productivity and skill formation are jointly determined: measuring one without the other conflates short-run gains with long-run costs.

The welfare implications of AI extend beyond labor productivity. Luo et al. (2025) find platforms may optimally restrict AI access to preserve human capital. Ong and Png (2026) show deskilling technology can increase labor supply by providing work amenity, highlighting a potential benefit our framework does not capture. Athey and Scott Morton (2025) examine welfare effects of AI market power. Our model builds on human capital theory (Becker, 1962), learning-by-doing (Arrow, 1962; Lucas, 1988), and learning curves (Thompson, 2010).[1] We extend Arrow's insight that production generates knowledge as a byproduct to show AI can sever this link.

A growing literature examines how AI threatens training and skill transmission. Garicano and Rayo (2025) show apprenticeships become unviable when AI automates entry-level work: if juniors generate no billable output, the economic foundation of apprenticeship collapses. Ide (2025) develops a growth model where AI reduces opportunities for tacit knowledge acquisition. Beane (2019, 2024) provide evidence that robotic surgery made trainees "optional," reducing hands-on practice tenfold. Brynjolfsson et al. (2025b) document early employment effects of AI, finding heterogeneous impacts across occupations. Our contribution is distinct: we study learning *within* jobs rather than access *to* jobs. The mechanisms compound

[1] Our learning function draws on Mincer (1974). The "competency trap" from Levinthal and March (1993) is related but concerns organizational learning.

– policies preserving entry-level employment will fail if the resulting work is pedagogically hollow.

Our training data mechanism connects to the computer science literature on model collapse (Shumailov et al., 2024; Alemohammad et al., 2024). Platforms like Stack Overflow provide both training data and mentorship networks; when users exit, they reduce fresh training data *and* degrade peer-learning, compounding the inefficiencies we identify.

2 The Model

2.1 Environment and Primitives

Time is discrete, indexed by $t \in \{0, 1, 2, \ldots\}$. A unit mass of firms, indexed by $i \in [0, 1]$, each employs one worker. We use lowercase (h, $\alpha \in [0, 1]$) for individual variables and uppercase (H, A) for aggregates.

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

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

Workers face an effort constraint: total effort across all worker-performed tasks is normalized to unity. When a firm adopts AI at intensity $\alpha \in [0, 1]$, it delegates tasks in $[0, \alpha]$ to AI while the worker performs tasks in $(\alpha, 1]$. Standard optimization shows the worker spreads effort uniformly across performed tasks, yielding worker output $h(1 - \alpha)^{1-\gamma}$.[2]

Substituting, period output takes the tractable form

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

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

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

2.2 Human Capital Dynamics

Human capital evolves according to

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

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

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

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

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

Assumption 1 (Learning Capacity)**.** The function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is twice continuously differentiable, strictly positive, bounded above, with $\varphi'(h) < 0$ and $\lim_{h\to\infty} \varphi(h) = 0$. These properties capture diminishing returns: experts face smaller learning gains as most relevant knowledge has been acquired.[3]

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

The parameter μ has clear empirical content.[4] We treat μ as exogenous, though it depends on AI design, workplace norms, and user incentives. Competitive pressure exacerbates low-μ outcomes; Appendix A analyzes this.

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

Remark 1 (Heterogeneous μ). In practice, μ varies across tasks and career stages. Such heterogeneity *strengthens* our results: workers using AI during low-μ phases accumulate less skill than those using it during high-μ phases, introducing additional path dependence.[5]

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

[4] Bastani et al. (2025) show GPT-4 access harms learning ($\mu < 1$), but pedagogically-designed tutors mitigate this (higher μ). Dell'Acqua (2022) document reduced effort with AI; Brynjolfsson et al. (2025a) find AI helping workers "move down the experience curve." The human factors literature documents "automation complacency" – reduced vigilance when automation handles tasks (Parasuraman and Riley, 1997; Sarter et al., 1997). Complacency is a short-run phenomenon; skill atrophy is its long-run consequence. When users disengage, they stop practicing, and capabilities degrade.

[5] The scalar μ can be interpreted as an adoption-weighted average $\bar{\mu} = \int_0^\alpha \mu(j)g(j)dj/G(\alpha)$. If workers delegate tasks where AI excels and learning value is low, $\bar{\mu}$ falls below the task-uniform average. Lifecycle heterogeneity is particularly relevant: if novices have low μ while experts have high μ, optimal policy may restrict AI for juniors while permitting it for seniors. Appendix A.7 verifies robustness to skill-varying μ.

Table 1: Notation Guide

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

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

2.3 The Firm's Dynamic Problem

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

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

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

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

Standard results ensure V exists, is unique, and is strictly increasing and concave in h.[6] The key trade-off is dynamic: higher adoption today raises current output but – when $\mu < 1$ – reduces future human capital.

Assumption 2 (Labor Market Structure)**.** Labor markets are competitive with general human capital (portable across employers). Wages equal marginal products, so workers with lower skills earn lower wages.[7]

The Bellman equation (5) takes output $Y(h, \alpha; A)$ as the firm's flow objective, which by construction internalizes skill dynamics. This is *not* a profit-maximizing firm in a spot

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

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

market for general skills – under Becker's canonical setup, such a firm would be indifferent to workers' human capital paths since $w_t = \text{MP}_t$. Rather, the problem is best interpreted as: (i) a representative worker choosing adoption to maximize lifetime income (under competitive wages, $w_t = Y_t$ up to constants), or (ii) a firm-worker pair bound by an implicit long-term contract. Assumption 2 then serves to map h into wages for the inequality analysis (Section 4.5), not to microfound the dynamic program. The key inefficiency arises not from a wedge between firm and worker incentives over skill atrophy – which by construction is internalized – but from spillovers across agents: when firm i's adoption degrades human capital, it harms learning at other firms through reduced mentorship (Section 5). With firm-specific human capital, firms would internalize even more of the skill atrophy effect (Acemoglu and Pischke, 1999), potentially reducing overadoption. The measurement results (Propositions 1–2) hold regardless of labor market structure because the counterfactual skill path remains endogenous.

3 Equilibrium Characterization

This section characterizes equilibrium adoption and establishes preliminary results that underpin our main findings. The key insight is that AI's effect on skill formation – captured by the pedagogical quality parameter μ – fundamentally shapes both adoption decisions and long-run outcomes.

Firms balance immediate output gains against future skill costs. When $\mu < 1$, AI substitutes for learning, creating a dynamic cost that patient firms internalize. In steady state, higher adoption leads to lower skills (Lemma 2), and the economy can settle into a "trap" where output is lower than under no adoption (Proposition 4). When $\mu \geq 1$, these dynamics reverse: AI augments learning, and no trap can occur. The remainder of this section formalizes these claims; readers primarily interested in measurement implications may proceed to Section 4 after noting that skill atrophy requires $\mu < 1$.

3.1 The Role of Pedagogical Quality

The firm's adoption decision balances immediate productivity gains against dynamic skill costs. When AI is sufficiently productive, some adoption is always optimal; complete delegation is never optimal because effort concentration on remaining tasks becomes increasingly valuable.[8]

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

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

where $\bar{h}$ is the steady-state human capital without AI, and $\bar{V}' \equiv V'(\bar{h})$ is the marginal value of human capital at that steady state. The left side is the static marginal benefit of adoption

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

at $\alpha = 0$; the right side is the discounted marginal learning cost. This ensures interior adoption $\alpha^* > 0$ at the boundary $(h, \alpha) = (\bar{h}, 0)$.[9]

The following proposition characterizes how pedagogical quality shapes adoption:

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

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

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

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

The proposition captures a fundamental asymmetry. In the substitution regime ($\mu < 1$), firms face an intertemporal trade-off: higher adoption raises current output but impairs skill development. Forward-looking firms internalize this cost and adopt less than myopic firms would. Unlike prior work focusing on which tasks machines perform (Autor et al., 2003; Acemoglu and Autor, 2011), we show automation can change the *supply* of skills by altering how they accumulate.[10]

Why does the market not simply select for patience? Several corrective mechanisms fail: patient firms are competitively punished in the short run before their strategy pays off (Proposition 10); spillovers mean private returns to patience understate social returns; and the measurement problem in Section 4 causes even planners to perceive skill-preserving policies as costly. The trap persists because rationality operates on distorted signals.

3.2 Steady-State Equilibria

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

$$\delta h^* = \lambda \ell(\alpha^*)\varphi(h^*) \tag{6}$$

balances depreciation against learning, where $\ell(\alpha) \equiv 1 - (1 - \mu)\alpha$ is the effective learning rate. We impose standard regularity conditions ensuring interior, stable steady states.[11]

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

[10] This connects to the "deskilling" literature (Braverman, 1974), but our framework allows for the opposite when $\mu > 1$. The comparative static $\partial\alpha^*/\partial\mu > 0$ is local; global results require additional curvature conditions stated in Appendix B.

[11] These technical requirements appear in Appendix B. In brief, they require interior steady states, local stability, static curvature dominating dynamic terms in the FOC, and a monotone policy function. They hold for generic parameter values.

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

Remark 2 (Stability). With φ strictly decreasing, the stationarity condition admits a unique steady state. Stability is guaranteed when depreciation dominates the learning feedback: $\delta > \lambda \ell(\alpha^*)|\varphi'(h^*)|$. Since $\varphi'(h^*) < 0$, this is equivalent to $\delta + \lambda \ell(\alpha^*)\varphi'(h^*) > 0$; the appendix formalizes the regularity conditions.

The lemma establishes that AI's long-run effect on skills depends entirely on whether it substitutes for or augments learning. This yields the following characterization of steady-state equilibria:

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

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

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

This dichotomy has a sharp implication: skill atrophy *requires* $\mu < 1$. If $\mu \geq 1$, skills cannot fall below the no-adoption benchmark $\bar{h}$, and AI's direct productivity contribution ensures $Y^* > \bar{h}$. The empirical question of which regime applies is first-order for policy.

We now establish that equilibrium is unique and globally stable – essential properties for welfare analysis and comparative statics.

Lemma 4 (Uniqueness and Global Stability). *Under Assumptions 1–3 and the regularity conditions stated in Assumption 8 (Appendix B):*

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

(ii) The steady-state equilibrium is unique.

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

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

Global stability follows from the contraction property under our regularity conditions. Part (iv) follows from the policy function's slope: $d\alpha^*/dh < 0$ when $\mu < 1$. Starting from $h_0 = \bar{h} > h^*$, skills decline monotonically toward h^*; since $\alpha_t = \alpha^*(h_t)$ and the policy function is decreasing, adoption rises monotonically toward α^*.

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

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

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

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

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

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

Result (i) echoes Acemoglu and Restrepo (2018): better automation increases automation, but here it *endogenously degrades* the human capital stock. Result (ii) implies short-termism exacerbates skill atrophy. Result (iii) implies occupations where learning is central face larger stakes. Result (iv) reflects offsetting forces: higher μ directly raises h^* but indirectly lowers it by inducing more adoption.[12]

Remark 3 (Robustness). The qualitative results do not depend on specific functional forms. What matters is diminishing returns to learning, $\mu < 1$, and spillovers creating a wedge between private and social returns. Appendix A.7 verifies robustness.

4 Mismeasurement of AI Productivity

Standard productivity studies estimate the causal effect of AI on output holding current skill fixed. We show this diverges from welfare-relevant comparisons when skill is endogenous to past AI use, and this divergence can reverse sign. The critique is not of empirical methods but of the welfare question those methods implicitly answer.

4.1 Spillover Bias

The choice of counterfactual fundamentally determines whether AI adoption appears beneficial or harmful.

Definition 1 (Alternative 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 $h_0 = \bar{h}$.

The cross-sectional counterfactual is the comparison made by most empirical studies, including RCTs that randomize AI access. The long-run counterfactual captures the welfare-relevant question. These diverge when aggregate AI adoption affects learning opportunities for non-users – through reduced mentorship, weaker knowledge spillovers, or degraded training institutions.

Assumption 4 (Learning Spillovers)**.** Individual learning depends on aggregate human capital: $L_i = [(1-\alpha_i) + \mu\alpha_i] \cdot \varphi(h_i) \cdot \psi(H)$, where $\psi(H) = (H/\bar{H})^\eta$ and $\eta \geq 0$ governs spillover intensity.

Evidence on learning spillovers comes from peer effects in education (Sacerdote, 2001) and coworker effects in firms (Mas and Moretti, 2009), with estimated elasticities in the range 0.05–0.2. We use $\eta = 0.15$ as our baseline.

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

[12] Appendix A shows impatient firms gain market share in the short run, potentially driving out patient firms before their restraint pays off.

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

4.2 State-Path Divergence

Spillover bias concerns how AI adoption by some workers degrades the counterfactual for others. A second bias operates even at the individual level: *path dependence in human capital* causes state-conditional productivity gains to diverge from welfare-relevant path comparisons.

Definition 2 (State-Conditional vs. Path Counterfactuals)**.** The *state-conditional counterfactual* holds human capital fixed:

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

The *path counterfactual* compares lifetime output under adoption versus the no-adoption path:

$$\Delta^{PATH}(\tilde{\beta}) = \sum_{t=0}^{\infty} \tilde{\beta}^t \left[Y(h_t^U, \alpha_t) - Y(h_t^{NA}, 0) \right],$$

where $\tilde{\beta}$ is the evaluator's discount factor.

