Skill Atrophy and AI Productivity Measurement

Tommaso Bondi* Gentry Johnson†

January 21, 2026

Abstract

How should we measure the productivity effects of generative AI? Recent experimental studies document substantial short-run improvements. 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 Ron Berman, Fabrizio Dell'Acqua, Sachin Gupta, Brett Hollenbeck, Vrinda Kadiyali, Jura Liaukonytė, Xueming Luo, Emaad Manzoor, Omid Rafieian, and Nathan Yang for helpful comments and discussions. All errors remain our own.

1 Introduction

Generative AI has delivered striking short-run productivity gains across a wide range of knowledge-intensive tasks. Customer service agents resolve more tickets per hour, consultants complete analyses faster, and junior developers ship code more quickly. These gains are especially pronounced for less-skilled workers, compressing the productivity distribution and raising average output – precisely in the tasks most central to early-career skill formation.

However, these estimates are – naturally – almost exclusively short-run. They measure output over weeks or months, not over the years or careers across which expertise is formed. This distinction matters because the tasks at which generative AI excels are often the very tasks through which humans develop skill. Junior developers build debugging ability by wrestling with broken code; legal associates develop judgment by drafting arguments from scratch; medical residents acquire diagnostic intuition by working through cases. When AI performs these formative tasks, immediate output rises. But if AI substitutes for the cognitive effort through which expertise develops, skill accumulation may slow.

This creates a fundamental measurement problem for evaluating generative AI over the longer term. When skill is endogenous to past technology use, productivity is no longer a sufficient statistic for long-run performance or welfare. A technology can raise output today while altering the state variable – human capital – that governs future productivity. In such environments, state-conditional productivity gains can coexist with declines in lifetime output. In the limit, cross-sectional comparisons can even have the wrong sign, showing gains precisely where long-run losses are largest.

Recent empirical work highlights the short-run side of this tradeoff. Brynjolfsson et al. (2025a) document average productivity gains of 14% in customer service, with effects exceeding 30% for novice workers. Dell'Acqua et al. (2023) report gains of roughly 40% for consultants on tasks within AI's capability frontier.[1] A consistent finding across these studies is that generative AI disproportionately benefits less-skilled workers, compressing the productivity distribution. These studies treat worker ability as fixed, and are therefore not designed to detect effects that operate through AI-induced, slower-moving changes in skill formation.

This paper develops a framework for understanding when and why short-run productivity gains from generative AI can diverge from long-run outcomes. We formalize a dynamic model with two mechanisms. First, *skill formation through learning-by-doing*: workers accumulate human capital by performing tasks, and delegating those tasks to AI reduces learning. The magnitude of this reduction depends on what we call the *pedagogical quality* of AI, denoted by μ. When AI provides answers directly ($\mu < 1$), it substitutes for cognitive effort and slows skill accumulation; when AI functions as a tutor that requires engagement ($\mu > 1$), it can instead augment learning.

Second, *training data degradation*: because generative AI is trained on human-generated content, reduced human practice degrades not only worker skills but also the future quality of AI itself.[2] We emphasize that this second channel – which is distinctive to generative AI

[1] Adoption has been rapid: 84% of developers now use or plan to use AI coding tools (Stack Overflow Developer Survey 2025).

[2] See Xu et al. (2024) for formal results on the irreducibility of hallucination in autoregressive language models.

– amplifies the magnitude of our findings, but is not required for them.

Our analysis delivers three sets of results. First, we identify two sources of mismeasurement in empirical estimates of AI productivity. One arises from *spillover bias*, which grows with industry-level AI saturation. As adoption spreads, non-users face degraded learning environments – reduced mentorship, weaker peer learning, and curricula adapted to AI-assisted workflows. Comparing AI users to these degraded non-users overstates the benefits of AI adoption. While negligible in early studies, this bias becomes substantial in high-adoption sectors.

The *state-path divergence* operates on a different margin: even comparing an AI user to that same user without AI overstates gains, because the user's current skill level reflects cumulative past adoption.[3] As skills atrophy, AI appears increasingly indispensable – not because it has become more productive, but because the outside option has deteriorated.

Second, we characterize when these biases are large versus small. The biases are largest when AI strongly substitutes for learning (low pedagogical quality), when learning-by-doing is central to skill formation, and when knowledge spillovers across workers are strong. They can reverse sign when AI augments learning ($\mu > 1$): AI users then accumulate skills faster, so comparing them to non-users would understate benefits. These results provide guidance for interpreting existing empirical estimates.

Third, we analyze welfare and policy. When human capital generates spillovers beyond its private value to the firm, decentralized AI adoption exceeds the social optimum. We characterize optimal corrective policies and analyze training mandates as a practical alternative to Pigouvian taxation. A key implication is that optimal policy may reduce measured productivity while improving welfare, creating a tension between efficiency and observable performance metrics.

Recent evidence supports the core mechanism and maps cleanly onto the two sources of bias we formalize. A recent randomized controlled trial (METR, 2025) finds that experienced developers are 19% slower when using AI tools, yet believe AI increases their productivity by 20% – a perception–reality gap consistent with state-path divergence: conditioning on current (atrophied) skill makes AI appear indispensable. Ethnographic work on robotic surgery documents how automation can render junior workers "completely optional," undermining the expert–novice relationships through which skill transfers across generations (Beane, 2019, 2024), consistent with spillover effects. Large-scale employment data tell a similar story: Brynjolfsson et al. (2025b) find that since late 2022, employment for workers aged 22–25 in AI-exposed occupations has declined by 13 percent relative to less-exposed fields, with declines concentrated in occupations where AI substitutes rather than augments human effort.

What is missing from this conversation is a formal framework that identifies *when* these forces dominate, *why* standard productivity measurement can fail, and *what* policy responses are warranted. Proposition 5 shows that spillover bias requires skill externalities across workers; absent such externalities, cross-sectional estimates are unbiased. More fundamentally, Proposition 6 formalizes the state-path divergence: even with a structurally correct

[3]This is not an omitted-variable or missing-state problem. Even with a correctly estimated production function and fully observed skill, conditioning on current skill leads to the wrong welfare ranking across technology paths. The bias arises because skill is endogenous to past adoption, not because skill is unobserved.

model and perfect observability of skill, comparing productivity along different technology paths overstates gains because current skill is itself endogenous to past adoption. The implication – that individually welfare-reducing technologies can appear indispensable under state-conditional measures – has not, to our knowledge, been formalized previously.

The paper proceeds as follows. Section 1.1 reviews the literature. Section 2 develops the baseline model of AI adoption with learning-by-doing. Section 3 characterizes equilibrium adoption and its dependence on key parameters. Section 4 analyzes the measurement problem, including both spillover bias and state-path divergence, and discusses implications for empirical research. Section 5 examines welfare and policy. Section 6 concludes. Appendix A develops extensions including heterogeneous workers, cohort effects, firm selection, and optimal policy instruments. Appendix B contains proofs.

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: most task frameworks treat skills as a stock that determines task 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. Eloundou et al. (2024) estimate 80% of the U.S. workforce could have at least 10% of tasks affected by LLMs; Acemoglu (2024) estimates modest TFP gains of 0.5–0.7% over ten years – both assuming no skill atrophy. Agrawal et al. (2018, 2019) characterize AI as reducing the cost of prediction and emphasize complementarities with human judgment; our framework identifies a tension – AI may complement the *use* of existing judgment while substituting for its *development*.

A growing empirical literature documents short-run productivity effects: Noy and Zhang (2023) find effects for writing tasks; Peng et al. (2023) document faster task completion with coding assistants; Dell'Acqua et al. (2023) identify a "jagged technological frontier" where AI helps on some tasks but hurts on others. Most relevant, Bastani et al. (2025) find that GPT-4 access *harms* educational outcomes – students use AI as a "crutch" and perform worse on subsequent assessments – but pedagogically-designed tutors mitigate this harm, providing direct support for our low-μ/high-μ distinction. Gaessler and Piezunka (2023) find chess computers *helped* players improve by substituting for scarce human partners ($\mu \geq 1$), though players failed to learn to exploit idiosyncratic human mistakes. More recent work documents deskilling: Budzyń et al. (2025) find endoscopists become less accurate after three months of AI assistance; Lee et al. (2025) document reduced critical thinking among AI-reliant workers; Dell'Acqua (2022) find recruiters with high-quality AI becoming less attentive. Ong and Png (2023) study a distinct phenomenon: automation deskilling *jobs* rather than workers, increasing labor supply by lowering work disutility.

The welfare implications of AI extend beyond labor productivity. Goldberg and Lam (2025) show human creators may exit creative goods markets even when their work is higher quality. Luo et al. (2025) find platforms may optimally restrict AI access to preserve human capital. Athey and Scott Morton (2025) examines welfare effects of AI market power; we share their concern with welfare-relevant counterfactuals but focus on skill dynamics. Our model builds on learning-by-doing (Arrow, 1962; Lucas, 1988) and learning curves (Thomp-

son, 2010).[4] We extend Arrow's insight that production generates knowledge as a byproduct to show AI can sever this link. Our welfare analysis relates to path dependence in technology adoption (David, 1985).

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 juniors to acquire tacit knowledge, generating socially excessive automation. Ide and Talamàs (2025) distinguish autonomous AI from non-autonomous AI (copilots) – with implications for whether AI substitutes for or augments learning. Beane (2019) provides ethnographic evidence: robotic surgery made trainees "optional," reducing hands-on practice tenfold. Our contribution is distinct: we study learning *within* jobs rather than access *to* jobs or the financing of training. The mechanisms compound – policies preserving training financing or 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; Dohmatob et al., 2024; Bertrand et al., 2024). del Rio-Chanona et al. (2024) find Stack Overflow activity declined 25% within six months of ChatGPT's release; Burtch et al. (2024) show newer users were most likely to exit. These platforms provide both training data for LLMs and mentorship networks – when newcomers exit, they reduce fresh training data *and* degrade the peer-learning environment, amplifying both sources of inefficiency 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, α) 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 $g(0) = 1$, $g(1) = \underline{g} \in (0, 1)$, 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 $[0, \alpha]$

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

to AI while the worker performs tasks in $(\alpha, 1]$. Standard optimization shows the worker spreads effort uniformly across performed tasks, yielding worker output $h(1-\alpha)^{1-\gamma}$.[5]

Substituting, period output takes the tractable form

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

where $G(\alpha) \equiv \int_0^\alpha g(j)\, dj$ represents cumulative AI output capacity, so $G'(\alpha) = g(\alpha)$ and $G''(\alpha) = g'(\alpha) < 0$. The first term captures AI's contribution, increasing in adoption intensity. The second term captures the worker's contribution, depending on human capital and tasks performed. The exponent $1-\gamma < 1$ reflects that when the worker performs fewer tasks, effort can be concentrated more effectively. The output function is linear in h, strictly concave in α for $h > 0$, and satisfies $\partial Y/\partial \alpha \to -\infty$ as $\alpha \to 1^-$, ensuring interior optima.

2.2 Human Capital Dynamics

The central innovation of our model is the specification of human capital dynamics that allows AI to either substitute for or augment learning. Human capital evolves according to the law of motion

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

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

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

where $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is a function of current human capital satisfying regularity conditions detailed below, and $\mu \geq -1$ is the *pedagogical quality* of AI.

The effective learning rate $\ell(\alpha) \equiv 1 - (1-\mu)\alpha$ must be non-negative for human capital to remain in $\mathbb{R}_+$. When $\mu \geq 0$, this holds for all $\alpha \in [0,1]$. When $\mu \in (-1, 0)$, we restrict attention to $\alpha < 1/(1-\mu)$. Our main results focus on $\mu \in [0, 1)$, the empirically relevant substitution regime; throughout, we maintain $\ell(\alpha) \geq 0$ so the law of motion (2) is well-defined.

This specification applies to the baseline model of Sections 2–3. In Section 5, we augment the learning function with aggregate human capital dependence: $L_i = [(1-\alpha_i) + \mu\alpha_i] \cdot \varphi(h_i) \cdot \psi(H)$, where $\psi(H)$ captures learning spillovers.

