Skill Atrophy and AI Productivity Measurement

Tommaso Bondi* Gentry Johnson†

Abstract

How should we measure AI's productivity effects? Standard estimates condition on proxies for current skill, but when AI substitutes for the cognitive effort that builds skill, skill itself becomes endogenous to past AI use. We formalize *pedagogical quality*, the fraction of learning-by-doing that survives AI delegation, and show it drives two biases: state-path divergence inflates within-worker panel estimates relative to the path counterfactual, and spillover bias contaminates control groups when learning is social. The biases generate a "scissors" pattern: panel and long-horizon RCT estimates diverge over time, so that measured productivity gains may increasingly favor adoption even as AI's long-run effect declines.

*Cornell Tech and SC Johnson School of Business, Cornell University. Email: tbondi@cornell.edu.

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

We thank Ajay Agrawal, Guy Aridor, Dirk Bergemann, Ron Berman, Luis Cabral, Avi Goldfarb, Brett Hemenway Falk, Judy Hanwen Shen, Brett Hollenbeck, David Holtz, Vrinda Kadiyali, Jacob LaRiviere, Jura Liaukonytė, Xueming Luo, Emaad Manzoor, John McHale, Ivan Png, Omid Rafieian, Joao Rodrigues, Eduard Talamàs, Alex Tamkin, Gerry Tsoukalas, Joel Waldfogel, Michael Waldman, and Nathan Yang for helpful comments and suggestions.

AI often raises short-run productivity: experimental estimates range from 14% to 55% at the task level (Brynjolfsson et al., 2025; Dell'Acqua et al., 2026; Noy and Zhang, 2023; Peng et al., 2023). But when AI substitutes for the practice through which workers build skill, it erodes human capital – and standard productivity estimates, which hold skill fixed, overstate AI's long-run effect.

Automation-induced skill decay is not new. Why does AI raise this concern more acutely than earlier automation? The concern is sharpest when delegated tasks are themselves learning-rich. Earlier waves of automation – assembly lines, ATMs, basic data processing – primarily displaced tasks with limited learning content. Generative AI is often used on activities such as drafting, diagnosis, and coding, where execution is also practice. Automating filing does not degrade legal reasoning; automating brief writing can, because constructing arguments is itself the practice that builds the capacity to evaluate them.

Evidence consistent with skill atrophy is accumulating. Bastani et al. (2025) find that unrestricted GPT-4 access reduces subsequent math performance. Shen and Tamkin (2026) find a skill reduction among software developers learning a new programming library. Budzyń et al. (2025) document endoscopist deskilling after three months of AI-assisted colonoscopy. Some workers appear to recognize the trade-off: a student in a related chess experiment reported, "I want to think for myself, not use the button" (Poulidis et al., 2025).

Yet existing productivity experiments identify the contemporaneous effect holding skill fixed; by design, they cannot identify the effect relative to the counterfactual skill path. AI may therefore look increasingly transformative precisely because we compare it to workers whose skills have atrophied, even if it looks modest relative to the counterfactual worker.

This measurement concern is distinct from a growing theoretical literature that identifies private–social wedges in AI adoption: demand externalities that drive firms into collective over-automation (Hemenway Falk and Tsoukalas, 2026), and public-good failures in the production of collective knowledge (Acemoglu et al., 2026b). Whether AI adoption is privately or socially desirable depends first on measuring its effect correctly. The same dynamics that drive those wedges also corrupt the estimands used to detect them.

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

The biases grow with adoption duration and coexist whenever $\mu < 1$ and spillovers are present. Because both inflate measured AI gains, a naive reading of the evidence would support continued or expanded adoption even as the underlying skill base deteriorates – creating a mismeasurement-policy feedback loop. Identifying the bias requires tracking not just current performance but the counterfactual skill path – what the worker would have accumulated had she never delegated.

Consider two questions about a worker who has used AI for years. *"How would she perform today if we switched AI off?"* (the within-worker panel, conditioning on proxies for current skill). *"How would she have performed if she had* never *used AI?"* (the long-horizon RCT, comparing against the counterfactual path). The questions sound nearly identical, but when AI erodes learning, their answers diverge.

The panel takes current skill as given – skill shaped by years of delegation – and sees AI becoming more valuable, because the worker's unaided capacity has eroded. The RCT compares against the worker she would have been, with skill built through unaided practice, and sees AI becoming less valuable, because the treated worker has lost ground relative to the control. Together they produce a *scissors pattern*: panel estimates of AI's productivity gain rise while long-horizon RCT estimates fall. This divergence is not a statistical artifact; it reflects a genuine difference in the question each design answers.

The scissors pattern is the model's cleanest testable prediction. It would not arise under benchmark habit formation, learning curves, and selection on early adopters: habit generates withdrawal penalties that dissipate, learning curves raise both estimands, and selection does not generate divergence in the formal estimands (Supplemental Appendix B). Among the benchmark mechanisms considered, only skill atrophy generates one estimand rising while the other falls.

The same mechanism has distributional implications. If workers differ in baseline skill, higher-skilled workers optimally delegate less in our setup, so short-run compression in measured productivity can mask longer-run divergence in skill trajectories.