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

We emphasize that this measurement problem does not depend on disagreement about discount factors. Even when $\tilde{\beta} = \beta$, the state-conditional gain Δ_t^{SC} overstates AI's welfare contribution because it *conditions on an endogenous state* – current skill h_t^U – rather than the counterfactual skill h_t^{NA}. The firm's choice is privately optimal given h_t, but Δ_t^{SC} conflates "valuable given current skill" with "valuable relative to never adopting."

Most empirical implementations estimate AI's contribution holding current worker state fixed – explicitly via controls for experience, tenure, or skill proxies, or implicitly by comparing the same worker before and after adoption. These designs recover Δ^{SC}: the effect of turning AI "on" at a given skill level. When the treatment changes the state variable, however, the welfare-relevant object is the total effect along the counterfactual path.

Proposition 2 (State-Path Divergence)**.** *Suppose $\mu < 1$. Then:*

(i) The ratio Δ_t^{SC}/h_t^U is strictly increasing in t.

[13] Throughout the main analysis, we maintain $\tilde{\beta} = \beta$ to isolate the role of spillover externalities. The case $\tilde{\beta} > \beta$ raises a distinct question – whether a paternalistic evaluator should override firms' time preferences – that we set aside. Our welfare results concern externalities, not paternalism.

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

Unlike spillover bias, this result requires no cross-agent externalities – it operates at the individual level through path dependence in human capital. As skills atrophy toward $h^* < \bar{h}$, the worker's AI-independent productivity falls, inflating the measured value of AI in state-conditional comparisons.[14]

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

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

The two biases differ in structure and remedy. Spillover bias is a cross-sectional externality calling for Pigouvian correction. State-path divergence is a longitudinal measurement error calling for counterfactual-aware research designs. Cohort comparisons and cross-country variation in adoption timing approximate the correct counterfactual more closely than state-conditional designs.

4.3 The Skill-Data Feedback Loop

The preceding analysis took AI quality as fixed. We now introduce a mechanism distinctive to generative AI: because these systems learn from human-generated content, widespread adoption can degrade the data on which future AI systems train.

The distinction from previous automation is stark. A calculator does not need humans to know arithmetic to compute 937×48; its accuracy is invariant to user skill. GPS navigation works identically whether or not drivers remember local streets. These technologies operate via fixed algorithms that neither learn from nor degrade with human practice. Generative AI is different: it learns from human output. If workers delegate tasks, they produce less original content, and the content they produce reflects diminished expertise. Both effects degrade training data.

We model AI productivity as evolving according to

$$A_{t+1} = (1 - \zeta)A_t + \zeta \cdot Q(H_t, \bar{\alpha}_t) \tag{7}$$

where $\zeta \in (0, 1)$ governs how quickly AI quality adjusts, $\bar{\alpha}_t$ is average adoption intensity, and $Q : \mathbb{R}_+ \times [0, 1] \to \mathbb{R}_+$ satisfies $\partial Q/\partial H > 0$ and $\partial Q/\partial \bar{\alpha} < 0$. The dependence on H captures that higher-skilled humans generate higher-quality training data; the dependence on $\bar{\alpha}$ captures that AI-generated content dilutes the human signal. This specification builds

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

on the computer science literature documenting "model collapse": recursive training on AI-generated content causes distributional tails to disappear, yielding increasingly homogeneous outputs (Shumailov et al., 2024).[15]

The law of motion implies $\partial A_{t+1}/\partial H_t > 0$ (human capital improves training data) and $\partial A_{t+1}/\partial \bar{\alpha}_t < 0$ (adoption degrades it). This creates a feedback loop with distinct effects on levels versus dynamics.

To characterize the joint dynamics, we now work with aggregate variables (uppercase H and $\bar{\alpha}$, which equal their individual counterparts in the representative-agent economy). The two-dimensional system is:

$$H_{t+1} = (1-\delta)H_t + \lambda \ell(\alpha^*(H_t, A_t))\varphi(H_t) \tag{8}$$

$$A_{t+1} = (1-\zeta)A_t + \zeta Q(H_t, \bar{\alpha}_t) \tag{9}$$

where $\alpha^*(H, A)$ is the optimal adoption policy. Stability of the joint system requires both eigenvalues of the Jacobian to lie inside the unit circle. The key observation is that $\partial H'/\partial A < 0$ (higher AI quality induces more adoption, reducing learning) while $\partial A'/\partial H > 0$ (higher human capital improves training data). This negative feedback is the stabilizing force.

Corollary 3 (Feedback Loop)**.** *With endogenous AI quality, let $H^*(A_0)$ denote steady-state skill when AI quality is exogenously fixed at A_0, and let (H^{**}, A^{**}) denote the joint steady state with endogenous AI quality. Then $H^{**} > H^*(A_0)$: endogenous AI quality raises steady-state human capital relative to the exogenous-A benchmark.*

The intuition follows from Corollary 1: $\partial \alpha^*/\partial A > 0$ and $\partial h^*/\partial A < 0$. When AI quality falls, firms adopt less, preserving more learning-by-doing. The feedback is self-correcting at the aggregate level, even though individual firms ignore their contribution to AI quality degradation. We formalize this training data externality in Section 5.

However, the stabilization masks a critical asymmetry: degradation is fast but recovery is slow. Corollary 3 concerns *steady-state levels* – the feedback loop raises where the economy settles. Proposition 3 below concerns *transition dynamics* – how long it takes to get there. The feedback stabilizes levels but cannot accelerate recovery, because human capital is the bottleneck.

Proposition 3 (Slow Recovery)**.** *Let $\tau_A \equiv 1/\zeta$ denote the AI retraining timescale and $\tau_H \equiv 1/\lambda$ the human learning timescale. Suppose a temporary shock increases adoption above α^{**} for duration T, after which adoption returns to α^{**}. Then:*

(i) Human capital and AI quality both decline during the shock.

(ii) After the shock, recovery to the original steady state (H^{**}, A^{**}) requires time T_R satisfying $T_R \geq c \cdot \tau_H$ for some $c > 0$, regardless of how fast AI can retrain.

(iii) When $\tau_H \gg \tau_A$ (human learning is slower than AI retraining), $T_R \gg T$: temporary shocks produce persistent effects.

[15] Follow-up work confirms this across settings: Alemohammad et al. (2024) document "Model Autophagy Disorder" in self-consuming generative models; Dohmatob et al. (2024) formalize degradation through scaling laws; Gerstgrasser et al. (2024) show collapse is avoidable when synthetic data *accumulates* alongside real data rather than replacing it.

The key insight is that AI quality depends on human capital through the training data channel: $A = Q(H, \alpha)$. Even if AI systems can be retrained rapidly (small τ_A), they cannot recover quality until the human capital stock recovers – and human capital accumulates slowly through learning-by-doing (large τ_H). The feedback loop creates a bottleneck: AI recovery waits on human recovery.

The asymmetry between degradation and recovery has practical consequences. A temporary period of overadoption can push the economy into a low-skill regime from which escape takes a generation (cf. David, 1985, on technological lock-in). Early intervention to prevent skill loss is far more effective than later attempts to reverse it.

The feedback loop also implies a natural deceleration in AI capability growth. Early in AI diffusion, when human skills remain high and AI-generated content is scarce, training data quality is high and AI improves rapidly. As adoption spreads and skills atrophy, training data quality degrades, slowing AI improvement – not from technical limits, but from the erosion of the human capital that feeds it.

Evidence is consistent with early-stage feedback effects. Stack Overflow activity declined 25% within six months of ChatGPT's release (del Rio-Chanona et al., 2024), with newer users most likely to exit (Burtch et al., 2024). Dell'Acqua (2022) document that workers using AI invest less cognitive effort – producing content that, if used for training, transmits less expertise.

4.4 When Bias Reverses Sign: The Skill Trap

The mismeasurement biases identified above operate whenever $\mu < 1$. We noted in Section 3 that steady-state output can fall below no-adoption levels when AI substitutes for learning. This section characterizes those instances by identifying necessary and sufficient conditions for what we call the *skill trap*: a scenario in which AI appears beneficial in cross-sectional comparisons while actually reducing long-run output.

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

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

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

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

Condition (T2) concerns productivity *levels*, not growth rates: the trap means AI users eventually produce less than they would have produced without AI. The trap is individually rational: firms optimize at every date, yet the equilibrium path delivers lower long-run output than no adoption.[16]

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

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

Proposition 4 (Existence of the Skill Trap). *Under Assumption 3 with initial condition $h_0 \leq \bar{h}$, the economy is in a skill trap if and only if:*

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

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

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

The trap requires *all three* conditions. Condition (ii) may not hold for highly capable AI – if $A \cdot G(1) > \bar{h}$, even complete skill atrophy leaves output above the no-adoption benchmark. When this holds, AI adoption is unambiguously beneficial: skills become obsolete but output remains high. The trap is thus a transitional phenomenon, applying when AI is productive enough to attract adoption but not yet capable enough to fully substitute for human expertise. With AI productivity A_t growing over time, condition (ii) eventually fails if $A_t \to \infty$, dissolving the trap. However, the mismeasurement biases (Propositions 1–2) operate under the weaker condition $\mu < 1$ alone, regardless of AI capability or firm patience. The trap clarifies when bias reverses sign; the biases themselves are general.

Corollary 4 (Sign Reversal). *When the economy is in a skill trap, the measured effect has the wrong sign:* $\Delta_t^{CS} > 0 > \Delta_t^{LR}$ *for t sufficiently large.*

In the skill trap, $\Delta^{LR} < 0$ follows directly from $Y^* < \bar{h}$. What spillovers provide is $\Delta^{CS} > 0$ despite this: learning spillovers degrade non-users' skills so that $h^{NU*} < \bar{h}$, allowing $Y^* > h^{NU*}$ even when $Y^* < \bar{h}$. Cross-sectional gains can coexist with long-run losses.

4.5 Cohort Effects and Wage Dynamics

The mismeasurement problems identified above have implications for the distribution of gains from AI across workers and over time. This section embeds our framework in a labor market where wages equal marginal products, generating predictions about how AI reshapes wages across ability levels, cohorts, and aggregate inequality.

A growing empirical literature documents that AI disproportionately benefits less-skilled workers in the short run. Brynjolfsson et al. (2025a) find productivity gains of 14% overall but exceeding 30% for novices in customer service; Noy and Zhang (2023) find larger effects for less experienced writers; Peng et al. (2023) document similar patterns for coding. This "democratization" has prompted optimism about reducing inequality (Autor, 2024). Our framework suggests a more complex dynamic: the short-run compression may reverse as skill atrophy accumulates.

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

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

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

Part (i) says high-ability workers bear the largest long-run costs – precisely those who benefit least from AI in short-run studies. These workers lose twice: in the short run, AI compresses their productivity advantage by disproportionately helping their less-skilled peers; in the long run, AI impedes their skill development, preventing them from reaching their full potential. We call this *ability reversal* because short-run and long-run effects have opposite signs for high-ability workers – they appear to benefit least in experiments but lose most over careers. This creates a political economy challenge: early AI adoption generates enthusiasm because those who benefit most visibly (low-ability workers gaining immediate productivity) are not those who bear the largest long-run costs.

Part (ii) says pre-AI cohorts become increasingly valuable. Early evidence is consistent with vintage effects: Beane (2019) documents that robotic surgery reduced trainee hands-on experience tenfold, with senior surgeons becoming increasingly valuable for complex cases requiring manual dexterity.

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.

Proposition 6 (Hump-Shaped Inequality)**.** *Suppose* $\mu < 1$ *and pre-AI cohorts retire at rate* $\nu > 0$*. Then wage variance* σ_t^2 *follows a hump-shaped path:*

(i) $\sigma_0^2 = 0$*: initially all workers are identical (pre-AI steady state).*

(ii) $d\sigma_t^2/dt > 0$ *for small* t*: as post-AI workers' skills diverge from pre-AI workers', variance rises.*

(iii) σ_t^2 *peaks at some* T^{max} *then declines as pre-AI cohorts retire entirely.*

Remark 5 (Reconciling Propositions 5 and 6). The vintage premium π_t tracks the bilateral gap between pre-AI and post-AI cohorts (monotonically increasing). The hump shape in σ_t^2 reflects *aggregate* inequality dynamics as cohort composition shifts: variance rises with the skill gap but falls as the high-skill cohort shrinks.

The hump shape implies that inequality peaks at an intermediate horizon – around year 25 under baseline parameters – then declines as pre-AI workers retire.

4.6 Quantifying the Bias

This section calibrates the model to experimental evidence to gauge the potential magnitude of mismeasurement. The exercise is illustrative; it demonstrates that bias can be economically meaningful under parameters anchored to experimental evidence.

The parameter μ is the key unknown. Direct identification requires panel data tracking AI usage and subsequent skill assessments – a demanding requirement that only recent

experiments satisfy. We treat μ as uncertain and report results across $\mu \in [0.3, 0.9]$, reflecting heterogeneity across AI-assisted tasks: autocomplete coding interfaces likely have lower μ than Socratic tutoring systems.

Bastani et al. (2025) find GPT-4 access reduces subsequent math test performance by 17%, implying $\mu \approx 0.83$. Shen and Tamkin (2026) find a nearly identical 17% reduction among software developers learning a new Python library – a different population, task domain, and research team, yet the same point estimate.[17] The convergence suggests $\mu \approx 0.83$ may be a robust central estimate for unrestricted AI assistance.