The pedagogical quality parameter μ determines when skill atrophy occurs. When $\mu < 0$, AI actively undermines learning – perhaps because AI-generated solutions obscure the reasoning that develops understanding. When $\mu = 0$, AI is pedagogically neutral: learning occurs only through tasks the worker performs. When $\mu \in (0, 1)$, AI partially augments learning: AI-performed tasks contribute to skill development, but less than worker-performed tasks. When $\mu \geq 1$, AI fully augments learning – perhaps because AI provides immediate feedback or serves as an intelligent tutor.

[5] 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 $\int_\alpha^1 h[1/(1-\alpha)]^\gamma dj = h(1-\alpha)^{1-\gamma}$.

Assumption 1 (Learning Capacity Function)**.** The function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ governing learning capacity exhibits diminishing returns: φ is twice continuously differentiable, strictly positive, bounded above, strictly decreasing ($\varphi'(h) < 0$ for all $h > 0$), and satisfies $\lim_{h\to\infty} \varphi(h) = 0$. These properties capture that learning-by-doing becomes less productive as expertise accumulates – deep experts face diminishing returns as most relevant knowledge has already been acquired.[6]

The key property of the learning function follows directly: $\partial L/\partial \alpha = (\mu - 1)\varphi(h)$, which is negative when $\mu < 1$, zero when $\mu = 1$, and positive when $\mu > 1$. This derivative governs whether AI adoption helps or hurts skill accumulation.

Remark 1 (Microfoundation for μ). The parameter μ can be derived from cognitive effort allocation. When AI handles a task, the worker's cognitive effort falls, freeing capacity for other uses. Let $\xi \in [0, 1]$ denote the fraction of freed effort allocated to learning-oriented activities (reviewing AI output, deliberate practice). Then $\mu = \xi$. If workers choose ξ to maximize immediate utility, and learning effort is costly, then $\xi^* < 1$ – explaining why $\mu < 1$ emerges endogenously. This cognitive-effort story microfounds $\mu \in [0, 1)$; the cases $\mu < 0$ (AI actively undermines learning) and $\mu > 1$ (AI enhances learning beyond what unaided practice provides) require different mechanisms, such as AI obscuring reasoning or serving as an intelligent tutor. We focus on $\mu \in [0, 1)$ in the baseline analysis; $\mu > 1$ and $\mu < 0$ appear only as reduced-form extensions capturing tutoring or obfuscation effects.

The parameter μ has clear empirical content.[7] We treat μ as exogenous in the baseline model. A limitation is that μ is really an equilibrium object – it depends on AI design, workplace norms, and user incentives – rather than a technological primitive. Competitive pressure exacerbates low-μ outcomes: firms that design AI for maximum immediate productivity (low μ) outperform rivals in the short run, even if this degrades long-run skill formation. Appendix A analyzes how this competitive dynamic shapes μ in equilibrium.

Settings where $\mu < 1$ is most likely include: (i) junior professional training, where formative struggle builds judgment; (ii) autocomplete-heavy workflows, where AI provides answers rather than scaffolding problem-solving; (iii) time-pressured environments, where users lack incentive to engage deeply with AI output. Settings where $\mu \geq 1$ may apply include: (i) AI-as-tutor applications explicitly designed to require user engagement; (ii) expert users who already possess the judgment to learn from AI suggestions; (iii) tasks where AI feedback accelerates learning.

Remark 2 (Heterogeneous μ). Our baseline treats μ as constant across tasks and time. In practice, μ likely varies: AI may be pedagogically valuable early in a career (when scaffolding helps) and substitutive later (when it crowds out refinement of judgment), or vice versa. Similarly, μ may vary across tasks within a period – a junior lawyer might use AI heavily for document review (low μ) but not for client interaction (high μ) – in which case the relevant μ is the adoption-weighted average across the task distribution. This heterogeneity would

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

[7] Evidence suggests μ varies across contexts. Bastani et al. (2025) show GPT-4 access harms learning ($\mu < 1$), but pedagogically-designed tutors mitigate this (higher μ). Dell'Acqua (2022) documents reduced effort with AI (low μ); Brynjolfsson et al. (2025a) find AI helping workers "move down the experience curve" (higher μ). The human factors literature documents related "automation complacency" effects (Parasuraman and Riley, 1997; Endsley, 2017).

strengthen our mismeasurement results by introducing additional path dependence. Workers who happened to use AI during low-μ phases would accumulate less skill than those who used it during high-μ phases, even holding total AI exposure constant. The scalar μ in our model can be interpreted as an adoption-weighted average over tasks and time; Appendix A.8 verifies that allowing $\mu(h)$ to vary with skill level does not qualitatively change results.

Table 1: Notation Guide

Symbol	Definition
h, H	Individual / aggregate human capital
h_t^U, h_t^{NU}, h_t^{NA}	Human capital at t for: users, non-users, no-adoption counterfactual
α	AI adoption intensity (fraction of tasks delegated to AI)
α^*	Optimal individual adoption (solves firm's FOC)
$\bar{\alpha}$	Aggregate/average adoption across firms
A	AI productivity level
$g(j)$, $G(\alpha)$	AI capability at task j / cumulative AI output: $G(\alpha) = \int_0^\alpha g(s)\,ds$
γ	Effort concentration parameter (diminishing returns)
μ	Pedagogical quality of AI (< 1: substitutes; ≥ 1: augments)
$\ell(\alpha)$	Effective learning rate: $\ell(\alpha) = 1 - (1-\mu)\alpha$
δ	Depreciation rate of human capital
λ	Learning intensity parameter
$\varphi(h)$	Learning capacity function
β	Discount factor
$\bar{h}$	No-adoption steady-state human capital: $\delta\bar{h} = \lambda\varphi(\bar{h})$
$h^*(\alpha)$	Steady-state human capital at adoption level α
h^{NU*}	Steady-state human capital of non-users (with spillover degradation)
θ	Human capital output spillover parameter
η	Curvature of output spillovers
$\psi(H)$	Learning spillover function (normalized: $\psi(\bar{H}) = 1$)
ζ	AI quality adjustment speed
$V(h)$, $W(H)$	Private / social value function
τ^*	Optimal Pigouvian tax
ρ	Mandatory training requirement

2.3 The Firm's Dynamic Problem

Firms maximize the present discounted value of output, taking as given the initial human capital stock and the path of AI productivity. Each period, the firm chooses adoption intensity α_t to balance current output against future skill accumulation. The discount factor $\beta \in (0, 1)$ governs how much weight firms place on future productivity relative to current output. Patient firms (high β) internalize skill costs more heavily; impatient firms prioritize immediate gains.

Formally, 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 that V exists, is unique, and is strictly increasing and concave in h.[8]

The key trade-off is dynamic: higher adoption today raises current output but – when $\mu < 1$ – reduces future human capital, which in turn reduces future output capacity.

Throughout, we impose several standard conditions. First, we assume competitive labor markets: firms cannot perfectly contract on workers' future skill levels or retain workers indefinitely, so they do not fully internalize human capital spillovers (discussed further in Section 5). Second, we impose a transversality condition ensuring bounded discounted value, which holds automatically given bounded human capital. Third, AI adoption affects output contemporaneously but affects skill accumulation only through the next-period transition. These assumptions are standard and relaxed in Appendix A.

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 1), 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.[9]

Assumption 2 (AI Productivity)**.** AI is sufficiently productive that adoption is attractive at the no-adoption steady state: $A \cdot g(0) > \bar{h}(1-\gamma)$, where $\bar{h}$ is the steady-state human capital without AI.

The following proposition characterizes how pedagogical quality shapes adoption:

[8] Existence and uniqueness follow from Stokey and Lucas (1989); human capital is bounded above by $\bar{h}$ (the no-adoption steady state), ensuring the problem is well-behaved. Supporting lemmas appear in Appendix B.

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

Proposition 1 (Role of Pedagogical Quality)**.** *The firm's optimal adoption $\alpha^*(h)$ is interior and depends critically on pedagogical quality μ:*

(i) When $\mu < 1$, adoption generates a dynamic skill cost. Optimal adoption is locally strictly increasing in μ around the stable steady state: better pedagogical quality permits more adoption without triggering skill atrophy.

(ii) When $\mu = 1$, AI is pedagogically neutral. Adoption is determined purely by the static output trade-off, with no dynamic considerations.

(iii) When $\mu > 1$, adoption generates a dynamic skill benefit. Firms may adopt more *than the static optimum because AI-assisted work enhances learning.*

Proof. The proof for this and all other results can be found in Appendix B. □

The proposition captures a fundamental asymmetry. In the substitution regime ($\mu < 1$), firms face a genuine intertemporal trade-off: higher adoption raises current output but reduces future productivity by impairing skill development. Forward-looking firms internalize this cost and adopt less than a myopic firm would. This connects to the "deskilling" literature in labor process theory (Braverman, 1974), but our framework allows for the opposite: when $\mu > 1$, AI enhances skill formation. Unlike prior work on automation that focuses 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.

This raises an obvious question: if patient firms internalize skill costs, why doesn't the market simply select for patience, or why don't firms commit to skill-preserving adoption levels? The answer is that several natural corrective mechanisms fail or backfire. Patience helps individually but not collectively: patient firms restrain adoption, but this restraint is competitively punished in the short run, and selection can eliminate patient firms before their strategy pays off (Proposition 14). Spillovers make things worse, not better: when human capital generates externalities, each firm's restraint benefits competitors, so the private return to patience understates the social return. And the measurement problem we identify in Section 4 means that even a social planner relying on standard productivity metrics would perceive skill-preserving policies as costly. The trap persists not because agents are irrational, but 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 $\ell(\alpha)$ is the effective learning rate defined in Table 1. We impose standard regularity conditions ensuring interior, stable steady states.[10]

[10]These are standard technical requirements: that steady states are interior, that local stability holds, that static curvature dominates dynamic terms in the FOC, and that the policy function is monotone. These ensure the value function is well-behaved and that comparative statics have unambiguous signs. They hold for generic parameter values; Appendix B states them formally.

Lemma 1 (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 3 (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:

Proposition 2 (Steady-State Characterization)**.** *The structure of steady-state equilibria depends critically on pedagogical quality:*

(i) ***Substitution regime*** *($\mu < 1$): Steady-state human capital satisfies $h^* < \bar{h}$ for any interior adoption $\alpha^* > 0$. AI adoption degrades skills.*

(ii) ***Augmentation regime*** *($\mu \geq 1$): Steady-state human capital satisfies $h^* \geq \bar{h}$. AI adoption does not reduce – and may increase – skills.*

This dichotomy has a sharp implication: skill atrophy *requires* that AI substitute for learning. If $\mu \geq 1$, then $h^* \geq \bar{h}$ – skills cannot fall below the no-adoption benchmark. Similarly, steady-state output can fall below $\bar{h}$ only when $\mu < 1$. When AI augments learning, skill preservation combined with AI's direct productivity contribution ensures $Y^* > \bar{h}$. The empirical question of which regime applies is therefore first-order for policy.

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

Proposition 3 (Comparative Statics)**.** *At a stable interior steady state with $\mu < 1$:*

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

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

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

(iv) Better pedagogical quality raises both skills and adoption: $\partial h^*/\partial \mu > 0$, $\partial \alpha^*/\partial \mu > 0$.

Result (i) echoes Acemoglu and Restrepo (2018): better automation technology increases automation. But unlike their framework where human capital is fixed, here the increased automation *endogenously degrades* the human capital stock. Result (ii) suggests that short-termism – whether from capital market pressure, managerial myopia, or high discount rates – exacerbates skill atrophy. Result (iv) is perhaps most policy-relevant: improving AI's pedagogical quality is doubly beneficial, raising both skills and (appropriately) adoption.[11]

[11] One might expect patient firms to gain long-run competitive advantage as their workers remain skilled. However, Appendix A shows the opposite: impatient firms gain market share in the short run, potentially driving out patient firms before their restraint pays off.

Remark 4 (Robustness to functional forms). The qualitative results – skill atrophy when $\mu < 1$, overadoption with spillovers, divergence between cross-sectional and long-run estimates – do not depend on the specific functional forms chosen. What matters is that learning-by-doing exhibits diminishing returns at high skill levels, that AI adoption reduces the rate of learning when $\mu < 1$, and that spillovers create a wedge between private and social returns to human capital. Appendix A.8 verifies robustness to alternative formulations.