Other evidence spans a wide range: Dell'Acqua (2022) document reduced cognitive effort with AI ($\mu < 0.5$); Budzyń et al. (2025) find endoscopist deskilling ($\mu \approx 0.6$–0.8); Gaessler and Piezunka (2023) find chess engines accelerated skill development ($\mu > 1$) – plausibly because chess feedback is immediate and unambiguous, with every game ending in a clear outcome and the engine providing continuous, objective evaluation. This contrasts with most knowledge work, where code that runs may still contain subtle bugs and a legal brief that reads well may contain flawed reasoning. The heterogeneity underscores that μ is context-dependent; we report results for multiple values rather than defending a single estimate.

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

Table 2: Outcomes by Pedagogical Quality μ

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

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

Qualitative conclusions are robust: bias is positive and economically meaningful for all $\mu < 1$, ranging from 2% at year 10 when $\mu = 0.9$ to 16% when $\mu = 0.3$.

Figure 1 plots measurement bias over time. At $\mu = 0.5$, bias exceeds 11% by year 10 and 17% by year 20; at $\mu = 0.3$, it reaches 16% and 25% respectively. When $\mu = 1$, no bias arises.

Figure 2 shows transition dynamics under different parameterizations. Panel (a) plots skill paths for varying μ: lower pedagogical quality leads to faster convergence to a lower steady state. Panel (b) shows how the vintage premium $\pi_t = \bar{h}/h_t$ evolves – the wage

[17] Both experiments study novices learning new material. Whether the convergent estimate reflects a general property of AI-assisted learning or something specific to novice acquisition remains open.

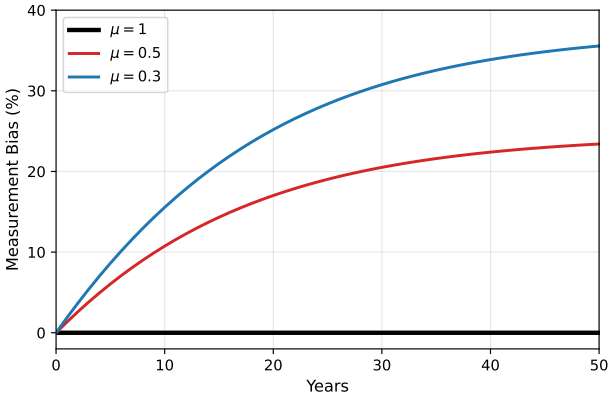

Figure 1: Measurement Bias Over Time

Note: Bias $= (\Delta_t^{SC} - \Delta_t^{LR})/\Delta_t^{LR}$. Parameters: $\delta = 0.05$, $\lambda = 0.15$, $\alpha = 0.5$, $\eta = 0.15$.

advantage of pre-AI cohorts grows monotonically as AI-era workers' skills atrophy. Panel (c) illustrates the hump-shaped inequality dynamics: wage variance rises as skill gaps widen, peaks around year 25, then falls as pre-AI cohorts retire.

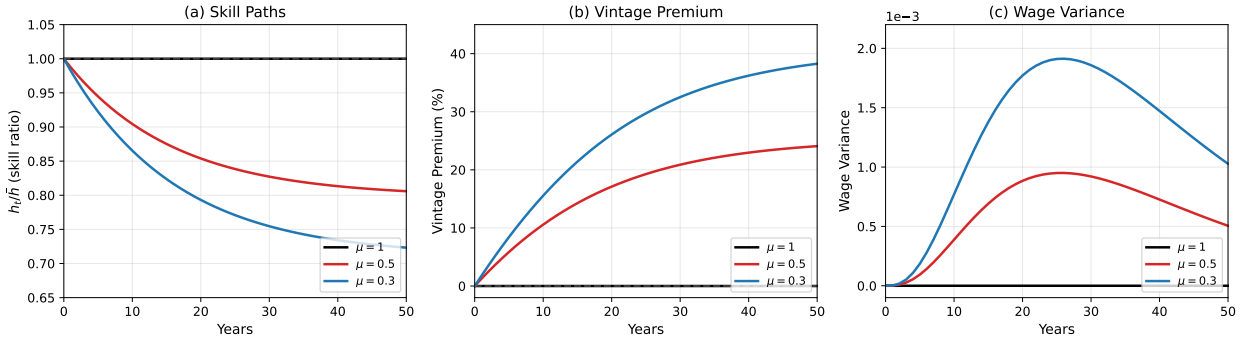

Figure 2: Transition Dynamics Under Skill Atrophy

Note: All panels show dynamics for $\mu \in \{1, 0.5, 0.3\}$ over 50 years. Parameters: $\delta = 0.05$, $\lambda = 0.15$, $\alpha = 0.5$, $\nu = 0.05$ (retirement rate). Panel (c) shows wage variance normalized by peak value; variance is hump-shaped, peaking around year 25.

Sensitivity analysis. Table 3 reports bias magnitudes under alternative parameterizations. The bias is increasing in adoption intensity α (more delegation means more forgone learning), decreasing in μ (lower pedagogical quality means faster atrophy), and increasing in learning intensity λ (when learning-by-doing matters more, its disruption is costlier). The bias is relatively insensitive to δ within plausible ranges, because depreciation affects both adoption and no-adoption paths similarly.

Wage and inequality implications. The skill gap translates into wage differentials under competitive labor markets. With wages proportional to marginal product, $w(h) \propto h$ in our baseline specification. A worker whose skill falls 20% below the no-adoption counterfactual ($h^* = 0.80\bar{h}$) earns 20% lower wages in steady state.

Table 3: Sensitivity of Measurement Bias to Parameter Values

	Bias at Year 10				
Parameter varied	Low	Med-Low	Baseline	Med-High	High
μ (1.0, 0.9, 0.7, 0.5, 0.3)	0.0%	2.0%	6.2%	10.7%	15.5%
α (0.3, 0.4, 0.5, 0.6, 0.7)	8.9%	9.7%	10.7%	12.0%	13.6%
λ (0.10, 0.125, 0.15, 0.175, 0.20)	6.9%	8.8%	10.7%	12.8%	15.1%
δ (0.03, 0.04, 0.05, 0.06, 0.07)	13.0%	11.6%	10.7%	10.1%	9.6%

Note: Each row varies one parameter while holding others at baseline values ($\mu = 0.5$, $\alpha = 0.5$, $\lambda = 0.15$, $\delta = 0.05$). Bias defined as the percentage by which state-conditional measurement overstates AI's welfare contribution relative to the path counterfactual: $(\Delta_t^{SC} - \Delta_t^{LR})/\Delta_t^{LR}$.

The vintage premium for pre-AI cohorts can be substantial. Let $\pi_t = w(\bar{h})/w(h_t^*)$ denote the ratio of pre-AI to post-AI wages. At year 10, $\pi_{10} \approx 1.106$ (a 10.6% premium); at year 20, $\pi_{20} \approx 1.171$ (a 17.1% premium). As pre-AI cohorts retire and post-AI workers converge to h^*, this premium grows to 25% in steady state. Under imperfect substitution between skill types, the premium can be amplified further: if pre-AI workers perform tasks that post-AI workers cannot, scarcity rents emerge.

For aggregate inequality, the hump-shaped pattern follows from the cohort dynamics. Initially, all workers have identical skills ($\sigma_0^2 = 0$). As AI adoption proceeds, post-AI workers' skills fall while pre-AI workers maintain $\bar{h}$, widening the gap and increasing variance. But as pre-AI cohorts retire, the workforce becomes increasingly homogeneous (at the lower skill level), and variance declines. The peak occurs around year 25 under baseline parameters, when the skill gap is large but pre-AI workers still constitute a substantial fraction of the workforce.

These dynamics complicate policy evaluation. A policymaker observing falling inequality in the first decade of AI adoption might conclude that AI is reducing skill gaps – the "democratization" narrative (Autor, 2024). But this compression is temporary, driven by the erosion of high-skill workers' advantages rather than the elevation of low-skill workers' capabilities. The long-run effect is a workforce with uniformly lower skills, punctuated by a shrinking cohort of pre-AI veterans commanding scarcity premiums.

4.7 Implications for Empirical Research

Our analysis has direct implications for how AI's productivity effects should be measured. Table 4 summarizes which results require which assumptions; Table 5 maps empirical strategies to their bias exposure.

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

Our analysis predicts that effect sizes should decline in longer panels, with faster decline

Table 4: Logical Dependence of Main Results

	$\mu < 1$	Spillovers	Feedback
Spillover bias (Prop. 1)	Yes	Yes	No
State-path divergence (Prop. 2)	Yes	No	No
Feedback stabilization (Cor. 3)	Yes	No	Yes
Skill trap (Prop. 4)	Yes	No	No

Table 5: Empirical Designs and Bias Exposure

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

where learning-by-doing is central. Cross-sectional estimates should systematically exceed within-worker panel estimates from the same setting. These predictions are testable as longitudinal data accumulate.

Data requirements for unbiased long-run estimation are demanding: direct assessments of human capital tracked over time (not just output), longitudinal records of AI usage intensity, measures of mentorship exposure and training environment quality, cohort identifiers relative to AI diffusion, and indicators distinguishing "autocomplete" from "tutor" AI interfaces.

5 Welfare and Policy

5.1 Sources of Inefficiency

Whether measurement biases correspond to welfare losses depends on whether decentralized adoption is efficient. A firm that recognizes AI will degrade its workers' future productivity can optimize intertemporally, trading current output against future human capital. If it bears the full cost of skill atrophy, the decentralized equilibrium is constrained efficient despite mismeasurement. Welfare loss requires an externality: human capital must have social value beyond what the adopting firm captures.

We augment the baseline model to allow learning to depend on aggregate human capital through $\psi(H)$, capturing mentorship and peer effects. The microfoundation (Appendix A.5) derives ψ from a matching model: workers who cannot solve a problem independently seek help from colleagues, and the probability of finding a capable mentor depends on the skill distribution. When aggregate human capital falls, mentorship becomes scarcer and all workers' learning suffers – including those at firms that did not adopt AI.

A social planner maximizes aggregate welfare $W = \sum_{t=0}^{\infty} \tilde{\beta}^t \int_0^1 Y(h_{i,t}, \alpha_{i,t}; A_t)\, di$, internalizing how adoption affects skill dynamics and training data feedback. Let α^D denote decentralized adoption and α^S the social optimum.

Proposition 7 (Human Capital Externality)**.** *With exogenous AI quality and common discounting ($\beta = \tilde{\beta}$), overadoption ($\alpha^D > \alpha^S$) occurs if and only if $\psi'(H) > 0$.*

The "if and only if" matters. Without spillovers, firms bear the cost of their workers' skill loss through lower future output. Spillovers break this logic: adoption imposes costs on other firms' workers that the adopting firm does not internalize. The overadoption result echoes a classic theme: when training generates positive externalities, decentralized investment is inefficiently low (Becker, 1962; Acemoglu and Pischke, 1999). Our contribution inverts this logic – here the externality arises from overinvestment in a technology that degrades training as a byproduct.

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

Proposition 8 (Training Data Externality)**.** *With endogenous AI quality $A_{t+1} = (1-\zeta)A_t + \zeta Q(H_t, \bar{\alpha}_t)$ where $\partial Q/\partial H > 0$ and $\partial Q/\partial \bar{\alpha} < 0$: individual adoption degrades future AI quality, generating overadoption. With both externalities present, total welfare loss exceeds the sum of individual effects.*

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

The magnitude of overadoption depends on the strength of spillovers. When $\psi'(H) = 0$ (no spillovers), decentralized adoption is constrained efficient. As spillovers strengthen, the wedge $\alpha^D - \alpha^S$ widens. At spillover elasticity $\eta = 0.15$, overadoption is approximately 8% of the efficient level.

5.2 Policy Responses

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

Proposition 9 (Training Mandates)**.** *Consider a training mandate $\rho \in [0, 1]$ requiring at least fraction ρ of work be performed without AI, constraining adoption to $\alpha \leq 1 - \rho$:*

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

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

Training mandates exist where skill maintenance is safety-critical. The FAA recommends pilots manually fly "at least periodically, the entire departure and arrival phases"; Casner et al. (2014) find cognitive skills for manual flying degrade with heavy automation. Medical residency programs mandate minimum procedure volumes without robotic assistance; Beane (2019) documents that residents in robot-heavy programs develop weaker unassisted skills. Luo et al. (2025) show that platforms may optimally restrict AI access to preserve human capital on their user base – a market-based analog to regulatory mandates.

The optimal mandate $\rho^* = 1-\alpha^S$ varies systematically with model parameters:

Corollary 5 (Comparative Statics of the Optimal Mandate)**.** *Let $\rho^* = 1-\alpha^S$ denote the first-best training mandate. Then:*

(i) $\partial\rho^*/\partial\eta > 0$: stronger spillovers require more restrictive mandates.

(ii) $\partial\rho^*/\partial\mu < 0$ for $\mu < 1$: lower pedagogical quality increases the optimal mandate. When $\mu \geq 1$, $\rho^* = 0$.

(iii) $\partial\rho^*/\partial\beta < 0$: more patient firms self-restrain, reducing the need for policy intervention.

The intuition is straightforward. Higher spillover intensity (larger η) means each unit of skill loss imposes larger costs on others, requiring more restriction. Lower μ means more harm per unit of AI use, warranting stronger mandates. Higher β means firms internalize more of the future skill cost, so the gap between private and social optima narrows. When $\mu \geq 1$, AI augments rather than substitutes for learning, and no mandate is needed.

Corollary 6 (AI Design)**.** *Let μ denote the pedagogical quality of AI. Compare Autocomplete design ($\mu = \mu_L < 1$) with Socratic design ($\mu = \mu_H \geq 1$):*

(i) Steady-state human capital is higher under Socratic design: $h^*(\mu_H) > h^*(\mu_L)$.

(ii) The welfare gain from raising μ exceeds the gain from an equivalent reduction in α.

(iii) Commercial incentives favor Autocomplete when users are myopic or do not internalize spillovers.[18]

Why do commercial incentives favor low-μ designs? Users selecting AI tools observe immediate productivity gains but not long-run skill effects. A tool that maximizes short-run output (Autocomplete) will outcompete one that preserves learning (Socratic) in market share, even if the latter generates higher lifetime welfare. This is analogous to the preference

[18]Within the formal model, a myopic user ($\beta = 0$) is indifferent over μ since μ affects only future h, not current Y; any forward-looking user strictly prefers higher μ. The market failure arises formally from myopia (underweighting future human capital) and spillover externalities. An informal behavioral channel – that users prefer Autocomplete because it "feels easier" or minimizes effort – is plausible but operates outside the formal framework, which includes no disutility-of-effort term.

for palatable over nutritious food: immediate utility dominates long-run health. The market failure is compounded when users are employees rather than residual claimants – they bear skill atrophy costs through lower future wages, but firms capture productivity gains. Misaligned incentives push adoption toward low-μ tools.

Training mandates and design policy are complements, not substitutes. Mandates address the *quantity* of AI use; design policy addresses its *quality*. The welfare gain from combining a modest mandate ($\rho = 0.2$) with improved design (μ: $0.5 \to 0.7$) exceeds the gain from either intervention alone. This complementarity suggests that policy should target both margins: restrict AI use in pedagogically critical settings while incentivizing Socratic AI design elsewhere.

Implementation faces practical challenges. Mandates require monitoring AI use, which may be difficult when AI is embedded in standard tools. Design regulation requires defining and measuring μ, which varies by task and user. Importantly, μ is not solely a property of the AI system – it reflects the human-AI interaction. The same underlying model can yield low μ for passive users who accept outputs uncritically and high μ for active users who engage with AI suggestions as a learning opportunity (as in the chess example, where engines produce $\mu > 1$ because players actively analyze engine recommendations). This complicates policy: regulating AI "design" may be less effective than shaping usage norms and incentives. Subsidies for high-μ AI development may be more feasible: governments could fund research into pedagogically-aware AI systems that encourage active engagement, or procurement rules could favor tools that preserve learning. Professional licensing bodies – already responsible for ensuring practitioner competence – could certify AI tools for use in training contexts.

Evidence supports the design channel. The experimental results in Section 4.6 demonstrate that interface design, not underlying capability, determines μ. Welfare-maximizing AI would function like training wheels: substantial assistance to novices, gradually withdrawing as competence develops.

Figure 3 illustrates. Panel (a) shows welfare as a function of adoption; the gap between α^D and α^S reflects overadoption. Panel (b) compares welfare paths: high-μ design dominates in the long run because it preserves skills without restricting adoption; training mandates trade off short-run output for skill preservation.

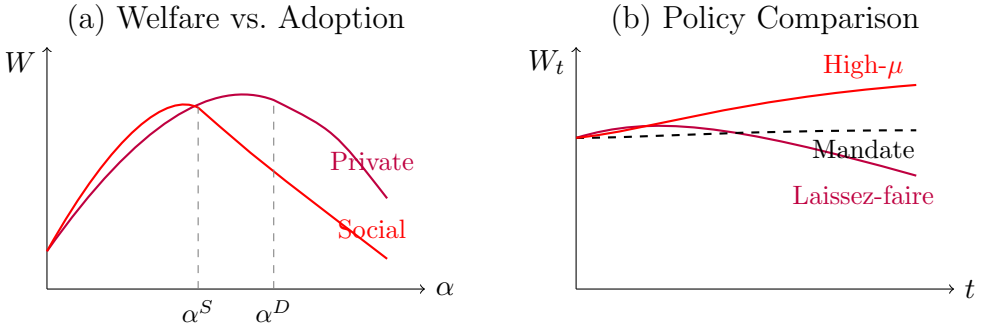

Figure 3: Welfare and Policy

Note: Panel (a): welfare as function of adoption; gap reflects externality. Panel (b): welfare paths under laissez-faire, training mandates ($\rho = 0.3$), and high-μ AI design.

6 Conclusion

This paper identifies two structural sources of mismeasurement in AI productivity studies. Spillover bias arises because non-users face degraded learning environments. State-path divergence arises because current skill reflects past AI use – standard estimates *condition on an endogenous state*. When $\mu < 1$, both biases cause estimates to overstate long-run benefits; when $\mu > 1$, both reverse sign and estimates understate benefits. The measurement problem is general; the direction of bias is parameter-dependent.

The magnitude and direction of these biases hinge on a single parameter: pedagogical quality μ. When $\mu < 1$, AI substitutes for deliberate practice, both biases grow with adoption, and the skill trap becomes possible. When $\mu > 1$, AI augments learning and the biases reverse – cross-sectional estimates then *understate* long-run benefits.[19] The majority of available evidence is consistent with $\mu < 1$: developers slower with AI yet believing otherwise (METR, 2025), endoscopists deskilling within months (Budzyń et al., 2025), workers investing less cognitive effort (Dell'Acqua, 2022). Two independent experiments yield $\mu \approx 0.83$ (Bastani et al., 2025; Shen and Tamkin, 2026).

Our analysis has limitations. The parameter μ is context-dependent and imprecisely estimated; we report results for a range rather than defending a point estimate. Workers might reallocate effort freed by AI to complex tasks, but evidence suggests otherwise (Lee et al., 2025). We treat μ as exogenous and abstract from task-level heterogeneity; in practice, AI may augment learning in some tasks while degrading it in others. The model assumes symmetric workers and abstracts from selection into AI use, though we discuss these extensions in Appendix A. Future work should extend the model to portfolios of skills, incorporate political economy considerations, and explore endogenous effort allocation. Panel data tracking AI usage and skill assessments would permit direct estimation of μ; the emerging experimental literature provides a template.

The two biases require different remedies. Spillover degradation is an externality amenable to Pigouvian correction: taxes on AI use that internalize the harm to others' learning environments, or subsidies for human-generated training data. State-path divergence is fundamentally a measurement problem requiring counterfactual-aware research designs – difference-in-differences and randomized trials that compare AI users to non-users will systematically overstate benefits when both groups' skills have been shaped by AI's presence in the economy. Cohort comparisons exploiting variation in AI exposure across entry cohorts, or cross-country comparisons leveraging differential adoption timing, may better approximate the welfare-relevant counterfactual. The slow-recovery result adds urgency to policy timing: preventing skill degradation is far easier than reversing it, suggesting that early intervention – before widespread adoption entrenches low-skill equilibria – may be far more effective than remediation after the fact.

The framework speaks to broader debates about technology and human capital. Automation has historically displaced workers from specific tasks while creating new ones that require different skills. Generative AI may be distinctive in that it targets the *process* of skill acquisition itself, not just the tasks that skills enable. If learning-by-doing is central to ex-

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

pertise development, and if AI substitutes for the cognitive effort that learning requires, then the long-run effects may differ qualitatively from previous waves of automation. The empirical question of whether μ exceeds or falls short of unity is thus first-order for understanding AI's ultimate impact on human capital.

More broadly, a technology can appear *increasingly indispensable* even when it is not improving, because past use has degraded the alternative. This "lock-in through atrophy" represents a novel form of technological dependence distinct from network effects or switching costs. The welfare-relevant counterfactual is not the worker's current state without the technology, but *the skill path that would have obtained* absent adoption. Our framework suggests that the most important effects of transformative technologies may be precisely those that standard productivity measurement cannot detect.

References

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

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

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.

Agrawal, A., J. Gans, and A. Goldfarb (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Review Press.

Agrawal, A., J. Gans, and A. Goldfarb (2019). Exploring the Impact of Artificial Intelligence: Prediction versus Judgment. *Information Economics and Policy* 47, 1–6.

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

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

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

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

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

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

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.

Beane, M. (2024). *The Skill Code: How to Save Human Ability in an Age of Intelligent Machines*. HarperCollins.

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

Braverman, H. (1974). *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century*. Monthly Review Press.

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

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

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

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

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

Chen, Y. J., J. Gong, J. Li, and Z. Zhao (2025). Better Technology, Worse Motivation: GenAI's Mediocrity Trap. SSRN Working Paper 5208163.

Cho, S. (2024). The Effect of Robot Assistance on Skills. SSRN Working Paper 4902149.

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

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

Dell'Acqua, F. (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.

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

Dohmatob, E., Y. Feng, P. Yang, F. Charton, and J. Kempe (2024). A Tale of Tails: Model Collapse as a Change of Scaling Laws. *Proceedings of the 41st International Conference on Machine Learning*, 11165–11197.

Eloundou, T., S. Manning, P. Mishkin, and D. Rock (2024). GPTs are GPTs: Labor Market Impact Potential of LLMs. *Science* 384(6702), 1306–1308.

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.

Gerstgrasser, M., R. Schaeffer, A. Dey, R. Rafailov, H. Sleight, J. Hughes, T. Korbak, R. Agrawal, D. Pai, A. Gromov, D. A. Roberts, D. Yang, D. L. K. Yamins, and S. Koyejo (2024). Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data. *arXiv preprint arXiv:2404.01413*.

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

Handa, K., A. Tamkin, M. McCain, S. Huang, E. Durmus, S. Heck, J. Mueller, J. Hong, S. Ritchie, T. Belonax, K. K. Troy, D. Amodei, J. Kaplan, J. Clark, and D. Ganguli (2025). Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations. *arXiv preprint arXiv:2503.04761*.

Hudgens, M. G. and M. E. Halloran (2008). Toward Causal Inference with Interference. *Journal of the American Statistical Association* 103(482), 832–842.

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

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

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

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

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

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

Manski, C. F. (1993). Identification of Endogenous Social Effects: The Reflection Problem. *Review of Economic Studies* 60(3), 531–542.

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

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

Mincer, J. (1974). *Schooling, Experience, and Earnings*. Columbia University Press.

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

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

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

Parasuraman, R. and V. Riley (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors* 39(2), 230–253.

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. *Quarterly Journal of Economics* 116(2), 681–704.

Sarter, N. B., D. D. Woods, and C. E. Billings (1997). Automation Surprises. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics* (2nd ed., pp. 1926–1943). Wiley.

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

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

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

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

Thompson, P. (2010). Learning by Doing. *Handbook of the Economics of Innovation* 1, 429–476.

Vázquez-Bare, G. (2023). Identification and Estimation of Spillover Effects in Randomized Experiments. *Journal of Econometrics* 237(1), 105237.

Ying, Q., W. Dong, and S. I. Fabrikant (2024). How Do In-Car Navigation Aids Impair Expert Navigators' Spatial Learning Ability? *Annals of the American Association of Geographers*, 1–22.

A Extensions

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

A.1 Firm Dynamics and Selection

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

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

Proposition 10 (Selection Effects)**.** *Under Assumption 5, during the transition from initial conditions:*

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

(ii) Let $s_{i,t}$ *denote firm* i*'s market share. During the transition phase when the static gain from AI dominates skill loss,* $\frac{d}{dt}\mathbb{E}[\beta_i | s_{i,t}] < 0$*: the output-weighted average patience declines.*[20]

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

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

A.2 Endogenous Certification and Skill Signaling

When AI makes it difficult to distinguish skilled from unskilled workers in ordinary output, markets for skill verification may emerge (Spence, 1973). This extension analyzes how certification institutions can partially mitigate the skill trap by preserving incentives for skill acquisition.

Assumption 6 (Hidden Skill)**.** Output is observable but the decomposition between AI and human contribution is not. A worker with human capital h using AI at intensity α produces output $Y(h, \alpha)$, but employers observe only Y, not h or α separately.

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

This assumption captures a key feature of AI-assisted work: the final product may look identical regardless of whether it was produced by a skilled worker with minimal AI assistance or an unskilled worker with heavy AI assistance. Traditional methods of evaluating worker quality – observing output, checking references, reviewing portfolios – become less informative when AI can augment any worker's apparent capabilities. The assumption connects to the broader literature on technology and skill observability (Autor et al., 2003).

Assumption 7 (Certification Technology)**.** A certification test measures human capital at cost $\kappa > 0$. The test accurately reveals h but cannot be taken with AI assistance (e.g., proctored professional licensing exams, in-person technical interviews).

Many existing professional certifications satisfy this assumption: medical boards, bar exams, CPA examinations, and technical interviews at major firms are conducted under conditions that preclude AI assistance. The rise of AI may increase demand for such certifications, or prompt the creation of new ones in fields where they did not previously exist.

Proposition 11 (Certification Equilibrium)**.** *Under Assumptions 6 and 7:*

(i) A separating equilibrium exists iff $w(h^{high}) - w(h^{low}) > \kappa$.

(ii) In the trap, certification value $V_t^{cert} \equiv w^C(h^{high}) - w_t^{NC}$ *is increasing in* t *as average skill* $\bar{h}_t$ *falls.*

(iii) Certification raises private returns to skill: $\frac{\partial V}{\partial h}\big|_{cert} > \frac{\partial V}{\partial h}\big|_{no\text{-}cert}$.