4 Skill Atrophy and the Mismeasurement of AI Productivity

This section presents our main results on mismeasurement. We begin by characterizing the skill trap, then derive two distinct sources of bias in productivity measurement.

As a benchmark, consider what happens if skills are exogenous – fixed endowments unaffected by technology use. Cross-sectional productivity comparisons would then correctly measure welfare effects: AI users would outperform non-users by exactly the amount AI contributes, and this gap would persist indefinitely. Similarly, if learning occurred independently of task performance, or if human capital generated no spillovers across workers, standard empirical designs would recover the welfare-relevant treatment effect. Our results identify precisely which of these conditions must fail, and how, for mismeasurement to arise.

4.1 Definition and Characterization

We formally define the skill trap and characterize conditions for its existence.

Definition 1 (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 (since $G(0) = \int_0^0 g(j)\, dj = 0$, we have $Y(h, 0) = h$).

(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, not merely that their productivity grows more slowly.

The trap is individually rational: firms optimize at every date, yet the equilibrium path delivers lower long-run output than no adoption. This relates to the "competency trap" in organizational learning (Levinthal and March, 1993). A natural question is why sophisticated firms don't treat skill losses as depreciation and restrain adoption accordingly. Our model includes this channel: patient firms (β high) *do* adopt less. The trap arises because labor mobility prevents firms from fully capturing returns to worker skills, competitive pressure causes firms that restrain adoption to lose market share, and discounting means even rational firms weight short-run gains heavily.

Proposition 4 (Existence of the Skill Trap)**.** *Under Assumption 2 with initial condition $h_0 \leq \bar{h}$, the economy is in a skill trap if and only if (steady-state comparison):*[12]

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

(ii) $A \cdot G(1) < \bar{h}$ *(AI cannot match fully-skilled labor);*

(iii) $\beta < \bar{\beta}$*, where* $\bar{\beta} \in (0, 1)$ *is the unique solution to the fixed-point equation:*

$$\Psi(\bar{\beta}) \equiv A \cdot G(\alpha^*(\bar{\beta})) + h^*(\alpha^*(\bar{\beta}))[1 - \alpha^*(\bar{\beta})]^{1-\gamma} - \bar{h} = 0 \tag{7}$$

where $\alpha^*(\beta)$ is the steady-state adoption at discount factor β and $h^*(\alpha)$ is the steady-state human capital function from Lemma 1.[13]

The intuition is straightforward. AI is always adopted because its marginal benefit exceeds the marginal learning cost at $\alpha = 0$. Initial adoption raises output because skilled workers are still abundant. But steady-state output $Y^*(\beta)$ is strictly increasing in β: patient firms restrain adoption enough to maintain higher skills, and the skill benefit dominates. The threshold $\bar{\beta}$ is where steady-state output under adoption exactly equals the no-adoption benchmark.

One might ask why workers cannot "practice without AI" to maintain skills. In principle they can, but this is costly and firms may not permit it. Moreover, identifying *which* skills to practice requires judgment that may itself have atrophied – creating a second-order skill loss that compounds the first.

Remark 5 (Parameter regions)*.* The three conditions for the trap have different empirical status. Condition (i), $\mu < 1$, is the central assumption throughout our analysis; it asserts that AI substitutes for rather than augments learning. Condition (ii), $A \cdot G(1) < \bar{h}$, requires that even full AI delegation cannot match the productivity of a fully-skilled human worker – a condition that holds when human judgment retains value on complex tasks. Condition (iii) determines the patience threshold $\bar{\beta}$, which depends on all model parameters. The mismeasurement results (Propositions 5 and 6) require only condition (i) plus, for the spillover bias, the presence of human capital externalities. The trap existence result additionally requires conditions (ii) and (iii). Importantly, even outside the trap – when $\beta > \bar{\beta}$ or $A \cdot G(1) > \bar{h}$ – the mismeasurement results continue to hold: cross-sectional estimates still overstate long-run effects when spillovers are present, and state-conditional gains still diverge from path welfare. The trap sharpens these results by generating sign reversal, but the qualitative biases are more general.

4.2 Short-Run versus Long-Run Productivity Effects

A central empirical implication of our model concerns how productivity gains from AI are measured. The choice of counterfactual fundamentally determines whether AI adoption appears beneficial or harmful.

[12]The proof requires standard regularity conditions ensuring a stable interior steady state – specifically, that $h^* \in (0, \bar{h})$, that local stability holds, and that the policy function is monotone. These hold for generic parameter values; see Appendix B.

[13]Existence and uniqueness of $\bar{\beta}$ follow from the monotonicity of steady-state output in β, which holds unconditionally at stable interior steady states; see Appendix B.

Definition 2 (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 AI users to the hypothetical path where AI was never adopted: $\Delta_t^{LR} = Y(h_t^U, \alpha_t) - Y(h_t^{NA}, 0)$.

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 of whether AI raises or lowers productivity relative to a world without the technology. These counterfactuals diverge when aggregate AI adoption affects learning opportunities for non-users – through reduced mentorship, weaker knowledge spillovers, or degraded training institutions.

Both counterfactuals answer legitimate questions. The cross-sectional counterfactual answers: should an individual firm adopt AI given that competitors may also adopt? The long-run counterfactual answers: does AI adoption improve welfare relative to a world without AI? These questions have different answers when spillovers are present. Existing empirical work answers the first question; our contribution is to highlight when the answer to the second question differs.

One might expect that if AI users consistently outperform non-users – as documented in study after study – then AI must be raising aggregate welfare. The next result shows this intuition can fail, and fail badly: cross-sectional gains can coexist with long-run losses, so that the measured effect has the wrong sign.

Proposition 5 (Divergence Between Counterfactuals)**.** *Suppose learning spillovers are present:* $\psi'(H) > 0$*. Then:*

(i) Cross-sectional estimates exceed long-run estimates: $\Delta_t^{CS} > \Delta_t^{LR}$ *for all* $t > 0$*.*

(ii) The gap is zero at $t = 0$ *and strictly increasing in* t *as aggregate human capital* H_t *declines (i.e., as AI adoption spreads and skills atrophy).*

(iii) When the economy is in a skill trap (Proposition 4), sign reversal occurs: $\Delta_t^{CS} > 0 > \Delta_t^{LR}$ *for* t *sufficiently large.*

Importantly, learning spillovers ($\psi'(H) > 0$) drive the cross-sectional vs. long-run wedge by degrading non-users' skill accumulation. Output spillovers ($\theta > 0$) affect welfare levels but cancel in the cross-sectional comparison since both user and non-user experience the same aggregate H_t.

When does sign reversal occur? In the skill trap, it requires two conditions: AI users must outperform degraded non-users ($Y^* > h^{NU*}$, ensuring $\Delta^{CS} > 0$), and learning spillovers must degrade non-users ($\psi'(H) > 0$, which combined with $Y^* < \bar{h}$ ensures $\Delta^{LR} < 0$). Output spillovers reinforce the negative long-run effect but are not necessary for the cross-sectional vs. long-run wedge.

When does the divergence matter? The gap requires two conditions: AI negatively affects skill formation ($\mu < 1$), and non-users' learning depends on aggregate skill levels (spillovers). If either fails, cross-sectional estimates correctly measure long-run effects. The bias is likely largest in high-adoption sectors with strong mentorship traditions – software development, with its heavy reliance on code review and pair programming, is a plausible

example. Within-firm studies comparing coworkers are most affected; cross-industry comparisons least affected.

The bias we identify differs from standard econometric concerns. It is not a general equilibrium wage adjustment: we hold prices fixed and identify mismeasurement in physical productivity. It is not cohort selection: we compare counterfactual paths for the same workers, not different populations. And it is not omitted variable bias in the standard sense: even a randomized experiment correctly identifying causal effects would face this problem, because the treatment (AI access) changes the state variable (skill) against which future AI benefits are measured. In the taxonomy of Abbring and Heckman (2007) – developed for dynamic treatment effects in labor economics – our bias arises because the treatment affects the evolution of the state, not because of selection on unobservables into treatment.

4.3 Path Dependence and State-Path Divergence

The spillover bias analyzed above concerns how AI adoption by some workers degrades the counterfactual for others. We now turn to our second mismeasurement result, which operates even at the individual level and *does not require spillovers*: path dependence in human capital causes state-conditional productivity gains to diverge from welfare-relevant path comparisons. Even a perfectly isolated worker can face this bias.

To clarify the structure: the paper identifies two distinct wedges. The first (Sections 4.1–4.2) is the *cross-sectional vs. long-run wedge*, driven by spillovers that degrade non-users' skill accumulation. The second (this section) is the *state-conditional vs. path wedge*, driven by path dependence in an individual worker's human capital. The first requires spillovers ($\psi'(H) > 0$); the second requires only skill atrophy ($\mu < 1$). Both cause measured productivity gains to overstate welfare effects, but through different mechanisms.

Definition 3 (State-Conditional vs. Path Counterfactuals)**.** The *state-conditional counterfactual* holds human capital fixed: $\Delta_t^{SC} = Y(h_t^{user}, \alpha_t) - Y(h_t^{user}, 0)$. The *path counterfactual* compares lifetime output under adoption versus the no-adoption path: $\Delta^{PATH}(\tilde{\beta}) = \sum_{\tau=0}^{\infty} \tilde{\beta}^{\tau}[Y(h_\tau^{user}, \alpha_\tau) - Y(h_\tau^{NA}, 0)]$, where $\tilde{\beta}$ is the evaluator's discount factor.

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

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 naturally recover Δ^{SC}: the effect of turning AI "on" at a given skill level. In causal-inference terms, this is a controlled direct effect holding a post-treatment variable (skill) fixed; but when the treatment changes that variable, the welfare-relevant object is the total effect along the counterfactual path. The wedge between these objects is not an econometric failure – it is an equilibrium implication of endogenous human capital.

Crucially, the measurement problem we identify 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 current skill h_t^{user} rather than the counterfactual skill

h_t^{NA}. The firm's choice is privately optimal given h_t, but empirical measurement using Δ_t^{SC} conflates "valuable given current skill" with "valuable relative to never adopting."

Proposition 6 (State-Path Divergence)**.** *Suppose $\mu < 1$ (no spillovers required). Then:*

(i) ***Skill convergence:*** $\lim_{t\to\infty} h_t^{user} = h^* < \bar{h} = \lim_{t\to\infty} h_t^{NA}$.

(ii) ***Growing relative value:*** *The ratio Δ_t^{SC}/h_t^{user} is strictly increasing in t, and $\lim_{t\to\infty} \Delta_t^{SC}/h_t^{user}$ can be arbitrarily large when $h^*/\bar{h}$ is small.*

(iii) ***Indispensability:*** *In the skill trap, $Y^* < \bar{h}$ yet $\Delta_t^{SC} > 0$ for all t: AI appears indispensable even when it reduces long-run output.*

(iv) ***Welfare reversal:*** *For any $\tilde{\beta} > \bar{\beta}$, $\Delta^{PATH}(\tilde{\beta}) < 0$: under more patient evaluation, the adoption path is welfare-inferior.*

The result admits a natural interpretation. As skills atrophy, the worker's AI-independent productivity h_t^{user} falls. This makes AI appear more valuable – not because AI has become more productive, but because the outside option has deteriorated. Continued use is optimal given current state, even if initial adoption was welfare-reducing.[14]

State-path divergence is not an externality problem but a *measurement* problem. It arises even when the agent fully internalizes their own skill dynamics and optimizes perfectly. The firm in our model correctly accounts for how current adoption affects future human capital; patient firms restrain adoption precisely because they value future skills. The bias occurs because empirical comparisons condition on current skill h_t, treating it as exogenous when it is in fact shaped by past adoption. A structurally correct model that estimates the production function $Y(h, \alpha)$ and computes AI's marginal product will still overstate welfare gains, because the counterfactual "this worker without AI" uses the worker's current (atrophied) skill rather than the skill they would have developed on the no-adoption path.

This distinction matters for policy. Classic externalities call for Pigouvian corrections that align private and social incentives. State-path divergence calls for measurement corrections that use welfare-relevant counterfactuals. Both problems can coexist – our spillover bias *is* a classic externality requiring corrective policy – but state-path divergence would persist even in a world with no spillovers and perfectly aligned incentives.