We emphasize that certification markets partially mitigate the skill trap by increasing private returns to skill, but certification addresses only the information problem, not the underlying human capital externality – it is a complement to, not substitute for, corrective policy. The proliferation of AI-era certifications may signal market recognition of the skill atrophy problem, even in the absence of formal policy intervention.

A.3 Adaptive Pedagogical AI Design

We analyze whether AI systems could be designed to mitigate skill atrophy by adjusting assistance based on user skill.

Definition 4 (Adaptive AI)**.** An adaptive AI system observes user skill h and chooses assistance level $\alpha(h)$ to maximize some objective:

- A *productivity-maximizing* AI chooses $\alpha^P(h) = \arg\max_\alpha Y(h, \alpha)$.
- A *learning-maximizing* AI chooses $\alpha^L(h) = \arg\max_\alpha L(\alpha, h; \mu)$.
- A *welfare-maximizing* AI chooses $\alpha^W(h)$ to maximize the present value of output plus human capital.

Proposition 12 (Optimal AI Design)**.** *Let* $\alpha^{opt}(h)$ *maximize* $V(h) = \sum_t \beta^t Y(h_t, \alpha_t)$ *subject to skill dynamics. Then:*

(i) $\alpha^{opt}(h) < \alpha^P(h)$ *for* $h < h^{threshold}$*, where* $\alpha^P(h) = \arg\max_\alpha Y(h, \alpha)$.

(ii) $\alpha^{opt}(h) \approx \alpha^{P}(h)$ *for* $h > h^{threshold}$.

(iii) $\frac{\partial h^{threshold}}{\partial \beta} > 0$ *and* $\frac{\partial h^{threshold}}{\partial \mu} < 0$.

The optimal AI design resembles "training wheels" that are removed as competence develops. This contrasts with standard AI optimization, which maximizes user productivity regardless of skill level. The model suggests that AI providers have incentives to over-assist users (since users prefer immediate productivity), creating a market failure in AI design: socially optimal AI would provide less assistance than privately optimal AI.

This market failure has a precise structure. Users choose AI systems based on immediate productivity gains, which are maximized by high-α Autocomplete interfaces. But lifetime welfare – accounting for skill formation – is maximized by lower-α Socratic interfaces during learning phases. The wedge between user preferences and social welfare widens when users are myopic (low β) or when AI-assisted work is particularly unformative for skill development (low μ). The problem is analogous to the tension between processed and nutritious food: immediate palatability conflicts with long-run health.

Concretely, contrast two interface paradigms: *Autocomplete* (AI provides complete solutions; user accepts or rejects; $\mu \approx 0$) versus *Socratic Tutor* (AI asks guiding questions, highlights errors without fixing them, requires user to articulate reasoning; μ potentially > 1). Current commercial incentives favor Autocomplete because users prefer immediate productivity (Dell'Acqua, 2022). But our analysis suggests Socratic interfaces preserve more human capital, even if measured adoption appears lower. The experimental results of Bastani et al. (2025) support this: pedagogically-designed AI tutors avoid the skill degradation observed with unrestricted AI access.

Several implementation approaches could address this market failure. Professional licensing bodies could mandate minimum engagement requirements during training periods, analogous to existing requirements for supervised practice hours. AI providers could be required to offer "learning mode" interfaces in educational and professional development contexts. Procurement policies for government and enterprise clients could favor AI systems with demonstrated pedagogical features. Tax incentives could subsidize development of high-μ AI designs, treating them as investments in human capital infrastructure rather than pure productivity tools.

A.4 Optimal Policy

This section provides formal results on optimal corrective policy when AI adoption generates externalities through human capital spillovers and training data degradation.

A.4.1 Pigouvian Taxation

The efficient corrective policy taxes AI use at a rate equal to the marginal external cost.

Proposition 13 (Optimal AI Tax)**.** *The optimal per-unit tax on AI adoption equals the*

marginal external cost evaluated at the current state (H, A)*:*

$$\tau^* = \underbrace{\beta \left[\frac{\partial W}{\partial H'} - V'(h') \right] \lambda(1-\mu)\varphi(H)\psi(H)}_{\textit{human capital externality}} + \underbrace{\beta \frac{\partial W}{\partial A'} \zeta |Q_\alpha|}_{\textit{training data externality}}$$

The first term is the wedge between social and private marginal values of human capital, times the marginal effect of adoption on skill formation. When spillovers are absent ($\theta = 0$, $\psi \equiv 1$), this term is zero because $\partial W/\partial H' = V'(h')$. The optimal tax is state-dependent (varying with H and A), exhibits corrective feedback (rising as H falls because skill scarcity raises the marginal value of remaining human capital), and evolves dynamically along equilibrium paths.

The corrective feedback property is notable: as human capital falls, the marginal value of remaining human capital rises, justifying higher taxes over time. This contrasts with standard Pigouvian taxes that are typically constant.

A.4.2 Competitive Dynamics

Individual firms face competitive pressure to adopt AI even when they recognize its long-run costs. Consider a symmetric duopoly where firm i's market share is $s_i = Y_i/(Y_i + Y_j)$. Each firm's first-order condition includes a business-stealing term $(\partial s_i/\partial \alpha_i) \cdot \Pi > 0$ that a joint maximizer would ignore. This generates overadoption: Nash equilibrium adoption α^N exceeds the joint-profit-maximizing level α^M. The competitive wedge compounds with the externalities analyzed in the main text; Appendix B provides the formal proof.

A.5 Microfoundations for Spillovers

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

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

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

$$\mathbb{E}[L_i] = \int_0^{h_i} \varphi(h_i)\, dF(s) + \int_{h_i}^{\bar{s}} \kappa\varphi(h_i)[1 - G_H(s)]\, dF(s) \tag{10}$$

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

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

Expected learning can be written as:

$$\mathbb{E}[L_i] = \varphi(h_i)\left[F(h_i) + \kappa\Psi(H)[1 - F(h_i)]\right] \tag{11}$$

Normalizing so that $\psi(\bar{H}) = 1$ at the no-adoption steady state, the term in brackets motivates a multiplicative *approximation* $L_i \approx \ell(\alpha_i)\varphi(h_i) \cdot \psi(H)$.[21] The key insight is that aggregate human capital affects individual learning through the availability of mentors: when H falls, the probability of finding a capable mentor declines, reducing learning for all workers – including those who do not adopt AI.

A.6 Microfoundations for Training Data Degradation

This section provides a formal microfoundation for the AI quality function $Q(H, \bar{\alpha})$ introduced in Section 5 and characterizes the feedback loop dynamics.

AI firm's data acquisition problem. Consider an AI firm that trains its model on a corpus of human-generated content. Each period, the firm observes output from a population of workers. Worker i produces content of quality $q_i = h_i \cdot (1-\alpha_i)^{\omega}$, where h_i is human capital, α_i is AI adoption intensity, and $\omega > 0$ governs how AI assistance affects output quality. The term $(1-\alpha_i)^{\omega}$ captures that AI-assisted output, while potentially correct, lacks the distinctive features (edge cases, creative solutions, expert judgment) that make training data valuable.

The AI firm's training corpus has two components: (1) human-generated content with quality distribution G_q, and (2) AI-generated content that has "leaked" into the training set. Let π_t denote the fraction of AI-generated content in the corpus at time t. The effective training signal is:

$$S_t = (1 - \pi_t) \cdot \underbrace{\int q_i\, dF_i}_{\text{human quality}} + \pi_t \cdot \underbrace{A_{t-1}}_{\text{AI quality}} \tag{12}$$

where A_{t-1} is previous-period AI quality. The AI-generated component contributes A_{t-1} because AI can only reproduce what it already knows – it cannot generate genuinely novel training signal.

Model collapse dynamics. Following Shumailov et al. (2024), recursive training on AI-generated content causes quality degradation. The intuition is that each generation of AI "compresses" the distribution, losing tail information. Formally, let $f : \mathbb{R}_+ \to \mathbb{R}_+$ denote the training function mapping signal quality to AI capability. If $A_t = f(S_t)$ where S_t is training signal quality, then:

$$A_{t+1} = f\left((1 - \pi_t)\bar{q}_t + \pi_t A_t\right) \tag{13}$$

[21] The exact expression $F(h_i) + \kappa\Psi(H)[1 - F(h_i)]$ depends on individual skill via $F(h_i)$ and cannot be literally factored into $g(h_i) \cdot \psi(H)$. The multiplicative form in the main text is a reduced-form approximation that captures the key qualitative feature: aggregate human capital affects individual learning through mentor availability. See Appendix A.7 for robustness to alternative specifications.

where $\bar{q}_t = \int q_i \, dF_i$ is average human output quality. When π_t is high (much AI content in training data), the model increasingly trains on its own outputs, causing the "autophagy" documented by Alemohammad et al. (2024).

Connecting to skill formation. Average human output quality is:

$$\bar{q}_t = \int h_i(1-\alpha_i)^\omega \, dF_i \approx H_t \cdot (1-\bar{\alpha}_t)^\omega \tag{14}$$

for symmetric adoption $\alpha_i = \bar{\alpha}$. This yields the reduced-form specification in the main text. To be precise about timing: define $\tilde{Q}(H, \bar{\alpha}) \equiv (1-\pi) \cdot H \cdot (1-\bar{\alpha})^\omega$ as the *human contribution* to training signal quality. The full law of motion for AI quality is:

$$A_{t+1} = (1-\zeta)A_t + \zeta \cdot [(1-\pi)\tilde{Q}(H_t, \bar{\alpha}_t) + \pi \cdot A_t] \tag{15}$$

which simplifies to $A_{t+1} = (1-\zeta(1-\pi))A_t + \zeta(1-\pi)\tilde{Q}(H_t, \bar{\alpha}_t)$. The term $\pi \cdot A_t$ captures AI-generated content in the training corpus, which reflects current AI quality (the AI can only reproduce what it already knows). For notational simplicity, the main text absorbs these terms into a single function $Q(H_t, \bar{\alpha}_t)$ satisfying $\partial Q/\partial H > 0$ (skilled humans produce better training data) and $\partial Q/\partial \bar{\alpha} < 0$ (adoption degrades output quality). The contamination rate π is itself endogenous to adoption: $\pi_t = \pi(\bar{\alpha}_t)$ with $\pi' > 0$, but we suppress this dependence for tractability.

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

$$H_{t+1} = (1-\delta)H_t + \lambda \ell(\bar{\alpha}_t)\varphi(H_t)\psi(H_t) \tag{16}$$

$$A_{t+1} = (1-\zeta)A_t + \zeta Q(H_t, \bar{\alpha}_t) \tag{17}$$

where $\bar{\alpha}_t = \alpha^*(H_t, A_t)$ is equilibrium adoption given state (H_t, A_t).

Proposition 14 (Feedback Loop Stability)**.** *The system (H_t, A_t) has a unique stable steady state (H^{**}, A^{**}) satisfying:*

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

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

(iii) The steady state (H^{**}, A^{**}) is globally stable when $\mu < 1$ and ζ is sufficiently small (AI quality adjusts slowly relative to human capital).

Comparative statics of the feedback loop. The feedback loop's stabilizing effect on levels – $H^{**} > H^*(A_0)$ – varies with parameters. Define the *stabilization gain* as $\Delta_S \equiv H^{**} - H^*(A_0) > 0$, the additional human capital preserved due to endogenous AI quality degradation. This gain is increasing in:

- ζ: faster AI quality adjustment strengthens the feedback

- ω: stronger quality degradation from AI-assisted output (larger $\partial Q/\partial \bar{\alpha}$)
- $|\partial \alpha^*/\partial A|$: stronger adoption response to AI quality

The stabilization is larger when AI systems are retrained frequently on recent data, when AI-assisted output is easily distinguishable from expert human output, and when firms strongly reduce adoption in response to lower AI quality.

Escape from the trap. Unlike pure skill atrophy, the training data channel offers a potential escape route: if AI firms can curate training data to exclude AI-generated content and prioritize high-skill human output, the degradation can be arrested. Formally, if $\pi_t \to 0$ (perfect filtering of AI content) and the firm overweights high-h workers' output, then A_t can be stabilized or even improved – but this requires the continued existence of a non-trivial high-skill population (e.g., pre-AI cohorts who maintain $\bar{h}$ or specially preserved expert groups). In the representative-agent model, all workers converge to h^*, so there is no high-skill tail to curate. The curation strategy thus implicitly assumes heterogeneity that goes beyond the baseline model. This suggests a role for data provenance systems, human-generated content certification, and premium markets for expert-produced training data. However, curation addresses *contamination* (the π channel) but not *depletion* (the H channel): as human skills atrophy, the supply of high-quality human content diminishes regardless of filtering efficacy. Technology-side fixes cannot substitute for human-side skill preservation.

A.7 Robustness to Functional Forms

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

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

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

Alternative spillover specifications. Replace the multiplicative specification $L_i = \ell(\alpha_i)\varphi(h_i)\psi(H)$ with an additive form $L_i = \ell(\alpha_i)\varphi(h_i) + \theta_L H$, where $\theta_L > 0$ captures direct knowledge spillovers. The overadoption result (Proposition 7) continues to hold: individual firms ignore their contribution to H, so private adoption exceeds social optima. The quantitative magnitude of the wedge changes, but the qualitative inefficiency result is robust.

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

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

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

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

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

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

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

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

Upper-tail spillover specification. As noted in the main text, the microfoundation in Appendix A.5 implies spillovers that depend on the full skill distribution, not merely the mean. We verify robustness to an alternative specification where spillovers depend on the upper tail:

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

where $h^{threshold}$ is a fixed mentorship threshold and $1 - G_H(h^{threshold})$ is the fraction of workers above it. As AI adoption causes skills to atrophy, more workers fall below the threshold, reducing $\tilde{\psi}$ and impairing learning for all workers. The comparative statics are identical to the mean-based specification: $\partial\tilde{\psi}/\partial\alpha < 0$ when $\mu < 1$, generating overadoption.

B Proofs

This appendix provides formal proofs for all results. Section B.1 states and proves technical lemmas; Section B.2 proves the main results. The skill trap proof (Proposition 4) appears before the spillover and state-path divergence proofs because Part (ii) of the latter references the trap characterization; otherwise proofs follow the order of the text. We begin by stating the regularity conditions maintained throughout.

Assumption 8 (Regularity)**.** The following conditions hold at steady state:

(i) **Interior steady state:** $h^* \in (0, \bar{h})$, where $\bar{h}$ is the no-adoption steady state.

(ii) **Stability:** $\delta > \lambda \ell(\alpha^*)|\varphi'(h^*)|$, ensuring $|T'(h^*)| < 1$.

(iii) **Curvature dominance:** $|Y_{\alpha\alpha}(h^*, \alpha^*)| > \beta |V'(h^*)| \lambda |1 - \mu| |\varphi'(h^*)|$.

(iv) **Monotone policy:** $d\alpha^*/dh$ has constant sign on $(0, \bar{h}]$.

These conditions ensure existence, uniqueness, and stability of equilibrium. Condition (i) places the steady state in the economically relevant region. Condition (ii) is standard local stability. Condition (iii) ensures static concavity dominates dynamic effects in the FOC. Condition (iv) rules out pathological non-monotonic policy functions.

Sufficient primitive conditions. Conditions (i)–(iv) hold when: (a) δ is bounded away from zero (skill depreciates); (b) φ is Lipschitz with $|\varphi'(h)| \leq M$ for some $M < \infty$; (c) $\beta < 1/(1+\delta)$ (firms are not too patient); and (d) $|g'(\alpha)|$ is bounded away from zero (AI capability declines with task complexity). Under these primitives, the set of parameter values violating (i)–(iv) has measure zero.

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

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

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

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

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

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

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

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

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

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

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

Lemma 9 (Stability Characterization)**.** *At a steady state h^*, local stability holds when $|T'(h^*)| < 1$, where $T'(h^*) = (1-\delta) + \lambda\ell(\alpha^*)\varphi'(h^*) + \lambda\ell'(\alpha^*)\frac{d\alpha^*}{dh}\varphi(h^*)$. Under Assumption 8(ii)–(iv), a sufficient condition is $\delta - \lambda\ell(\alpha^*)|\varphi'(h^*)| > 0$: the stability term dominates the policy feedback term, which is bounded under curvature dominance.*

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

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

The first two terms give $(1-\delta) + \lambda\ell(\alpha^*)\varphi'(h^*)$. Since $\varphi'(h^*) < 0$ by Assumption 1, this is less than $(1-\delta) < 1$. The third term – the policy feedback – has magnitude bounded by Assumption 8(iii)–(iv): curvature dominance ensures $|d\alpha^*/dh|$ is small, and monotone policy ensures it has constant sign. Combining, $|T'(h^*)| < 1$ when $\delta - \lambda\ell(\alpha^*)|\varphi'(h^*)| > 0$. □

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

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

Step 1: Existence and location of steady state. By Lemma 2, there exists a unique $h^* > 0$ satisfying the stationarity condition. Under Assumption 8(i), $h^* \in (0, \bar{h})$.

Step 2: Local stability. The derivative $T'(h^*) = (1-\delta) + \lambda\ell(\alpha^*)\varphi'(h^*) + \lambda\ell'(\alpha^*)\frac{d\alpha^*}{dh}\varphi(h^*)$. Under Assumption 8(ii), $(1-\delta) + \lambda\ell(\alpha^*)\varphi'(h^*) < 1$. The third term is bounded under Assumption 8(iii)–(iv). Thus $|T'(h^*)| < 1$, establishing local asymptotic stability.

Step 3: Global convergence from $(0, \bar{h}]$. For $h \in (0, \bar{h}]$, we show $T(h)-h$ has constant sign on each side of h^*. At $h = \bar{h}$: $T(\bar{h}) = (1-\delta)\bar{h} + \lambda\ell(\alpha^*(\bar{h}))\varphi(\bar{h})$. Since $\ell(\alpha) < 1$ when $\alpha > 0$ and $\mu < 1$, and since $\delta\bar{h} = \lambda\varphi(\bar{h})$ defines $\bar{h}$, we have $T(\bar{h}) < \bar{h}$. At h^*: $T(h^*) = h^*$. By continuity and the intermediate value theorem, for $h \in (h^*, \bar{h}]$, we have $T(h) < h$, so the sequence is decreasing. Local stability then implies $h_t \to h^*$. □

Lemma 11 (Jacobian Non-Singularity)**.** *Under Assumption 8, 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 8(ii)
- $D_{h\alpha} = \lambda(1-\mu)\varphi(h) > 0$ since $\mu < 1$ and $\varphi(h) > 0$
- $D_\alpha = Y_{\alpha\alpha} + \beta V''(h')[\lambda(\mu-1)\varphi(h)]^2$
- $D_{\alpha h} = -(1-\gamma)(1-\alpha)^{-\gamma} + \beta\lambda(\mu-1)\left[V''(h')\frac{\partial h'}{\partial h}\varphi(h) + V'(h')\varphi'(h)\right]$

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

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

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

Note that $\frac{\partial h'}{\partial h}$ multiplies only the $V''(h')$ term, not the $V'(h')\varphi'(h)$ term – this follows from the chain rule since $V'(h')$ depends on h through h', while $\varphi'(h)$ depends directly on h. The first term $-(1-\gamma)(1-\alpha)^{-\gamma} < 0$. For the bracketed expression when $\mu < 1$: $V''(h') \leq 0$ by concavity, $\frac{\partial h'}{\partial h} > 0$, $\varphi(h) > 0$, so the first bracketed term is non-positive. For the second term: $V'(h') > 0$, $\varphi'(h) < 0$ by Assumption 1, so $V'(h')\varphi'(h) < 0$. Thus the bracket is non-positive. With $(\mu-1) < 0$, we have $\beta\lambda(\mu-1) \cdot (\text{non-positive}) \geq 0$, making the second term non-negative. The sign of $D_{\alpha h}$ depends on which effect dominates.

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

Under Assumption 8(iv) (monotone policy), $D_{\alpha h}$ has constant sign on $(0, \bar{h}]$. If $D_{\alpha h} > 0$, then $D_{h\alpha}D_{\alpha h} > 0$ and hence $-D_{h\alpha}D_{\alpha h} < 0$, so both terms in $\det(\mathbf{J}) = D_h D_\alpha - D_{h\alpha}D_{\alpha h}$ are negative and $\det(\mathbf{J}) < 0$ unconditionally. If $D_{\alpha h} \leq 0$, then $D_{h\alpha}D_{\alpha h} \leq 0$ and hence $-D_{h\alpha}D_{\alpha h} \geq 0$, so $\det(\mathbf{J})$ is the sum of a negative term ($D_h D_\alpha < 0$) and a non-negative term

$(-D_{h\alpha}D_{\alpha h} \geq 0)$; in this case, the sign of $\det(\mathbf{J})$ is ambiguous unless we impose $|D_h D_\alpha| > |D_{h\alpha}D_{\alpha h}|$. This latter condition is implied by Assumption 8(iii): when static curvature dominates, the cross-partial products are second-order relative to $|Y_{\alpha\alpha}|$.

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

B.2 Proofs of Main Results

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

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

Substituting the derivatives and rearranging:

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

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

Lemma 2 (Steady-State Human Capital Function).
(i) Define $\Phi(h; \alpha) \equiv \delta h - \lambda\ell(\alpha)\varphi(h)$ where $\ell(\alpha) = 1 - (1-\mu)\alpha$. For existence, we require $\ell(\alpha) > 0$, which holds for all $\alpha \in [0, 1)$ when $\mu \geq 0$ since $\ell(\alpha) \geq 1 - \alpha > 0$.

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

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

(ii) At $\alpha = 0$: $\ell(0) = 1$, so (6) becomes $\delta h = \lambda\varphi(h)$, which defines $\bar{h}$.

(iii)–(iv) Implicitly differentiating (6):

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

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

Lemma 3 (Steady-State Characterization).
The characterization follows directly from the properties of the steady-state human capital function $h^*(\alpha)$ established in Lemma 2. □

Lemma 4 (Uniqueness and Global Stability).

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

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

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

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

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

- At h close to 0: The stationarity condition $\delta h = \lambda \ell(\alpha)\varphi(h)$ with $\varphi(0) > 0$ requires α close to $1/(1-\mu) > 1$ for $\mu \in (0, 1)$, which is outside $[0, 1]$. Thus for any $\alpha \in [0, 1]$, stationarity requires $h > 0$. Meanwhile, the optimal policy has $\alpha^*(h) \to \alpha^{max} < 1$ as $h \to 0$ (AI remains valuable even at low skill).
- At $h = \bar{h}$: Stationarity with $\alpha = 0$ gives $\delta \bar{h} = \lambda \varphi(\bar{h})$, which defines $\bar{h}$. Thus $h_S(0) = \bar{h}$. The optimal policy has $\alpha^*(\bar{h}) > 0$ by Assumption 3.

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

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

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

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

Step 1: The policy function is strictly decreasing. By Assumption 8(iv), $d\alpha^*/dh$ has constant sign. We show this sign is negative when $\mu < 1$. The FOC for optimal adoption is $Y_\alpha(h, \alpha) + \beta V'(h') \cdot \partial h'/\partial \alpha = 0$. The cross-partial $\partial^2/\partial h \partial \alpha$ of the Bellman objective includes the term $-(1-\gamma)(1-\alpha)^{-\gamma} < 0$ from $Y_{h\alpha}$. Since higher h raises output more when α is lower, and since the dynamic cost $\beta V'(h')\lambda(1-\mu)\varphi(h)$ is positive when $\mu < 1$, the optimal response to higher h is lower α. Thus $d\alpha^*/dh < 0$.

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

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

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

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

Necessity of Substitution for Skill Atrophy.

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

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

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

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

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

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

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

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

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

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

Corollary 1 (Comparative Statics).

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

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

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

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

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

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

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

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

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

Solving:

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

The denominator is positive by Assumption 8. The numerator has two terms:

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

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

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

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

Proposition 4 (Existence of Skill Trap).

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

Step 1: Condition (T1) holds. By Assumption 3, $A > \bar{h}(1-\gamma)$. By Lemma 8, $\alpha^*(h) > 0$ for all $h \in (0, \bar{h}]$. Since $h_0 \leq \bar{h}$ and human capital remains bounded in $(0, \bar{h}]$ along any equilibrium path (Lemma 7), we have $\alpha_t > 0$ for all t.

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

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

Differentiating at $\alpha_0 = 0$: $\partial Y_0/\partial\alpha|_{\alpha=0} = A \cdot g(0) - h_0(1-\gamma) = A - h_0(1-\gamma) > 0$ by Assumption 3. Since the firm chooses $\alpha_0^* > 0$ (Lemma 8) and payoff is strictly concave in α (Lemma 6), we have $Y_0 > Y_0^{NA}$.

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

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

$$\frac{dh^*}{d\alpha} = \frac{-\lambda(1-\mu)\varphi(h^*)}{\delta - \lambda \ell(\alpha)\varphi'(h^*)} \tag{18}$$

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

Differentiating W:

$$W'(\alpha) = Ag(\alpha) + \frac{dh^*}{d\alpha}(1-\alpha)^{1-\gamma} - h^*(1-\gamma)(1-\alpha)^{-\gamma} \tag{19}$$

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

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

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

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

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

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

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

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

$$V'(h^*) = \frac{(1-\alpha^*)^{1-\gamma}}{1 - \beta(1-\delta) - \beta\lambda\ell(\alpha^*)\varphi'(h^*)} \tag{20}$$

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

Substituting into $W'(\alpha^*)$. Recall from (18):

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

Substituting (20) and this expression into $W'(\alpha^*)$:

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

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

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

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

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

since all terms in the denominator are positive.

Therefore $W'(\alpha^*) < 0$ *unconditionally* at any stable interior steady state with $\mu < 1$. The sign does not require any additional assumption beyond those already imposed (Assumptions 3–8).

Remark 6. The monotonicity result $W'(\alpha^*) < 0$ does not require any additional assumption beyond Assumptions 3–8. The impatience condition in Remark 7 ensures well-behaved comparative statics but is not needed for the sign of $W'(\alpha^*)$.

Remark 7 (Impatience Condition). The following condition, while not required for $W'(\alpha^*) < 0$, ensures well-behaved comparative statics: $\delta - \lambda\ell(\alpha^*)\varphi'(h^*) < (1-\beta)/\beta$. This holds when firms are sufficiently impatient relative to depreciation. A sufficient primitive condition is $\beta < 1/(1+\delta+\lambda\bar{m})$ where $\bar{m} = \sup_h |\varphi'(h)|$.

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

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

Steady-state output is strictly increasing in firm patience.