The divergence is largest when AI substitutes strongly for learning, careers are long, and adoption is high. For policy, the implication is stark: a regulator considering AI restrictions will face inflated cost estimates, since workers whose skills have atrophied show large losses when measured against their current state rather than the path where skills never atrophied. Cohort comparisons and cross-country variation in adoption timing approximate the correct counterfactual more closely than state-conditional designs.

The two biases differ in important ways. The *spillover bias* is a cross-sectional error arising from externalities: it compares users to degraded non-users, requires skill spillovers ($\psi'(H) > 0$), and is characterized in Proposition 5. The *state-path divergence* is a longitudinal

[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.

error arising from path dependence: it compares a user to their own counterfactual, requires only skill atrophy ($\mu < 1$), and is characterized in Proposition 6.
The key distinction: even with zero spillovers, state-path divergence persists because each worker's skill is endogenous to their own past AI use. Conversely, spillover bias can arise even for workers who never adopt, if aggregate adoption degrades their learning environment.

To summarize the conceptual structure: the skill trap (Proposition 4) is an equilibrium characterization describing what happens; the spillover bias (Proposition 5) is both a measurement and an inefficiency result; and state-path divergence (Proposition 6) is a pure measurement result that operates even absent any externality. This last point is arguably our most novel contribution: standard measurement fails not because incentives are misaligned, but because conditioning on current skill treats an endogenous state as exogenous.

4.4 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. This creates a feedback loop that amplifies skill atrophy.

The training data mechanism is an amplification channel, not a necessary condition for our core results. The spillover bias (Proposition 5) requires human capital spillovers but not training data degradation. The state-path divergence (Proposition 6) requires only $\mu < 1$; it holds even with no spillovers and fixed AI quality. The feedback loop strengthens both results quantitatively, but the qualitative mismeasurement phenomena survive even if data curation fully mitigates model collapse.

The distinction from previous automation technologies is stark. Calculators and spreadsheet software operate via fixed algorithms; their accuracy does not depend on how much arithmetic humans have recently practiced. GPS navigation works identically whether or not drivers remember local streets – and while GPS may atrophy navigational skills, it does not *learn from* human navigation, so no feedback loop exists. Generative AI is fundamentally different: it learns from human output. If workers delegate more tasks to AI, they produce less original content, and the content they do produce may be lower quality. Both effects degrade training data for future AI systems.

We model AI productivity as evolving according to

$$A_{t+1} = (1 - \zeta)A_t + \zeta \cdot Q(H_t, \bar{\alpha}_t) \tag{8}$$

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 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]

[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.

Proposition 7 (The Skill-Data Feedback Loop)**.** *With endogenous AI quality, the system* (H_t, A_t) *exhibits a feedback loop:*

(i) Higher aggregate human capital produces better training data ($\partial A_{t+1}/\partial H_t > 0$*), while higher adoption degrades data quality (*$\partial A_{t+1}/\partial \bar{\alpha}_t < 0$*).*

(ii) The state-conditional value of AI, $Y(h, \alpha^*) - Y(h, 0)$*, remains positive even as A falls, because the outside option* $Y(h, 0) = h$ *deteriorates with skill atrophy.*

(iii) The joint steady state (H^{**}, A^{**}) *satisfies* $H^{**} < H^*(A_0)$*, where* $H^*(A_0)$ *is steady-state skill with exogenous AI quality* $A_0 = Q(\bar{H}, 0)$*: the feedback loop amplifies skill atrophy beyond what fixed AI quality would generate.*

The amplification in part (iii) arises because lower H degrades training data, which lowers A, which sustains high adoption via the dependence mechanism, which keeps H low. The feedback operates through the joint dynamics, not through partial-equilibrium comparative statics.

4.5 Implications for Empirical Research

Our analysis has direct implications for how AI's productivity effects should be measured. Table 2 summarizes which results require which assumptions; Table 3 maps empirical strategies to their bias exposure.

Table 2: Logical Dependence of Main Results

	$\mu < 1$	Spillovers	Feedback
Skill trap (Prop. 4)	Yes	No	No
Spillover bias (Prop. 5)	Yes	Yes	No
State-path divergence (Prop. 6)	Yes	No	No
Feedback amplification (Prop. 7)	Yes	No	Yes
Overadoption (Prop. 9)	Yes	Yes	No
Training data externality (Prop. 10)	Yes	No	Yes

The choice of research design fundamentally determines exposure to the biases we identify. Within-firm comparisons – including randomized controlled trials that assign AI access to some workers but not others – face maximum spillover bias when coworkers share mentorship networks and training resources. The bias is attenuated when comparing across industries (where spillovers are weaker) and minimal when comparing pre-AI to post-AI cohorts (which approximates the path counterfactual). Staggered adoption designs occupy an intermediate position: they control for time-invariant worker heterogeneity but remain vulnerable to spillover effects that operate within industries.

Existing evidence should be interpreted through the lens of our framework. The striking productivity gains documented in recent studies – 14% for customer service agents (Brynjolfsson et al., 2025a), 40% for consultants on frontier tasks (Dell'Acqua et al., 2023) – are not wrong, but they answer a different question than long-run welfare effects. These estimates

Table 3: 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
Cross-industry comparison	Low	Moderate	Weak spillovers across industries
Pre/post AI cohort	Low	Low	Approximates path counterfactual
Cross-country (staggered adoption)	Low	Low	Natural experiment on paths
AI-free training periods	Low	Low	Directly tests skill formation
Expert users, routine tasks	Low	Low	Skill formation not at stake

reflect the value of AI conditional on current skill levels, measured over horizons too short for substantial skill atrophy to accumulate. Our analysis predicts that effect sizes should decline in longer panels as skills degrade, that the decline should be faster in occupations where learning-by-doing is central, and that cross-sectional estimates should systematically exceed within-worker panel estimates from the same setting.

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. Few existing datasets contain these variables; their collection should be a priority for future research.

5 Welfare Analysis and Optimal Policy

The preceding section characterized mismeasurement as a positive phenomenon. We now turn to normative analysis: when is decentralized AI adoption inefficient, and what policies could improve welfare?

The learning spillovers introduced in Section 4 to explain mismeasurement also create inefficiency. When a firm's AI adoption degrades the learning environment for workers at other firms, the firm ignores this external cost. But inefficiency can arise even without spillovers across firms: training data degradation creates a second externality that operates across all users of AI systems. We analyze both sources of inefficiency, then discuss policy

implications.

Proposition 8 (Sources of Inefficiency)**.** *In the decentralized equilibrium:*

(i) Human capital spillovers ($\theta > 0$ or $\psi'(H) > 0$) are necessary and sufficient for overadoption when training data is exogenous.

(ii) Training data degradation ($\partial Q/\partial \bar{\alpha} < 0$) is independently sufficient for overadoption, even absent human capital spillovers.

(iii) When both externalities are present, total welfare loss exceeds the sum of individual effects due to feedback amplification.

5.1 Human Capital Spillovers and Overadoption

Section 4 introduced human capital spillovers to explain why cross-sectional estimates diverge from long-run effects. The same spillovers generate inefficiency: firms undervalue skill accumulation because some benefits accrue to other firms (through labor mobility and knowledge diffusion) or to workers themselves (through option value in future employment).

Definition 4 (Social Welfare and Optimality)**.** Social welfare is $W \equiv \sum_{t=0}^{\infty} \beta^t \int_0^1 Y_i(h_{i,t}, \alpha_{i,t}; A_t)\, di$. The socially optimal adoption α^S maximizes W subject to skill dynamics; the decentralized adoption α^D solves the firm's problem (4). The economy exhibits overadoption if $\alpha^D > \alpha^S$.

Recall the spillover structure from Section 4: output includes a public-good term θH^η, and learning depends on aggregate human capital through $\psi(H)$. The output spillover captures knowledge diffusion; the learning spillover captures mentorship availability.

Proposition 9 (Human Capital Externality)**.** *Under competitive labor markets:*

(i) The decentralized equilibrium exhibits overadoption ($\alpha^D > \alpha^S$) if and only if human capital spillovers are present ($\theta > 0$ or $\psi'(H) > 0$). When spillovers are absent, the decentralized equilibrium is constrained efficient.

(ii) The proportional wedge between private and social marginal returns to skill accumulation is $\theta\eta H^{\eta-1}/V'(h') + \psi'(H)/\psi(H)$.

This result clarifies that overadoption is not automatic – it requires an externality. The mere fact that AI adoption affects human capital does not by itself generate inefficiency. What generates inefficiency is that human capital has social value beyond its private value to the firm. This could arise through various channels: knowledge spillovers as workers move between firms, mentorship and training of junior workers, contributions to industry-wide knowledge stocks, or agglomeration effects in labor markets. The quantitative importance of our results depends on spillover magnitude, which varies across occupations and industries – strong in professions with apprenticeship traditions, weaker in isolated or standardized work.

5.2 Training Data Externality

The feedback loop of Section 4.4 creates a second source of inefficiency, independent of human capital spillovers.

Proposition 10 (Training Data Externality)**.** *With endogenous AI quality:*

(i) Individual adoption degrades future AI quality ($\partial A_{t+1}/\partial \bar{\alpha}_t = \zeta \cdot \partial Q/\partial \bar{\alpha} < 0$), but atomistic firms ignore this aggregate effect, generating overadoption: $\alpha^D > \alpha^S$.

(ii) When both human capital spillovers and training data effects are present, total welfare loss exceeds the sum of individual effects due to feedback amplification.

Unlike the human capital externality, which could in principle be addressed through bilateral contracts between firms and workers, the training data externality operates across all users of AI systems. No individual firm can internalize the effect of its adoption on AI quality experienced by other firms. The externality is maximally diffuse (each firm's contribution is infinitesimal), delayed (today's output enters tomorrow's training data), and practically irreversible (filtering contaminated content is costly). This creates a novel commons problem: human-generated training data is a shared resource that AI adoption depletes.

Software development illustrates the quantitative importance of this channel. Stack Overflow activity declined 25% within six months of ChatGPT's release (del Rio-Chanona et al., 2024), with newer users – who would have contributed fresh perspectives – most likely to exit (Burtch et al., 2024). The resulting training data increasingly reflects AI's existing capabilities rather than the frontier of human problem-solving.

5.3 Policy Implications

When either externality is present, decentralized adoption exceeds the social optimum. Standard Pigouvian logic suggests taxing AI use at the marginal external cost:

$$\tau^* = \underbrace{\beta \frac{\partial W}{\partial H'} \lambda (1-\mu) \varphi(H)}_{\text{human capital externality}} + \underbrace{\beta \frac{\partial W}{\partial A'} \zeta |Q_\alpha|}_{\text{training data externality}}$$

The optimal tax has a notable property: it rises as human capital falls, because skill scarcity raises the marginal value of remaining human capital. This corrective feedback contrasts with standard Pigouvian taxes that are typically constant.

Given practical difficulties in measuring AI use and estimating shadow values, quantity restrictions may be more implementable. A training mandate requiring fraction ρ of work be performed without AI constrains adoption to $\alpha \leq 1 - \rho$. Such mandates have precedents in professional licensing: pilots must log manual flight hours despite autopilot availability; surgical residents must perform procedures unassisted; apprenticeship systems require hands-on work before certification. Extending such requirements to AI-intensive professions – mandatory "unassisted coding" periods, AI-free drafting rotations – would preserve skill formation while allowing AI use in production.

A further consideration is competitive dynamics. Even firms that recognize AI's long-run costs face pressure to adopt: unilateral restraint means losing market share to competitors.

This creates a prisoner's dilemma that compounds the externality problems. Appendix A.5 shows that competitive equilibrium involves overadoption relative to joint profit maximization, and that this competitive effect compounds with human capital spillovers and training data externalities.

The key policy insight is that optimal intervention may reduce *measured* productivity while improving welfare. A regulator relying on standard productivity metrics would perceive skill-preserving policies as costly, even when they raise welfare. This political economy challenge is compounded by the mismeasurement problems we identify: the cost-benefit analyses on which policy decisions rest are themselves biased toward overestimating AI's benefits.

6 Conclusion

This paper identifies two structural sources of mismeasurement in AI productivity studies when adoption affects skill formation. The *spillover bias* arises because non-users in high-adoption environments face degraded learning opportunities; the *state-path divergence* arises because current skill is endogenous to past AI use. Both cause cross-sectional estimates to overstate long-run benefits, and can reverse the sign of measured effects.

The model generates testable predictions. Effect sizes should decline over time as skill atrophy accumulates, with faster decline where learning-by-doing is central. Pre-AI cohorts should command growing wage premiums. Cross-sectional estimates should exceed panel estimates from the same setting. These predictions are conditional on $\mu < 1$; when AI augments learning, they reverse.

Early evidence is consistent with these concerns. Bastani et al. (2025) find GPT-4 access harms learning outcomes, but pedagogically-designed tutors mitigate the harm. METR (2025) find developers are slower with AI yet believe they are faster. Budzyń et al. (2025) document deskilling among endoscopists after three months of AI assistance. Lee et al. (2025) find higher AI reliance associated with reduced critical thinking among knowledge workers.

Autor (2024) offers an optimistic vision: AI could democratize elite expertise. Our analysis clarifies when this applies. Democratization requires $\mu \geq 1$ – AI must augment rather than substitute for learning. When AI provides scaffolding requiring cognitive engagement, it extends expertise. But when AI provides answers directly – the autocomplete model dominating current deployments – it substitutes for the struggle through which expertise develops. The empirical question of which regime applies is first-order for whether AI will rebuild or hollow out the middle class.

Our analysis has limitations that merit discussion. Most centrally, we assume freed time is not productively reinvested. In our model, when AI handles routine tasks, the worker performs fewer tasks and learns less. But workers might reallocate effort to more complex tasks, or use freed time for deliberate practice. If such reallocation is complete and effective, our $\mu < 1$ assumption would not hold. However, emerging evidence suggests reallocation is incomplete: surveys indicate AI-assisted workers spend freed time on leisure rather than skill development, consistent with a revealed preference for immediate utility over human capital investment. The "jagged technological frontier" documented by Dell'Acqua et al. (2023) –

where workers struggle to identify which tasks benefit from AI – compounds this problem: effective reallocation requires meta-knowledge that may itself atrophy with AI dependence.

We also treat pedagogical quality μ as exogenous, though it is really an equilibrium object shaped by AI design, workplace norms, and user incentives. Appendix A shows competitive pressure favors low-μ designs: firms that maximize immediate user productivity outcompete those that preserve learning, even if users would collectively prefer the latter. This selection effect likely strengthens overadoption results. Finally, we abstract from wage adjustments, though skill scarcity should generate growing wage premia for pre-AI cohorts. Such adjustments could partially mitigate the trap by increasing returns to skill preservation, but they also redistribute – rather than eliminate – the welfare costs of skill degradation.

Despite these limitations, the mismeasurement problems we identify follow necessarily from assumptions that are increasingly well-supported empirically. The two sources of bias – spillover degradation and state-path divergence – operate through distinct mechanisms and require different remedies: the former is a classic externality amenable to Pigouvian correction; the latter is a measurement problem requiring counterfactual-aware research designs.

The mismeasurement problems we identify grow worse over time. As adoption spreads, the spillover bias intensifies; as skills atrophy, the state-path divergence widens. Early intervention – before skill degradation becomes entrenched and before the counterfactual workforce has disappeared – is therefore more effective than late intervention. The framework we develop provides tools for identifying when the risk of mismeasurement is greatest, and for designing policies and research strategies that account for it.

References

Abbring, J. H. and J. J. Heckman (2007). Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation. *Handbook of Econometrics* 6B, 5145–5303.

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 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.

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.

Bertrand, Q., J. Bose, A. Duplessis, M. Jiralerspong, and G. Gidel (2024). On the Stability of Iterative Retraining of Generative Models on their Own Data. *International Conference on Learning Representations*.

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.

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

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.

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.

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.

Endsley, M. R. (2017). From Here to Autonomy: Lessons Learned from Human-Automation Research. *Human Factors* 59(1), 5–27.

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

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

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

Goldberg, S. and H. T. Lam (2025). Generative AI in Equilibrium: Evidence from a Creative Goods Marketplace. Working Paper.

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

Ide, E. and E. Talamàs (2025). Artificial Intelligence in the Knowledge Economy. *Journal of Political Economy* 133(12), 4041–4078.

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.

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 (2023). Technology Deskills Jobs, Reduces the Disutility of Work, Particularly Among the Low-Skilled. SSRN Working Paper 4666472.

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.

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.

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.

Xu, Z., S. Jain, and M. Kankanhalli (2024). Hallucination is Inevitable: An Innate Limitation of Large Language Models. arXiv preprint arXiv:2401.11817.

A Extensions

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

A.1 Endogenous Wages and Skill Premia

We embed the baseline model in a labor market equilibrium to generate predictions about wage dynamics. The key question is how AI adoption affects the distribution of wages across workers with different skill levels, and how this distribution evolves over time.

Assumption 3 (Labor Market Structure)**.** Workers are hired in a competitive labor market. Firm i's output depends on worker skill according to $Y_i = f(h_i)(1-\alpha_i)^{1-\gamma} + A \cdot G(\alpha_i)$, where $f(h)$ is strictly increasing and concave. Wages equal marginal products: $w(h_i, \alpha_i) = f'(h_i)(1-\alpha_i)^{1-\gamma}$, which depends on both worker skill and the firm's adoption intensity.

The key insight from this specification is that AI adoption affects wages through two channels. First, the direct productivity effect: AI raises output for given skill, which raises wages. Second, the skill atrophy effect: AI reduces skill accumulation, which lowers future wages. The balance between these effects determines whether AI raises or lowers wages in the long run.

Consider two types of workers: high-skill workers with $h = \bar{h}$ (who developed their skills before AI diffusion) and low-skill workers with $h = h^*$ (who developed skills with AI assistance). The skill premium is $\pi = w(\bar{h})/w(h^*)$.

The following three propositions characterize how AI affects different worker groups and how these effects evolve over time.

Proposition 11 (Ability Reversal)**.** *Let workers differ in learning ability θ_i, with high-ability workers accumulating more human capital absent AI. Then:*

(i) High-ability workers lose the most from skill atrophy ($\partial(h_t^{NA} - h_t^{user})/\partial\theta > 0$) because they would accumulate the most human capital on the no-adoption path.

(ii) AI benefits low-ability workers in the short run (immediate productivity gains) but harms high-ability workers in the long run (foregone skill accumulation), generating a reversal in the distribution of gains over time.

The ability reversal creates a political economy challenge: the workers who benefit most from AI in the short run (low-ability workers experiencing "democratization") are not those who bear the largest long-run costs (high-ability workers who would have accumulated substantial human capital).

Proposition 12 (Cohort Effects)**.** *Let h_t^{pre} denote human capital of pre-AI cohorts and h_t^{post} denote human capital of post-AI cohorts. Suppose $\mu < 1$. Then:*

(i) $h_t^{post} < h_t^{pre}$ for all t sufficiently large: post-AI cohorts accumulate less skill.

(ii) Pre-AI cohorts command a "vintage premium" that grows as the skill gap widens, peaks when the last pre-AI cohort nears retirement, then collapses as fully AI-trained workers dominate the labor market.

The cohort effect generates testable predictions: wage premiums for pre-AI trained workers should emerge and grow over time in occupations where AI substitutes for learning.

Proposition 13 (Wage Inequality Dynamics)**.** *Let $\pi_t \equiv w(\bar{h})/w(h_t^*)$ denote the skill premium at time t. Then:*

(i) $\pi_{t+1} < \pi_t$ *for t small (short-run compression).*

(ii) $\pi_{t+1} > \pi_t$ *for t large (long-run scarcity).*

(iii) $\lim_{t\to\infty} \pi_t = \infty$ *if pre-AI cohorts retire and $h^* < \bar{h}$.*

Corollary 1 (U-Shaped Inequality)**.** *Let $\sigma_t^2 \equiv Var(w_t)$ denote wage inequality. Then:*

(i) There exists $T^* > 0$ such that $\sigma_{t+1}^2 < \sigma_t^2$ for $t < T^*$ (short-run compression) and $\sigma_{t+1}^2 > \sigma_t^2$ for $t > T^*$ (long-run scarcity).

(ii) The turning point T^* is decreasing in both the speed of skill convergence (larger $\delta + \lambda\ell(\alpha^*)|\varphi'(h^*)|$) and the elasticity of substitution between skilled and unskilled labor.

This U-shaped pattern has important implications for policy evaluation. Early studies of AI adoption will observe falling inequality – the "democratization" narrative that has dominated public discussion (see Autor, 2024, for the optimistic case). But this may reverse as skill atrophy accumulates. Policymakers who evaluate AI based only on short-run inequality effects may be misled about long-run consequences.

A.2 Firm Dynamics and Selection

When firms differ in their discount factors, AI adoption generates selection effects that amplify aggregate skill loss. This extension shows how competitive dynamics can worsen the skill trap beyond what any individual firm's optimization would generate.

Assumption 4 (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.

The heterogeneity in patience could arise from differences in ownership structure (public vs. private firms), managerial incentives (short-term vs. long-term compensation), access to capital (firms facing borrowing constraints behave more myopically), or corporate culture (firms that invest in training vs. those that hire experienced workers).

Proposition 14 (Selection Effects)**.** *Under Assumption 4:*

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

(ii) Let $s_{i,t}$ denote firm i's market share. Then $\frac{d}{dt}\mathbb{E}[\beta_i|s_{i,t}] < 0$: the output-weighted average patience declines over time.

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

This extension highlights a collective action dimension of the skill trap. Patient firms that would prefer a low-adoption equilibrium may be unable to sustain their position against impatient competitors. The market selects for short-term productivity, driving out the firms whose long-run orientation would otherwise preserve human capital. The selection effect creates a form of "Gresham's law" for human capital: bad (impatient) firms drive out good (patient) firms. This amplifies the overadoption problem identified in the main text.

A.3 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. This extension analyzes how certification institutions can partially mitigate the skill trap by preserving incentives for skill acquisition.

Assumption 5 (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.

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.

Assumption 6 (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).

Proposition 15 (Certification Equilibrium)**.** *Under Assumptions 5 and 6:*

(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: $\frac{\partial V}{\partial h}\big|_{cert} > \frac{\partial V}{\partial h}\big|_{no\text{-}cert}$.

Certification markets partially mitigate the skill trap by increasing private returns to skill. However, certification addresses only the information problem, not the underlying human capital externality – it is a complement to, not substitute for, corrective policy.

A.4 Adaptive Pedagogical AI Design

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

Definition 5 (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 16 (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.

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. But our analysis suggests Socratic interfaces preserve more human capital, even if measured adoption appears lower. Professional licensing could mandate minimum engagement requirements during training periods.

A.5 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.5.1 Pigouvian Taxation

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

Proposition 17 (Optimal AI Tax)**.** *The optimal per-unit tax on AI adoption equals the marginal external cost evaluated at the current state (H, A):*

$$\tau^* = \underbrace{\beta \frac{\partial W}{\partial H'} \lambda (1-\mu) \varphi(H)}_{\text{human capital externality}} + \underbrace{\beta \frac{\partial W}{\partial A'} \zeta |Q_\alpha|}_{\text{training data externality}}$$

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.5.2 Training Mandates

Given practical difficulties with measuring and taxing AI use, quantity restrictions may be more implementable.

Definition 6 (Training Mandate)**.** A training mandate $\rho \in [0, 1]$ requires that at least fraction ρ of work be performed without AI assistance, constraining adoption: $\alpha \leq 1 - \rho$.

Training mandates have precedents in professional licensing (continuing education requirements) and apprenticeship systems.

Proposition 18 (Welfare Effects of Training Mandates)**.** *Consider a training mandate $\rho \in [0, 1]$ requiring at least fraction ρ of work be performed without AI:*

(i) The mandate is welfare-improving if and only if $\rho \leq 1 - \alpha^S$, where α^S is socially optimal adoption. The optimal mandate sets $\rho^* = 1 - \alpha^S$, exactly implementing the social optimum.

(ii) Under the optimal mandate, measured productivity may fall while welfare rises because human capital is preserved.

Implementation faces several challenges: gaming (workers using AI covertly), heterogeneity (optimal ρ varies across occupations), and international competition (domestic mandates may disadvantage firms against foreign competitors).