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

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

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

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

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

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

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

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

Step 7: Necessity. (a) If $\mu \geq 1$: as shown above (Necessity of Substitution), $h^* \geq \bar{h}$. For the trap to fail, we need $Y^* \geq \bar{h}$. We have $Y^* = AG(\alpha^*) + h^*(1-\alpha^*)^{1-\gamma}$. Since $h^* \geq \bar{h}$ and $AG(\alpha^*) > 0$ for $\alpha^* > 0$, a sufficient condition for $Y^* \geq \bar{h}$ is:

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

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

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

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

Proposition 1 (Spillover Bias).

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

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

The cross-sectional counterfactual is:

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

The long-run counterfactual is:

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

The difference is:

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

since $h_t^{NU} < \bar{h}$ for $t > 0$. The gap is zero at $t = 0$ (before adoption affects non-users) and strictly increasing in t as h_t^{NU} falls further below $\bar{h}$. □

Proposition 2 (State-Path Divergence).

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

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

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

As $h_t^U \to h^*$, the absolute gain $\Delta_t^{SC} \to A \cdot G(\alpha^*) - h^*[1 - (1-\alpha^*)^{1-\gamma}]$, which is bounded. The *relative* gain Δ_t^{SC}/h_t^U satisfies:

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

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

Claim 2: Ratio eventually increases. Along the transition path, h_t^U is decreasing (since $h_0 = \bar{h} > h^*$ and the system converges monotonically) while α_t is increasing. The ratio $\Delta_t^{SC}/h_t^U = AG(\alpha_t)/h_t^U - [1 - (1-\alpha_t)^{1-\gamma}]$ has a first term that is strictly increasing (numerator bounded, denominator falling) and a second term that is strictly decreasing (being subtracted, and increasing in α_t). The sign of the change depends on relative magnitudes. However, as $t \to \infty$, we have $h_t^U \to h^* > 0$ and $\alpha_t \to \alpha^*$, so both terms converge to finite limits. What matters for the proposition is that Δ_t^{SC}/h_t^U can become arbitrarily large for parameterizations where $h^*/\bar{h}$ is small – this follows from the limit analysis above regardless of period-by-period monotonicity.

Part (ii): From Proposition 4, when the economy is in a skill trap, steady-state output satisfies $Y^* < \bar{h} = Y^{NA}$. Yet for any t sufficiently large that h_t^U is near h^*, we have $\Delta_t^{SC} > 0$ (AI raises current output given current skills). This is the core of state-path divergence: $\Delta_t^{SC} > 0$ while $Y^* < \bar{h}$. □

Corollary 2 (Welfare Reversal Under Patient Evaluation).

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

Corollary 3 (Feedback Loop: Stabilizing Force on Levels).

By Corollary 1(i), $\partial\alpha^*/\partial A > 0$ and $\partial h^*/\partial A < 0$: higher AI quality induces more adoption and lower steady-state skills.

Consider two systems:

- System (a): Fixed AI quality $A = A_0 = Q(\bar{H}, 0)$ (quality when humans are fully skilled and AI is unused). Steady state $H^*(A_0)$ satisfies $\delta H = \lambda \ell(\alpha^*(H; A_0))\varphi(H)$.
- System (b): Endogenous AI quality. Steady state (H^{**}, A^{**}) satisfies both the skill stationarity condition and $A^{**} = Q(H^{**}, \alpha^{**})$.

In system (b), if $\alpha^{**} > 0$ and $H^{**} < \bar{H}$ (skill atrophy occurs), then:

$$A^{**} = Q(H^{**}, \alpha^{**}) < Q(\bar{H}, 0) = A_0$$

since $\partial Q/\partial H > 0$ and $\partial Q/\partial \alpha < 0$. The feedback loop degrades AI quality.

The key observation is that this degradation partially protects human capital. Define $A_0 \equiv Q(\bar{H}, 0)$ as the AI quality when humans are fully skilled. Since $A^{**} < A_0$ (established above) and $\partial h^*/\partial A < 0$ (from Corollary 1(i)), it follows that:

$$H^{**} = h^*(A^{**}) > h^*(A_0) = H^*(A_0)$$

Lower AI quality induces less adoption, which reduces skill atrophy. □

Proposition 3 (Slow Recovery).

Consider the joint dynamics of human capital and AI quality:

$$H_{t+1} = (1 - \delta)H_t + \lambda \ell(\alpha_t)\varphi(H_t)$$
$$A_{t+1} = (1 - \zeta)A_t + \zeta Q(H_t, \alpha_t)$$

Define timescales $\tau_H \equiv 1/\lambda$ and $\tau_A \equiv 1/\zeta$.

Part (i): Decline during shock. When $\alpha > \alpha^{**}$, learning is reduced: $\partial L/\partial \alpha = (\mu - 1)\varphi(H) < 0$ for $\mu < 1$. Human capital declines because depreciation exceeds reduced learning. AI quality declines because $\partial Q/\partial \alpha < 0$ and $\partial Q/\partial H > 0$ (lower H also degrades training data).

Part (ii): Recovery bottleneck. Suppose the shock ends at time T. For tractability, assume adoption returns to α^{**}. Post-shock, $H_T < H^{**}$ and $A_T < A^{**}$. AI quality dynamics become:

$$A_{t+1} - A^{**} = (1 - \zeta)(A_t - A^{**}) + \zeta[Q(H_t, \alpha^{**}) - Q(H^{**}, \alpha^{**})]$$

Since $\partial Q/\partial H > 0$ and $H_t < H^{**}$, the bracketed term is negative: AI cannot recover until H recovers. The binding constraint is human capital accumulation, which evolves as:

$$H_{t+1} - H^{**} \approx (1 - \delta + \lambda \ell(\alpha^{**})\varphi'(H^{**}))(H_t - H^{**})$$

Convergence rate is governed by λ, not ζ. Hence recovery time $T_R \geq c \cdot \tau_H$ for some constant $c > 0$ that depends on the magnitude of the shock but not on ζ.

Part (iii): Persistence. When $\tau_H \gg \tau_A$, human learning is the slow variable. A shock of duration T creates a human capital deficit $\Delta H \propto T$. Recovery requires $T_R \sim \tau_H \cdot f(\Delta H/H^{**})$ where f is increasing. For substantial shocks, $T_R \gg T$: temporary shocks produce persistent effects because human expertise accumulates slowly while AI can only recover as fast as its training data improves. □

Corollary 4 (Sign Reversal).
In the skill trap, $Y^* < \bar{h}$ by Proposition 4, so $\Delta^{LR} = Y^* - \bar{h} < 0$. For $\Delta^{CS} > 0$, we need $Y^* > h^{NU*}$. Learning spillovers ensure $h^{NU*} < \bar{h}$: non-users' steady-state skill satisfies $\delta h^{NU*} = \lambda\varphi(h^{NU*})\psi(H^*)$ with $\psi(H^*) < 1$, implying $h^{NU*} < \bar{h}$. When $Y^* > h^{NU*}$ (AI users outperform degraded non-users) but $Y^* < \bar{h}$ (AI users underperform the no-adoption benchmark), we have $\Delta^{CS} > 0 > \Delta^{LR}$. □

Proposition 7 (Human Capital Externality).
The social planner maximizes $\sum_t \beta^t[Y(H_t, \alpha_t; A) + \theta H_t^\eta]$ subject to $H_{t+1} = (1-\delta)H_t + \lambda L(\alpha_t, H_t; \mu) \cdot \psi(H_t)$, where $\psi(H)$ captures learning spillovers.

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

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

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

Proposition 8 (Training Data Externality).
Part (i): With endogenous AI quality, $A_{t+1} = (1-\zeta)A_t + \zeta Q(H_t, \bar{\alpha}_t)$ with $\partial Q/\partial \bar{\alpha} < 0$. Each atomistic firm i chooses α_i taking $\bar{\alpha}$ as given. The private FOC is:

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

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

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

Proposition 9 (Training Mandates).
Without policy, the decentralized equilibrium features adoption $\alpha^D > \alpha^S$ (by Proposition 7). A mandate ρ constrains $\alpha \leq 1 - \rho$.

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

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

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

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

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

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

Corollary 5 (Comparative Statics of the Optimal Mandate).

Since $\rho^* = 1 - \alpha^S$, it suffices to show the comparative statics for α^S. The social optimum satisfies the first-order condition $\partial W/\partial\alpha = 0$, where $W = \sum_t \beta^t [Y_t + \theta H_t^\eta]$. By the implicit function theorem:

Part (i): Higher η increases the marginal social cost of skill loss ($\theta\eta H^{\eta-1}$), reducing α^S, hence $\partial\rho^*/\partial\eta = -\partial\alpha^S/\partial\eta > 0$.

Part (ii): Lower μ reduces effective learning per unit of AI use ($\partial\ell/\partial\mu = \alpha > 0$), increasing the skill cost of adoption. The social planner responds by reducing α^S, so $\partial\rho^*/\partial\mu = -\partial\alpha^S/\partial\mu < 0$ for $\mu < 1$. When $\mu \geq 1$, learning is non-decreasing in α, externalities vanish, and $\alpha^S = \alpha^D$ implies $\rho^* = 0$.

Part (iii): Higher β raises the weight on future skill in W. Since the private optimum already reflects discounting at rate β, the gap $\alpha^D - \alpha^S$ narrows as β increases (more patient firms internalize more of the future cost), so $\partial\rho^*/\partial\beta < 0$. □

Corollary 6 (AI Design).

Part (i): From Corollary 1, steady-state human capital h^* satisfies $\delta h^* = \lambda\ell(\alpha^*;\mu)\varphi(h^*)$ where $\ell(\alpha;\mu) = (1-\alpha) + \mu\alpha$. For fixed α, $\partial\ell/\partial\mu = \alpha > 0$, so higher μ raises learning and thus h^* (direct effect). However, by Corollary 1(iv), $\partial\alpha^*/\partial\mu > 0$ when $\mu < 1$ – higher pedagogical quality induces more adoption – and $\partial h^*/\partial\alpha < 0$ by Lemma 2. Thus the indirect effect through α^* *partially offsets* the direct effect: higher μ raises adoption, which lowers h^*. The net effect of μ on h^* is positive under conditions established in Corollary 1, but the offset means the gain in h^* is smaller than if adoption were held fixed.

Part (ii): Welfare is $W = \sum_t \beta^t [Y(h_t, \alpha_t) + \theta H_t^\eta]$. At steady state:

$$\frac{\partial W}{\partial \mu} = \frac{\partial W}{\partial h^*}\frac{\partial h^*}{\partial \mu} + \frac{\partial W}{\partial \alpha^*}\frac{\partial \alpha^*}{\partial \mu}$$

The first term is positive (higher μ raises h^*, which raises welfare). The second term captures the adoption response: higher μ changes optimal α^*, but since μ directly improves learning quality, the welfare gain from μ exceeds what could be achieved by equivalently constraining α.

Part (iii): Commercial AI maximizes user adoption, which depends on immediate productivity gains. The market failure in μ arises from forward-looking users with intermediate $\beta \in (0,1)$ who value future skills but underweight them relative to the social optimum. Specifically: (a) users with $\beta > 0$ strictly prefer higher μ (it raises future h without affecting current Y), but private demand for high-μ AI is insufficient when users ignore spillovers to

others' learning; (b) AI firms competing for market share cater to users' private willingness to pay, which underweights the social return to skill preservation. The knife-edge case $\beta = 0$ is instructive but not the source of market failure: myopic users are indifferent over μ (it doesn't affect their payoff), so there's no private demand for high-μ AI from this group, but also no distortion since they don't value skills anyway. The market failure requires $\beta > 0$ combined with spillover externalities or social discounting ($\tilde{\beta} > \beta$). Under these conditions, equilibrium $\mu^D < \mu^S$. □

Corollary (Inequality Dynamics).

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

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

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

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

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

Proposition 5 (Ability Reversal and Vintage Premium).

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

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

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

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

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

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

Differentiating with respect to θ:

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

The second term is positive since $\varphi(h_t^{NA}) > \ell(\alpha_t)\varphi(h_t^U)$ (the no-adoption path has higher effective learning). For the third term: $\varphi' < 0$ by Assumption 1, $\partial h_t^{NA}/\partial\theta > 0$, and $\partial h_t^U/\partial\theta > 0$. Since $h_t^{NA} > h_t^U$ (the no-adoption path yields higher skill), the sign of the third term depends on the curvature of φ. The induction succeeds when φ is log-concave (i.e., $\varphi''/\varphi - (\varphi'/\varphi)^2 \leq 0$), which ensures that the "direct scaling effect" (higher θ scales up the learning differential) dominates the "convergence effect" (higher θ pushes both paths into regions where φ is lower). The functional form $\varphi(h) = c/(1+h)$ used in calibration satisfies

this condition. Thus $\partial\Delta_{t+1}/\partial\theta > 0$, completing the induction. The intuition: ability scales learning, so high-ability workers forgo more learning when AI substitutes for practice.

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

Proposition 6 (Hump-Shaped Inequality).

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

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

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

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

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

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

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

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

Corollary 4 (Sign Reversal).

By Definition 3, the skill trap requires $Y_t < Y_t^{NA}$ for large t, so $\Delta_t^{LR} = Y_t - Y_t^{NA} < 0$.