A.5.3 Competitive Dynamics

Individual firms face competitive pressure to adopt AI even when they recognize its long-run costs.

Proposition 19 (Competitive Overadoption)**.** *Consider a symmetric duopoly where firms A and B choose adoption intensities α_A and α_B simultaneously. Market share depends on relative productivity: firm i's share is $s_i = Y_i/(Y_A + Y_B)$. Let α^M denote the adoption level that maximizes joint profits and α^N denote the Nash equilibrium adoption level. Then:*

(i) $\alpha^N > \alpha^M$: competitive equilibrium involves overadoption relative to joint profit maximization.

(ii) The overadoption margin $\alpha^N - \alpha^M$ is increasing in the elasticity of market share with respect to productivity.

(iii) This competitive effect compounds with human capital spillovers ($\theta > 0$) and training data externalities (endogenous A). When multiple externalities are present, total overadoption exceeds what any single externality would generate.

Each firm's best response involves adopting more intensively than the other, creating a prisoner's dilemma. Both firms would prefer coordinated restraint, but unilateral restraint means losing market share.

A.6 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:

$$\mathbb{E}[L_i] = \int_0^{h_i} \varphi(h_i)\, dF(s) + \int_{h_i}^{\bar{s}} \kappa\varphi(h_i)[1 - G_H(s)]\, dF(s) \tag{9}$$

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:

$$\mathbb{E}[L_i] = \varphi(h_i)\,[F(h_i) + \kappa\Psi(H)[1 - F(h_i)]] \tag{10}$$

Normalizing so that $\psi(\bar{H}) = 1$ at the no-adoption steady state, we obtain the multiplicative form $L_i = \varphi(h_i) \cdot \psi(H)$ where $\psi(H)$ is increasing in H. 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.7 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:

$$S_t = (1 - \pi_t) \cdot \underbrace{\int q_i \, dF_i}_{\text{human quality}} + \pi_t \cdot \underbrace{A_{t-1}}_{\text{AI quality}} \tag{11}$$

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:

$$A_{t+1} = f\left((1 - \pi_t)\bar{q}_t + \pi_t A_t\right) \tag{12}$$

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:

$$\bar{q}_t = \int h_i(1 - \alpha_i)^\omega \, dF_i \approx H_t \cdot (1 - \bar{\alpha}_t)^\omega \tag{13}$$

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:

$$A_{t+1} = (1 - \zeta)A_t + \zeta \cdot [(1 - \pi)\tilde{Q}(H_t, \bar{\alpha}_t) + \pi \cdot A_t] \tag{14}$$

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:

$$H_{t+1} = (1 - \delta)H_t + \lambda \ell(\bar{\alpha}_t)\varphi(H_t)\psi(H_t) \tag{15}$$

$$A_{t+1} = (1 - \zeta)A_t + \zeta Q(H_t, \bar{\alpha}_t) \tag{16}$$

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

Proposition 20 (Feedback Loop Stability)**.** *The system (H_t, A_t) has a unique stable steady state (H^{**}, A^{**}) satisfying:*

(i) $H^{**} < H^*$*, where* H^* *is the steady state with fixed AI quality: the feedback loop worsens skill atrophy.*

(ii) $A^{**} < A_0$*, where* $A_0 = Q(\bar{H}, 0)$ *is AI quality when humans are fully skilled and no AI is used: equilibrium AI quality is below its potential.*

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

When does the data channel dominate? Define the *data contribution* to the skill trap as $\Delta_D \equiv H^* - H^{**}$, the additional skill loss attributable to training data degradation. This is increasing in:

- ζ: faster AI quality adjustment amplifies the feedback loop
- ω: stronger quality degradation from AI-assisted output
- π: higher AI content contamination in training corpora
- $|\partial\alpha^*/\partial A|$: stronger adoption response to AI quality

The data channel is relatively more important when AI systems are retrained frequently on recent data, when AI-assisted output is easily distinguishable from expert human output (so ω is large), and when AI-generated content proliferates rapidly in training corpora.

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. 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.8 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 9) 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.6 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. We begin by stating the regularity conditions maintained throughout the proofs; these are standard technical conditions ensuring well-behaved steady states and are satisfied for generic parameter values.

Assumption 7 (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.

Remark 6 (Sufficient Primitive Conditions). The regularity conditions are satisfied for standard functional forms. Conditions (i)–(ii) hold when $\varphi(h) = \varphi_0/(1 + h/\xi)$ with ξ sufficiently large relative to δ/λ, ensuring the steady state is interior and stable. Condition (iii) holds when output curvature $|Y_{\alpha\alpha}| = h\gamma(1-\gamma)(1-\alpha)^{-\gamma-1}$ is large relative to the dynamic feedback, which is ensured by γ bounded away from 0 and 1, moderate β, and $|1 - \mu|$ not too large. Condition (iv) holds generically; failure requires knife-edge parameter values where the policy function is non-monotonic. Numerically, these conditions can be verified by checking that the Jacobian of the steady-state system has eigenvalues inside the unit circle.

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 2 (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 3 (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 4 (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 5 (Optimal Adoption is Interior)**.** *Under Assumption 2, $\alpha^*(h) \in (0, 1)$ for all $h \in (0, \bar{h}]$.*

Proof. At $\alpha \to 1$: $\partial Y/\partial\alpha \to -\infty$ (Lemma 3), so $\alpha^* < 1$. At $\alpha = 0$: marginal benefit is $A - h(1-\gamma) > 0$ by Assumption 2, so $\alpha^* > 0$. □

Lemma 6 (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 7(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 7(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 7 (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 1, there exists a unique $h^* > 0$ satisfying the stationarity condition. Under Assumption 7(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 7(ii), $(1-\delta) + \lambda\ell(\alpha^*)\varphi'(h^*) < 1$. The third term is bounded under Assumption 7(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 8 (Jacobian Non-Singularity)**.** *Under Assumption 7, 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 7(ii)
- $D_{h\alpha} = \lambda(1-\mu)\varphi(h) > 0$ since $\mu < 1$ and $\varphi(h) > 0$
- $D_\alpha = Y_{\alpha\alpha} + \beta V''(h')[\lambda(\mu-1)\varphi(h)]^2 + \beta V'(h')\lambda(\mu-1)\varphi'(h)$
- $D_{\alpha h} = -(1-\gamma)(1-\alpha)^{-\gamma} - \beta[V''(h')\lambda(1-\mu)\varphi(h) + V'(h')\lambda(1-\mu)\varphi'(h)]\frac{\partial h'}{\partial h}$

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. The third term $\beta V'(h')\lambda(\mu-1)\varphi'(h)$: with $\mu < 1$ (so $\mu - 1 < 0$) and $\varphi'(h^*) < 0$ (by Assumption 1), this equals $\beta V'(h^*) \cdot (\text{negative}) \cdot (\text{negative}) > 0$. Thus the third term partially offsets the first two.

Under Assumption 7(iii), we have:

$$|Y_{\alpha\alpha}(h^*,\alpha^*)| > \beta|V'(h^*)|\lambda|1-\mu||\varphi'(h^*)| \tag{17}$$

This ensures the static concavity term dominates the positive dynamic term, yielding $D_\alpha < 0$.

Signing $D_{\alpha h}$: We have $D_{\alpha h} = -(1-\gamma)(1-\alpha)^{-\gamma} - \beta[V''(h')\lambda(1-\mu)\varphi(h) + V'(h')\lambda(1-\mu)\varphi'(h)]\frac{\partial h'}{\partial h}$. The first term $-(1-\gamma)(1-\alpha)^{-\gamma} < 0$. For the bracketed expression: $V''(h') \leq 0$ by concavity, $\lambda(1-\mu)\varphi(h) > 0$ when $\mu < 1$, and $V'(h')\lambda(1-\mu)\varphi'(h)$ has sign (positive) $\cdot$ (positive) $\cdot$ (negative) < 0 since $\varphi'(h) < 0$ by Assumption 1. Thus the bracket is negative. Since $\partial h'/\partial h = (1-\delta) + \lambda\ell(\alpha)\varphi'(h) > 0$ under Assumption 7(ii), the second term is positive. 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 7(iv) (monotone policy), $D_{\alpha h}$ has constant sign on $(0, \bar{h}]$. If $D_{\alpha h} \leq 0$, then $-D_{h\alpha}D_{\alpha h} \geq 0$, so $\det(\mathbf{J}) = (\text{negative}) + (\text{non-negative}) < 0$. If $D_{\alpha h} > 0$, then $-D_{h\alpha}D_{\alpha h} < 0$, so $\det(\mathbf{J}) < 0$ provided $|D_h D_\alpha| > |D_{h\alpha}D_{\alpha h}|$. This latter condition is implied by Assumption 7(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

Proposition 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 1 (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 $\ell(\alpha) > 0$. When $\mu \geq 0$, we have $\ell(\alpha) \geq 1-\alpha > 0$ for all $\alpha \in [0,1)$. When $\mu \in (-1,0)$, we have $\ell(\alpha) > 0$ iff $\alpha < \alpha_0 \equiv 1/(1-\mu) \in (1/2,1)$. For $\alpha \geq \alpha_0$ with $\mu < 0$, effective learning becomes zero or negative, and no positive steady state exists – skills decline without bound. In what follows, we restrict attention to (μ,α) pairs satisfying $\ell(\alpha) > 0$; this holds automatically when $\mu \geq 0$ (the empirically relevant case) or when adoption is not too extreme.

Under this restriction, 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 $\mu \geq 1$. □

Proposition 2 (Steady-State Characterization).

The characterization follows directly from the properties of the steady-state human capital function $h^*(\alpha)$ established in Lemma 1. □

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 3. 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 1, 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 AI is sufficiently productive that this holds, $Y^* \geq \bar{h} = Y^{NA}$, and the trap cannot exist when $\mu \geq 1$. □

Proposition 3 (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 8, $\det(\mathbf{J}) \neq 0$. 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: Both effects reduce the "cost" of adoption. 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 $\partial h^*/\partial\mu > 0$: 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: $\partial h^*/\partial\mu = [\lambda\alpha^*\varphi(h^*) - \lambda(1-\mu)\varphi(h^*)(\partial\alpha^*/\partial\mu)]/[\delta - \lambda\ell(\alpha^*)\varphi'(h^*)]$. The denominator is positive by Assumption 7. The numerator's first term $\lambda\alpha^*\varphi(h^*) > 0$; the second term involves $\partial\alpha^*/\partial\mu > 0$ multiplied by $-(1-\mu)\varphi(h^*) < 0$ when $\mu < 1$. Thus both terms in the numerator are positive, yielding $\partial h^*/\partial\mu > 0$: higher pedagogical quality raises steady-state human capital. □

Proposition 4 (Existence of Skill Trap).

We verify each condition of Definition 1 and establish uniqueness of $\bar{\beta}$.

Step 1: Condition (T1) holds. By Assumption 2, $A > \bar{h}(1-\gamma)$. By Lemma 5, $\alpha^*(h) > 0$ 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 4), 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 2. Since the firm chooses $\alpha_0^* > 0$ (Lemma 5) and payoff is strictly concave in α (Lemma 3), 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 2–7 by Lemma 5 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 6). 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 4, 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 2–7).

Remark 7. The monotonicity result $W'(\alpha^*)<0$ does not require any additional assumption beyond Assumptions 2–7. The impatience condition in Remark 8 ensures well-behaved comparative statics but is not needed for the sign of $W'(\alpha^*)$.

Remark 8 (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 Proposition 3(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 $\beta\to 1$: We show $\alpha^*(\beta)\to 0$ and hence $Y^*(\beta)\to\bar{h}$. From the steady-state FOC:

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

As $\beta \to 1$, the RHS grows (patient firms weight future skills heavily). For the FOC to hold, either $\alpha^* \to 0$ (reducing the LHS) or $h^* \to \bar{h}$ (increasing the skill cost term). Under $\mu < 1$, the stationarity condition $\delta h^* = \lambda \ell(\alpha^*)\varphi(h^*)$ with $\ell(\alpha) = 1 - (1-\mu)\alpha$ implies that $h^* \to \bar{h}$ requires $\alpha^* \to 0$ (since $\delta \bar{h} = \lambda \varphi(\bar{h})$ defines $\bar{h}$). Thus both occur jointly: $\alpha^*(\beta) \to 0$ and $h^*(\beta) \to \bar{h}$ as $\beta \to 1$. Consequently $Y^*(\beta) \to \bar{h}$, so $\Psi(1^-) = 0$.

As $\beta \to 0$: myopic firms maximize current output. The static FOC $Ag(\alpha) = h(1-\gamma)(1-\alpha)^{-\gamma}$ determines adoption. As $\beta \to 0$, firms ignore future skill costs, so $\alpha^*(\beta) \to \alpha^{myopic}$ where α^{myopic} maximizes $Y(h, \alpha)$ for fixed h. Since $Y_\alpha \to -\infty$ as $\alpha \to 1$ (Lemma 3), $\alpha^{myopic} < 1$. However, as $\beta \to 0$, the steady-state skill $h^*(\alpha)$ falls toward zero because the firm does not internalize skill atrophy. Specifically, from the stationarity condition, $h^* \to 0$ as $\alpha^* \to \bar{\alpha}$ where $\ell(\bar{\alpha})\varphi(0)$ balances depreciation at a very low skill level. With $h^* \approx 0$ and $\alpha^* < 1$, we have $Y^* \approx A \cdot G(\alpha^*)$. By condition (ii), $A \cdot G(1) < \bar{h}$, and since $G(\alpha^*) < G(1)$, we have $Y^* < A \cdot G(1) < \bar{h}$, so $\Psi(0^+) < 0$.

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 2, $Ag(0) > \bar{h}(1-\gamma)$. Since $g(\alpha) \geq \underline{g} > 0$ for all α and $[1-(1-\alpha)^{1-\gamma}] \leq (1-\gamma)\alpha$ for α small (by convexity), Assumption 2 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; but when $\mu \geq 1$, the equilibrium α^* is bounded away from 1 because higher adoption does not degrade skills. 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 9 (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 5 (Divergence Between Counterfactuals).

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 human capital spillovers, output includes a spillover term θH_t^η that depends on aggregate human capital, and learning includes a multiplicative spillover term $\psi(H_t)$. In the skill trap, aggregate H_t falls over time as adopters' skills decline. This affects non-users through two channels.

First, even if non-users' own human capital accumulation rate remains positive, with the learning spillover $\psi(H_t)$ declining as H_t falls, their actual skill accumulation is impaired: $h_{t+1}^{NU} = (1-\delta)h_t^{NU} + \lambda\varphi(h_t^{NU}) \cdot \psi(H_t)$. By Lemma 9, $\psi(H_t) < \psi(\bar{H}) = 1$ when $H_t < \bar{H}$, so non-users accumulate skills more slowly than they would in the no-adoption counterfactual. Formally, both paths start from $h_0^{NU} = h_0^{NA} = h_0$, but since $\psi(H_t) < 1$ for all $t > 0$, the non-user's human capital transition is uniformly dominated: $h_{t+1}^{NU} < h_{t+1}^{NA}$ for all $t \geq 0$. By induction, $h_t^{NU} < h_t^{NA}$ for all $t > 0$.

Second, their *effective output* is lower than it would be in a world without AI adoption, because the output spillover term $\theta H_t^\eta < \theta \bar{H}^\eta$.

The cross-sectional counterfactual compares user output to non-user output in the same (AI-adopting) economy:

$$\Delta_t^{CS} = [A \cdot G(\alpha_t) + h_t^U(1-\alpha_t)^{1-\gamma} + \theta H_t^\eta] - [h_t^{NU} + \theta H_t^\eta] = A \cdot G(\alpha_t) + h_t^U(1-\alpha_t)^{1-\gamma} - h_t^{NU}$$

The output spillover terms cancel in the cross-sectional comparison.

The long-run counterfactual compares to a world without AI:

$$\Delta_t^{LR} = [A \cdot G(\alpha_t) + h_t^U(1-\alpha_t)^{1-\gamma} + \theta H_t^\eta] - [\bar{h} + \theta \bar{H}^\eta]$$

The difference is:

$$\Delta_t^{CS} - \Delta_t^{LR} = [\bar{h} + \theta\bar{H}^\eta] - [h_t^{NU} + \theta H_t^\eta] = \underbrace{(\bar{h} - h_t^{NU})}_{\text{skill degradation}} + \underbrace{\theta(\bar{H}^\eta - H_t^\eta)}_{\text{spillover loss}} \geq 0$$

The first term captures skill degradation of non-users through the learning spillover channel (positive when $\psi(H_t) < 1$). The second term captures the output spillover loss from aggregate skill decline. Both terms grow as adoption spreads and H_t falls, so Δ_t^{CS} increasingly overstates benefits relative to Δ_t^{LR}. □

Sign Reversal Condition.

At steady state, define $Y^* \equiv A \cdot G(\alpha^*) + h^*(1-\alpha^*)^{1-\gamma}$. Then:

$$\Delta^{CS} = Y^* - h^{NU*}$$
$$\Delta^{LR} = Y^* - \bar{h} - \theta(\bar{H}^\eta - H^{*\eta})$$

For sign reversal ($\Delta^{CS} > 0 > \Delta^{LR}$), we need both $Y^* > h^{NU*}$ and $\Delta^{LR} < 0$.

For $\Delta^{LR} < 0$: rearranging $Y^* < \bar{h} + \theta(\bar{H}^\eta - H^{*\eta})$ gives $\theta > (Y^* - \bar{h})/(\bar{H}^\eta - H^{*\eta})$. In the trap, $Y^* < \bar{h}$ (Proposition 4), so the numerator $(Y^* - \bar{h})$ is negative. Since $H^* < \bar{H}$ and $\eta > 0$, the denominator $(\bar{H}^\eta - H^{*\eta})$ is positive. Thus the RHS is negative, and *any* $\theta > 0$ satisfies this inequality. Hence in the trap with any positive spillovers, $\Delta^{LR} < 0$ automatically.

For $\Delta^{CS} > 0$: this requires $Y^* > h^{NU*}$, i.e., AI users must outperform degraded non-users. This is the binding condition for sign reversal.

Combining: in the skill trap, sign reversal occurs if and only if (i) spillovers are present ($\theta > 0$ or $\psi'(H) > 0$), ensuring $\Delta^{LR} < 0$, and (ii) $Y^* > h^{NU*}$, ensuring $\Delta^{CS} > 0$. □

Proposition 6 (State-Path Divergence).

Part (i): By Lemma 1, $h_t^{user} \to h^* < \bar{h}$ as $t \to \infty$ when $\mu < 1$ and $\alpha^* > 0$. Thus $Y(h_t^{user}, 0) = h_t^{user} \to h^* < \bar{h} = Y(h_t^{NA}, 0)$. The state-conditional counterfactual for the AI user deteriorates relative to the never-adopted path.

Part (ii): We establish three formal claims about Δ_t^{SC}.

Claim 1: Convergence to lower steady state. Under $\mu < 1$ and the stationary optimal policy $\alpha_t = \alpha^*(h_t)$, starting from $h_0 \in (0, \bar{h}]$, the sequence $\{h_t^{user}\}$ converges to $h^* < \bar{h}$. By Lemma 7, convergence follows from local stability of h^* and the global properties of the transition map on $(0, \bar{h}]$. Specifically, define $T(h) = (1-\delta)h + \lambda\ell(\alpha^*(h))\varphi(h)$. At $h = \bar{h}$: since $\alpha^*(\bar{h}) > 0$ (Lemma 5) and $\ell(\alpha) < 1$ when $\mu < 1$ and $\alpha > 0$, we have $T(\bar{h}) = (1-\delta)\bar{h} + \lambda\ell(\alpha^*)\varphi(\bar{h}) < (1-\delta)\bar{h} + \lambda\varphi(\bar{h}) = \bar{h}$, where the last equality uses the definition of $\bar{h}$. At h^*: $T(h^*) = h^*$ by definition. By continuity of T and the intermediate value theorem, for any $h \in (h^*, \bar{h}]$, we have $T(h) < h$. Combined with local stability (Assumption 7(ii)), the sequence converges to h^*. By Lemma 1, $h^* < \bar{h}$ when $\mu < 1$ and $\alpha^* > 0$.

Claim 2: Bounded absolute gain, potentially large relative gain. The state-conditional gain is $\Delta_t^{SC} = Y(h_t^{user}, \alpha_t) - Y(h_t^{user}, 0) = A \cdot G(\alpha_t) + h_t^{user}(1-\alpha_t)^{1-\gamma} - h_t^{user}$. Rewriting:

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

As $h_t^{user} \to h^*$, the absolute gain $\Delta_t^{SC} \to A \cdot G(\alpha^*) - h^*[1 - (1-\alpha^*)^{1-\gamma}]$, which is bounded. The *relative* gain $\Delta_t^{SC}/Y(h_t^{user}, 0) = \Delta_t^{SC}/h_t^{user}$ satisfies:

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

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. Thus AI's contribution relative to unassisted productivity grows as skills atrophy, and can be arbitrarily large for parameterizations yielding sufficiently low h^*.

Claim 3: Steady-state output falls below no-adoption benchmark. From Proposition 4, when $\beta < \bar{\beta}$, steady-state output satisfies $Y^* < \bar{h} = Y^{NA}$. This means AI users in steady state produce less than they would have produced on the no-adoption path. Yet for any t sufficiently large that h_t^{user} is near h^*, we have $\Delta_t^{SC} > 0$ (AI raises current output given current skills). This is the core of state-path divergence: the state-conditional comparison $\Delta_t^{SC} > 0$ while steady-state output comparison $Y^* < \bar{h}$.

Part (iii): At any time t after sufficient atrophy, $Y(h_t^{user}, 0) = h_t^{user} < Y(h_t^{user}, \alpha_t)$, so removing AI access would reduce current output. The worker is "dependent" on AI in this state-conditional sense, even though $Y^* < \bar{h}$ implies they would produce more in steady state on the never-adopted path.

Part (iv): Consider the path counterfactual $\Delta^{PATH}(\tilde{\beta}) = \sum_{\tau=0}^{\infty} \tilde{\beta}^\tau [Y_\tau^{user} - Y_\tau^{NA}]$. For the firm's own discount factor β, revealed preference implies $\Delta^{PATH}(\beta) \geq 0$. However, when

$\tilde{\beta} > \bar{\beta}$, more weight is placed on long-run outcomes where $Y_t^{user} < Y_t^{NA}$ (for $t > T^*$). Since $Y^* < \bar{h}$, the tail of the sum is negative, and for $\tilde{\beta}$ sufficiently large, $\Delta^{PATH}(\tilde{\beta}) < 0$. This formalizes the welfare divergence: an evaluator more patient than the firm judges the adoption path welfare-inferior. □

Proposition 7 (The Skill-Data Feedback Loop).

Part (i): From the AI quality law of motion, $A_{t+1} = (1-\zeta)A_t + \zeta Q(H_t, \bar{\alpha}_t)$ where $\partial Q/\partial H > 0$. Thus $\partial A_{t+1}/\partial H_t = \zeta \cdot \partial Q/\partial H > 0$.

Part (ii): Similarly, $\partial Q/\partial \bar{\alpha} < 0$ by assumption, so $\partial A_{t+1}/\partial \bar{\alpha}_t = \zeta \cdot \partial Q/\partial \bar{\alpha} < 0$.

Part (iii): Consider optimal adoption $\alpha^*(h, A)$ solving the FOC:

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

By the implicit function theorem:

$$\frac{\partial \alpha^*}{\partial A} = \frac{g(\alpha^*)}{-\partial^2(\cdot)/\partial \alpha^2} > 0$$

since the second-order condition ensures the denominator is positive. Higher AI quality increases adoption.

For the effect of h, the FOC gives:

$$\frac{\partial \alpha^*}{\partial h} = \frac{(1-\gamma)(1-\alpha^*)^{-\gamma} + \beta V''(h')\lambda(1-\mu)\varphi(h) \cdot \partial h'/\partial h + \beta V'(h')\lambda(1-\mu)\varphi'(h)}{-\partial^2(\cdot)/\partial \alpha^2}$$

The denominator is positive by the second-order condition. The numerator's first term is positive. The second term involves $V'' \leq 0$ (concavity) and is non-positive. The third term involves $\varphi'(h) < 0$ by Assumption 1.