For the cross-sectional estimate: $\Delta_t^{CS} = Y(h_t^U, \alpha_t) - Y(h_t^{NU}, 0)$. With learning spillovers (Assumption 4), non-users face degraded learning environments because aggregate skill H_t has fallen. Thus $h_t^{NU} < h_t^{NA}$ (non-users in an AI economy have lower skill than they would in a no-AI economy). Since Y is increasing in h:

$$\Delta_t^{CS} = Y(h_t^U, \alpha_t) - Y(h_t^{NU}, 0) > Y(h_t^U, \alpha_t) - Y(h_t^{NA}, 0) = \Delta_t^{LR}$$

When $\Delta_t^{LR} < 0$ (trap) and spillovers are sufficiently strong, we can have $\Delta_t^{CS} > 0 > \Delta_t^{LR}$: cross-sectional estimates show positive effects while long-run effects are negative. □

Proposition 10 (Selection Effects).

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

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

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

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

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

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

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

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

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

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

Proposition 11 (Certification Equilibrium).

Part (i): Consider a candidate separating equilibrium with threshold h^*: workers with $h \geq h^*$ certify, others do not. From the main model, wages equal marginal products: $w(h, \alpha) = (1-\alpha)^{1-\gamma} \cdot h$. For certified workers, employers observe h directly and pay $w^C(h) = (1-\alpha)^{1-\gamma} \cdot h$. For uncertified workers, employers pay expected productivity:

$$w^{NC} = (1-\alpha)^{1-\gamma} \cdot \mathbb{E}[h | h < h^*] = (1-\alpha)^{1-\gamma} \cdot \frac{\int_0^{h^*} s\, dG(s)}{G(h^*)}$$

Certification is individually rational for worker with skill h if $(1-\alpha)^{1-\gamma} h - \kappa \geq w^{NC}$, i.e., $h \geq h^*$ where h^* solves $(1-\alpha)^{1-\gamma} h^* - \kappa = w^{NC}(h^*)$. This fixed point exists and is unique under standard regularity conditions on G.

Part (ii): In the absence of certification, wages equal $w = (1-\alpha)^{1-\gamma} \cdot \mathbb{E}[h]$ for all workers. With certification, $w^C(h) = (1-\alpha)^{1-\gamma} \cdot h$, so high-skill workers reveal type and earn $(1-\alpha)^{1-\gamma} h > (1-\alpha)^{1-\gamma}\mathbb{E}[h]$ when $h > \mathbb{E}[h]$. The return to skill investment increases because skill becomes observable.

Part (iii): Private return to skill with certification is $\partial w^C/\partial h = (1-\alpha)^{1-\gamma} > 0$. Without certification, wages pool across unobservable skill levels: $\partial w/\partial h = 0$ since all uncertified workers receive the same pooled wage regardless of their true h. The higher private return under certification induces more skill investment, partially offsetting AI-induced atrophy. □

Corollary (Certification as Partial Remedy).

Certification increases the private return to skill by making skill observable, but does not affect the externality: each firm still ignores how its workers' skills benefit other firms through spillovers (θH^η) and learning spillovers ($\psi(H)$). The social FOC includes $\partial W/\partial H' \cdot \partial H'/\partial \alpha$, which exceeds the private marginal cost whether or not certification exists. Hence $\alpha^D > \alpha^S$ persists, though the gap may narrow. □

Proposition 12 (Optimal AI Design).

The welfare-maximizing AI designer solves:

$$\max_{\mu} W(\mu) = \sum_{t=0}^{\infty} \beta^t Y(h_t(\mu), \alpha^*(h_t, \mu))$$

subject to the equilibrium skill dynamics $h_{t+1} = (1-\delta)h_t + \lambda \ell(\alpha^*(h_t,\mu))\varphi(h_t)$.

Part (i): Differentiating: $\frac{dW}{d\mu} = \sum_t \beta^t \left[\frac{\partial Y}{\partial h}\frac{\partial h_t}{\partial \mu} + \frac{\partial Y}{\partial \alpha}\frac{\partial \alpha^*}{\partial \mu}\right]$. From Corollary 1, $\partial \alpha^*/\partial \mu > 0$ and $\partial h^*/\partial \mu > 0$ when $\mu < 1$. Both effects work in the same direction: higher μ is welfare-improving.

Part (ii): Private firm i maximizes $\pi_i = Y_i - c(\mu)$ where $c(\mu)$ is the cost of designing high-μ AI. The FOC is $\partial Y_i/\partial \mu = c'(\mu)$. Since $\partial Y/\partial \mu > 0$, firms do choose positive μ, but they ignore the externality on aggregate human capital. The social planner's FOC includes $\partial W/\partial H \cdot \partial H/\partial \mu > \partial Y_i/\partial \mu$, implying $\mu^S > \mu^D$.

Part (iii): The result $\mu^S > \mu^D$ (Socratic AI is socially preferred to Autocomplete) reflects underweighting of future skill development by users. Equivalently: if we define "pedagogical quality" as μ (higher is better for learning), then $\mu^S > \mu^D$ – the social optimum has higher pedagogical quality than the decentralized choice. Alternatively, if one defines "ease of use" as $\phi = 1 - \mu$ (higher ϕ means easier interface), then $\phi^S < \phi^D$ – the socially optimal AI is *less easy* (more pedagogically demanding) than what users would choose. The intuition is clear: users underweight future skill development, so they accept interfaces that are too easy and insufficiently pedagogical. The social optimum imposes more "productive friction" than users would choose voluntarily. □

Proposition 13 (Optimal AI Tax).

The social planner's problem is:

$$W(H, A) = \max_{\alpha} \{Y(H, \alpha; A) + \theta H^\eta + \beta W(H', A')\}$$

subject to $H' = (1-\delta)H + \lambda[1-(1-\mu)\alpha]\varphi(H)\psi(H)$ and $A' = (1-\zeta)A + \zeta Q(\alpha, H)$. Note that the learning spillover $\psi(H)$ enters the human capital transition, and the AI quality transition reflects endogenous data quality.

The social FOC is:

$$Y_\alpha + \beta \frac{\partial W}{\partial H'} \cdot \frac{\partial H'}{\partial \alpha} + \beta \frac{\partial W}{\partial A'} \cdot \frac{\partial A'}{\partial \alpha} = 0$$

Since $\frac{\partial H'}{\partial \alpha} = \lambda(\mu - 1)\varphi(H)\psi(H) < 0$ when $\mu < 1$ and $\frac{\partial A'}{\partial \alpha} = \zeta Q_\alpha < 0$, both dynamic costs enter negatively. Rearranging:

$$Y_\alpha = \beta \frac{\partial W}{\partial H'} \cdot \lambda(1-\mu)\varphi(H)\psi(H) + \beta \frac{\partial W}{\partial A'} \cdot \zeta |Q_\alpha|$$

The left side is the marginal output benefit. The right side sums the marginal costs through human capital ($\frac{\partial W}{\partial H'} > 0$, so the first term is positive when $\mu < 1$) and AI quality ($\frac{\partial W}{\partial A'} > 0$, so the second term is positive when adoption degrades training data).

The private FOC is $Y_\alpha = \beta V'(h')\lambda(1-\mu)\varphi(h)\psi(H)$, which ignores spillovers (θH^η) and AI quality effects.

The optimal tax τ^* equates private and social marginal costs:

$$\tau^* = \underbrace{\beta \frac{\partial W}{\partial H'} \lambda(1-\mu)\varphi(H)\psi(H) - \beta V'(h')\lambda(1-\mu)\varphi(h)\psi(H)}_{\text{HC externality}} + \underbrace{\beta \frac{\partial W}{\partial A'} \zeta \left| \frac{\partial Q}{\partial \alpha} \right|}_{\text{Training data externality}}$$

The first component captures the difference between social and private valuation of human capital (arising from spillovers). The second captures the training data effect, which firms ignore entirely.

Corrective feedback: As α increases, H falls (in the substitution regime). With $\theta > 0$, $\frac{\partial W}{\partial H}$ is increasing in the spillover contribution, which rises as H falls (scarcity increases marginal value). Thus τ^* rises with α. □

Competitive Overadoption (Appendix Result).

Consider a symmetric duopoly with firms A and B. Firm i's payoff is $\pi_i = s_i(\alpha_i, \alpha_j) \cdot \Pi(Y_i, Y_j) - c(\alpha_i)$, where $s_i = Y_i/(Y_i + Y_j)$ is market share, Π is total industry profit, and $c(\alpha)$ captures the human capital cost of adoption.

Firm i's FOC:

$$\frac{\partial s_i}{\partial \alpha_i}\Pi + s_i \frac{\partial \Pi}{\partial \alpha_i} = c'(\alpha_i)$$

The first term, $\frac{\partial s_i}{\partial \alpha_i}\Pi > 0$, represents the competitive motive: higher adoption steals market share from the rival.

A joint-profit maximizer chooses α^M to maximize total profits net of costs: $\max_\alpha[\Pi(\alpha, \alpha) - 2c(\alpha)]$ subject to both firms adopting identically. The FOC is $\frac{\partial \Pi}{\partial \alpha} = 2c'(\alpha)$, which at symmetric adoption simplifies to $\frac{1}{2}\frac{\partial \Pi}{\partial \alpha} = c'(\alpha)$ per firm. This omits the competitive term $\frac{\partial s_i}{\partial \alpha_i}\Pi$ because the joint maximizer internalizes that market share gains are zero-sum.

Since $\frac{\partial s_i}{\partial \alpha_i}\Pi > 0$ at any symmetric equilibrium, Nash equilibrium adoption α^N satisfies a FOC with a larger LHS than joint maximization, implying $\alpha^N > \alpha^M$.

Part (ii): The competitive term $\frac{\partial s_i}{\partial \alpha_i}\Pi$ is proportional to $\frac{\partial s_i}{\partial \alpha_i}$. With $s_i = Y_i/(Y_i + Y_j)$, we have:

$$\frac{\partial s_i}{\partial \alpha_i} = \frac{Y_j \cdot \frac{\partial Y_i}{\partial \alpha_i}}{(Y_i + Y_j)^2}$$

The magnitude of this business-stealing term varies through $\frac{\partial Y_i}{\partial \alpha_i}$ and the output levels, not through a separate competition-intensity parameter (which is fixed at $\varepsilon = 0.5$ at the symmetric equilibrium under this specification). Higher $\frac{\partial Y_i}{\partial \alpha_i}$ (greater productivity gain from adoption) increases the competitive motive, widening the gap $\alpha^N - \alpha^M$.

Part (iii): With human capital spillovers, firm i's human capital accumulation depends on aggregate H: $h_{i,t+1} = (1-\delta)h_{i,t} + \lambda\ell(\alpha_i)\varphi(h_i)\psi(H)$. When firm j adopts heavily, H falls, which reduces $\psi(H)$ and impairs firm i's skill accumulation even if i restrains.

The total externality combines: (a) the spillover externality (each firm's adoption degrades the skill ecosystem for others); and (b) the competitive externality (each firm's adoption steals market share). When both operate, firm i adopts heavily both because it undervalues human capital (spillover) and because restraint loses market share (competition). The effects compound because higher adoption by j both harms i's workers and forces i to match adoption to survive.

Formally, let α^S denote the social optimum, α^D the decentralized (single-firm) solution ignoring competition, and α^N the competitive Nash equilibrium. By definition:

$$\alpha^N - \alpha^S = (\alpha^D - \alpha^S) + (\alpha^N - \alpha^D) \equiv \Delta^{spill} + \Delta^{comp}$$

This is an identity, not a behavioral claim. The economic content is that $\Delta^{spill} > 0$ (spillover distortion) and $\Delta^{comp} > 0$ (competitive distortion), both pushing toward overadoption.

The distortions *interact* in that neither can be computed in isolation: α^D depends on the skill level that prevails under spillovers, and α^N depends on both spillovers and competitive dynamics. To formalize interaction, define counterfactual benchmarks: let $\alpha^{spill-only}$ be the equilibrium with spillovers but no competition (single firm or coordinated adoption), and $\alpha^{comp-only}$ be the equilibrium with competition but no spillovers. Then:

$$\alpha^N - \alpha^S > (\alpha^{spill-only} - \alpha^S) + (\alpha^{comp-only} - \alpha^S)$$

The excess reflects the *interaction*: spillover-induced skill degradation from high α_j makes firm i's workers less productive, increasing i's incentive to rely on AI, which amplifies i's competitive adoption. □

Proposition 14 (Feedback Loop Stability).

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

Part (ii): For uniqueness, note that the joint steady-state conditions define a continuous map. The H steady-state locus is downward-sloping in (H, A) space (higher A induces more adoption, which lowers steady-state H), while the A steady-state locus $A = Q(H, \alpha^*(H, A))$

is upward-sloping (higher H improves training data quality). The single crossing implies a unique intersection. For local stability, let (H^*, A^*) be a steady state. Consider perturbation $(H^* + \epsilon, A^* + \delta)$. The dynamics are:

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

where the Jacobian entries are:

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

The characteristic polynomial is $\lambda^2 - (J_{11} + J_{22})\lambda + (J_{11}J_{22} - J_{12}J_{21}) = 0$. For stability, we verify the Schur conditions: (i) $|\det J| < 1$; (ii) $|\text{tr}J| < 1 + \det J$. Under Assumption 8(ii), $J_{11} < 1$. Since $J_{12} < 0$ and $J_{21} > 0$, the off-diagonal product $-J_{12}J_{21} > 0$ raises the determinant; this relaxes the trace inequality (ii) but tightens requirement (i). Stability ultimately follows from ζ being small: when ζ is small, $J_{22} \approx 1 - \zeta$, $J_{21} \approx 0$, and the eigenvalues are approximately J_{11} and $1 - \zeta$, both with modulus less than 1. The slow AI adjustment rate ensures the feedback loop does not destabilize the system.

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