Since $\varphi'(h) < 0$ everywhere, the third term is negative (since $(1-\mu) > 0$ when $\mu < 1$), reinforcing the negative second term. The sign of the numerator is thus ambiguous in general, depending on whether the positive first term dominates.

However, the *dependence effect* operates through a different channel. As skills atrophy and h falls toward h^*, workers don't increase adoption *because* $\partial \alpha^*/\partial h < 0$; rather, they maintain high adoption because their outside option (productivity without AI) has deteriorated. The state-conditional comparison $Y(h, \alpha) - Y(h, 0)$ grows as h falls, making AI appear increasingly essential even if optimal α doesn't change much. This is the "indispensability" result of Proposition 6, not a claim about the comparative static $\partial \alpha^*/\partial h$.

When both A and h fall: $\partial \alpha^*/\partial A > 0$ suggests lower A reduces adoption. However, the dependence mechanism operates through the *level* of the outside option, not through $\partial \alpha^*/\partial h$. As skills atrophy, the worker's AI-independent productivity $Y(h, 0) = h$ falls. Even if optimal adoption α^* doesn't increase, the *value* of AI to the worker – measured by $Y(h, \alpha) - Y(h, 0)$ – grows because the denominator (unassisted productivity) shrinks. Workers with low h thus remain dependent on AI despite its degradation because their outside option has deteriorated. This is the state-path divergence formalized in Proposition 6.

Part (iv): Consider steady states of two systems:

System (a): Fixed AI quality $A = A_0$. Steady state H^* solves $\delta H = \lambda \ell(\alpha^*(H; A_0))\varphi(H)$.

System (b): Endogenous AI quality. Steady state (H^{**}, A^{**}) solves:

$$\delta H^{**} = \lambda \ell(\alpha^{**})\varphi(H^{**})$$
$$A^{**} = Q(H^{**}, \alpha^{**})$$

Suppose $A_0 = Q(\bar{H}, 0)$, the AI quality when humans are fully skilled and no AI is used. In system (b), if $\alpha^{**} > 0$ and $H^{**} < \bar{H}$ (the trap conditions), then:

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

since $\partial Q/\partial H > 0$ and $\partial Q/\partial \alpha < 0$.

With lower A^{**}, one might expect lower adoption. But from part (iii), workers with low H^{**} maintain high adoption despite lower AI quality. The steady-state condition $\delta H^{**} = \lambda \ell(\alpha^{**})\varphi(H^{**})$ with high α^{**} (low ℓ) implies lower H^{**} than in system (a) where A_0 is higher.

Formally, at the steady state of system (b), the FOC with lower A^{**} shifts the marginal benefit curve down. If workers were at H^* (from system (a)), they would reduce adoption. But with endogenous A, the system settles at a lower H^{**} where the marginal *learning* cost of adoption – the right-hand side of the FOC, $\beta V'(h')\lambda(1-\mu)\varphi(h)$ – is also lower (because both $V'(h')$ and $\varphi(h)$ fall with h in the relevant region). This lower marginal learning cost sustains high α despite lower A. Hence $H^{**} < H^*$. □

Proposition 8 (Sources of Inefficiency).

Part (i): Consider the social planner's problem with human capital spillovers but exogenous training data ($A_t = A$ for all t). The planner maximizes $\sum_t \beta^t [Y(H_t, \alpha_t; A) + \theta H_t^{\eta}]$ subject to $H_{t+1} = (1-\delta)H_t + \lambda \ell(\alpha_t)\varphi(H_t)\psi(H_t)$. The FOC includes the social value of human capital, which exceeds the private value when $\theta > 0$ or $\psi'(H) > 0$. This wedge implies $\alpha^S < \alpha^D$ (Proposition 9). Conversely, when $\theta = 0$ and $\psi(H) \equiv 1$, the FOCs coincide and the decentralized equilibrium is efficient.

Part (ii): Consider the case $\theta = 0$, $\psi(H) \equiv 1$, but $A_{t+1} = (1-\zeta)A_t + \zeta Q(H_t, \bar{\alpha}_t)$ with $\partial Q/\partial \bar{\alpha} < 0$. The social planner internalizes the effect of adoption on future AI quality: $\partial W/\partial \alpha$ includes the term $\beta(\partial W/\partial A') \cdot \zeta \cdot (\partial Q/\partial \bar{\alpha}) < 0$. This additional cost is absent from the private FOC. Hence $\alpha^S < \alpha^D$ even without human capital spillovers.

Part (iii): The divergence between counterfactuals (Proposition 5) arises because non-users' skill formation depends on aggregate human capital through $\psi(H)$. If $\psi(H) \equiv 1$, non-users' learning is unaffected by aggregate adoption, eliminating the divergence mechanism.

Part (iv): When both externalities operate, the feedback loop (Proposition 7) creates amplification: lower H reduces Q, which reduces A, which induces higher α, which further reduces H. The welfare loss from each externality is magnified by the presence of the other. □

Proposition 9 (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 10 (Training Data Externality).

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 $\bar{\alpha} = \int \alpha_i \, di$ affects AI quality. The social FOC includes an additional term:

$$\beta \frac{\partial W}{\partial A'} \cdot \zeta \frac{\partial Q}{\partial \bar{\alpha}} < 0$$

since $\frac{\partial W}{\partial A'} > 0$ (higher AI quality raises welfare) and $\frac{\partial Q}{\partial \bar{\alpha}} < 0$ (more adoption degrades training data). This additional cost implies $\alpha^S < \alpha^D$.

When both human capital spillovers and training data effects are present, the total marginal external cost at a given state (H, A, α) is:

$$\tau^{total}(H, A) = \underbrace{\beta \frac{\partial W}{\partial H'} \lambda(1-\mu)\varphi(H)[\psi(H) - 1]}_{\tau^{HC}(H,A)} + \underbrace{\beta \frac{\partial W}{\partial A'} \zeta |Q_{\bar{\alpha}}|}_{\tau^{data}(H,A)}$$

Both terms are positive. While this expression is additive at any given state, the *equilibrium* wedge exhibits complementarity. Specifically, let $\tau^{HC}_{partial}$ denote the optimal tax in a model with only HC spillovers (holding A fixed) and $\tau^{data}_{partial}$ the optimal tax with only training data effects (holding H fixed). Then:

$$\tau^{total}(H^{**}, A^{**}) > \tau^{HC}_{partial}(\bar{H}, A_0) + \tau^{data}_{partial}(\bar{H}, A_0)$$

where (H^{**}, A^{**}) is the joint equilibrium and $(\bar{H}, A_0)$ is the no-externality benchmark. The strict inequality arises because the HC channel lowers H below $\bar{H}$, which increases both $\partial W/\partial H'$ (scarcity raises marginal value) and $|Q_{\bar{\alpha}}|$ (lower skills worsen data degradation). The externalities reinforce each other in general equilibrium. □

Propositions 11, 12, and 13 (Ability Reversal, Cohort Effects, and Wage Inequality).

Under Assumption 3, wages equal marginal products: $w(h) = f'(h)(1-\alpha)^{1-\gamma}$ where f is output per unit human capital. This specification assumes workers are perfect substitutes

within skill levels. For the scarcity results, we additionally assume an aggregate production function with imperfect substitution across worker vintages: $Y = F(N^{pre} \cdot h^{pre}, N^{AI} \cdot h^{AI})$ where F exhibits diminishing marginal products. Under CES aggregation with elasticity $\sigma < \infty$, relative wages depend on both marginal products and relative supplies, generating the scarcity channel.

Proposition 11: High-ability workers have higher $\bar{h}$ and thus higher foregone learning $\bar{h} - h^*$ when AI substitutes for skill formation. The short-run gain from AI is proportional to current output, which is higher for high-ability workers, but the long-run loss is proportional to foregone human capital accumulation, which is also higher. The net effect depends on discounting: with sufficient patience, high-ability workers are harmed more.

Proposition 12: At $t = 0$, pre-AI workers have $h^{pre} = \bar{h}$ and post-AI workers begin accumulating with $\mu < 1$. As AI-trained workers' skills converge to $h^* < \bar{h}$, the vintage gap $h^{pre} - h_t^{post}$ grows. With $N_t^{pre} = N_0^{pre} e^{-\nu t}$ (retirement at rate ν), scarcity drives up the premium.

Proposition 13: The premium $\pi_t = w^{pre}/w_t^{post}$. *Part (i):* Initially, AI compresses wages by raising w_t^{post} for low-skill workers. *Part (ii):* As $h_t^{post} \to h^* < h^{pre}$, the wage gap widens. *Part (iii):* As $N_t^{pre} \to 0$, remaining pre-AI workers become arbitrarily scarce and $\pi_t \to \infty$. □

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 14 (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): The bias is $\text{Cov}(s_{i,t}, h_{i,t})$. Since $s_{i,t}$ is high when α_i is high (short-run productivity), while $h_{i,t}$ is low when α_i is high (skill atrophy), this covariance is negative. Cross-sectional estimates weighted by market share understate skill degradation. □

Proposition 15 (Certification Equilibrium).

Part (i): Consider a candidate separating equilibrium with threshold h^*: workers with $h \geq h^*$ certify, others do not. Employers observe h directly for certified workers and pay $w^C(h) = f'(h)$. For uncertified workers, employers pay expected productivity:

$$w^U = \mathbb{E}[f'(h)|h < h^*] = \frac{\int_0^{h^*} f'(s)dG(s)}{G(h^*)}$$

Certification is individually rational for worker with skill h if $f'(h) - c \geq w^U$, i.e., $h \geq h^*$ where h^* solves $f'(h^*) - c = w^U(h^*)$. This fixed point exists and is unique under standard regularity conditions.

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

Part (iii): Private return to skill with certification is $\partial w^C/\partial h = f'(h) > 0$, since f is increasing in h. Without certification, wages pool across unobservable skill levels: $\partial w/\partial h = 0$. The higher private return under certification, $f'(h) > 0 = \partial w/\partial h$, 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 16 (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 Proposition 3, $\partial \alpha^*/\partial \mu > 0$ 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): Define "frustration" as $\phi = 1/\mu$ (inverse pedagogical quality). Users prefer low ϕ (easy AI), but welfare-maximizing $\phi^S < \phi^D$: socially optimal AI is more frustrating than what users would choose. □

Proposition 17 (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:

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

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

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

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

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

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. □

Proposition 18 (Welfare Effects of Training Mandates).

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

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

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

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

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

$$W = \sum_{t} \beta^t [Y_t + \theta H_t^{\eta}]$$

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

Proposition 19 (Competitive Overadoption).

Part (i): 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:

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

A higher elasticity of market share with respect to productivity increases this term, 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. Define the spillover distortion as $\Delta^{spill} \equiv \alpha^D - \alpha^S > 0$ and the competitive distortion as $\Delta^{comp} \equiv \alpha^N - \alpha^D > 0$. The total distortion is:

$$\alpha^N - \alpha^S = \Delta^{spill} + \Delta^{comp} + \underbrace{\mathcal{I}}_{\text{interaction}}$$

where $\mathcal{I} > 0$ is the interaction term arising because the two distortions are not additively separable. The interaction term is positive because the 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 20 (Feedback Loop Stability).
Part (i): With endogenous A, the steady-state H solves:

$$\delta H = \lambda \ell(\alpha^*(H, A(H)))\varphi(H)\psi(H)$$

where $A(H) = Q(H, \alpha^*(H, A(H)))$ in steady state. Since $\partial\alpha^*/\partial A > 0$ (Proposition 3) and $\partial A/\partial H > 0$ (Proposition 7), higher H raises A, which raises α^*, which lowers H'. This additional feedback (absent when A is fixed) shifts the steady-state condition, yielding $H^{**} < H^*$.

Part (ii): 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 $\mathbf{J}$ depends on model parameters. Stability requires both eigenvalues of $\mathbf{J}$ to have modulus less than 1.

Part (iii): With ζ small (slow AI adjustment), $J_{21} \approx \zeta \cdot \partial Q/\partial H$ and $J_{22} \approx 1 - \zeta$. The eigenvalues approach those of the H-only system (which is stable by Lemma 6) plus one eigenvalue near 1. Slow AI adjustment ensures the A dynamics do not destabilize the system. □