The dynamic framework we employ serves as a conservative benchmark. The problem is fundamentally intertemporal: delegation raises productivity today but reduces the skill stock available tomorrow. A static model would capture the state-conditional productivity gain from AI at a point in time, but not the wedge between estimands that opens as skill erodes – nor the feedback loop that makes that wedge grow, as lower skill raises the return to delegation, which further erodes skill, generating a self-reinforcing spiral between AI dependence and human capital loss. Characterizing this spiral, the steady-state it converges

to, and the scissors prediction it implies requires a dynamic model.

A behavioral model could generate over-adoption through inattention or myopia. We show neither is necessary: in our model, the decision maker is a rational, forward-looking agent – an individual worker or a firm – who fully internalizes the future skill cost of delegation. This modeling choice minimizes delegation: a less patient agent, or one who underestimates the learning cost $1 - \mu$, would adopt more, widening the biases. The model establishes the measurement problem as a feature of the technology, not a consequence of behavioral or organizational error.

A numerical illustration makes magnitudes concrete. For instance, at $\mu = 0.5$, the measurement bias exceeds 11% of the initial productivity gain within a decade, and the scissors pattern emerges: over 20 years, the panel estimand rises roughly 9% while the RCT estimand falls roughly 9%, though both evaluate the same technology's effect.

I Related Literature

This paper contributes to several adjacent literatures; for recent surveys, see Acemoglu (2025) and Agrawal et al. (2026). The task-based framework of Acemoglu and Restrepo (2018, 2020) models automation as machines performing tasks previously done by humans, taking human capital as fixed; Jones and Tonetti (2026) find that weak links – tasks still performed by slowly-improving labor – constrain the aggregate effects of even dramatic automation. Agrawal et al. (2026) emphasize complementarities between AI and human judgment. Ide and Talamàs (2025) study how AI reorganizes hierarchical knowledge firms; in their framework human capital is a static endowment, whereas in ours it evolves endogenously through the work AI displaces. AI may complement the *use* of judgment while substituting for its *development*.

A growing empirical literature documents the short-run productivity effects of AI: Brynjolfsson et al. (2025) for call centers, where the gains include durable worker learning consistent with $\mu \approx 1$ in a structured-feedback setting; Noy and Zhang (2023) for writing; Peng et al. (2023) for coding; Dell'Acqua et al. (2026) for consulting. The "jagged frontier" identified by Dell'Acqua et al., in which AI substantially helps on some tasks but hurts on others, reflects heterogeneous static AI capability across tasks (heterogeneous $g(j)$ in our framework). Gans (2026) formalises this jagged frontier as irregular coverage over a task space and derives an inspection-paradox wedge between benchmark reliability and user-experienced reliability; the measurement concern there is cross-sectional – which tasks a user actually faces at a

point in time – whereas ours is dynamic, concerning how the worker's skill evolves along the adoption path. A distinct and less studied source of heterogeneity is that the learning content of delegated tasks may also vary: μ may differ across task types, with the sharpest skill atrophy arising when AI encroaches on the most learning-rich tasks. Otis et al. (2023) find heterogeneous effects among Kenyan entrepreneurs: AI mentorship helped high performers but *hurt* low performers. Gaessler and Piezunka (2023) find chess computers accelerated skill development ($\mu > 1$), plausibly because chess provides immediate, objective feedback – though a related experiment in which AI provided on-demand move suggestions finds the opposite effect (Poulidis et al., 2025), illustrating that μ depends on how the tool is deployed, not just on the domain. The majority of recent work documents deskilling, including among endoscopists (Budzyń et al., 2025), robot-assisted workers (Beane, 2019), students (Bastani et al., 2025), and knowledge workers (Dell'Acqua, 2022; Lee et al., 2025; Shen and Tamkin, 2026).

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

Concurrent work by Acemoglu et al. (2026b) studies how agentic AI degrades collective knowledge in a Bayesian social learning framework; their mechanism operates through information aggregation, ours through individual human capital dynamics. Our distinctive contribution is the measurement focus: rather than asking whether AI helps or hurts in the long run, we characterize how standard causal estimands diverge from the path counterfactual along the skill transition.

A parallel strand of concurrent work identifies private–social wedges in AI adoption. Hemenway Falk and Tsoukalas (2026) study demand externalities: foresighted firms collectively over-automate because each ignores the drag that its layoffs impose on the aggregate demand base. The economic primitives are related to ours but the questions differ: they characterize conditions under which private optimization yields collective welfare losses, whereas we characterize how standard causal estimands misrepresent the productivity gain

itself, even when the decision-maker internalizes the skill trajectory in full. The biases we identify therefore persist in settings where no such wedge arises – in our benchmark the agent is long-lived and fully forward-looking, yet panel and RCT estimands still diverge.

At the same time, the two papers are natural complements. Eroding human capital and the compressed wages or layoffs that eventually track it are two sides of the same coin: the first reflects what AI adoption does to the skill stock, the second reflects how that stock is priced in the labor market. We isolate an upstream mechanism that standard productivity measurement fails to register; Hemenway Falk and Tsoukalas (2026) trace its downstream labor-market footprint, amplified by demand externalities each firm alone ignores. Quantifying the skill-erosion channel requires a measurement baseline that current designs do not deliver.

Recent work examines how AI threatens training and skill transmission: Garicano and Rayo (2025) show apprenticeships become unviable when AI automates entry-level work, and Beane (2019) finds robotic surgery reduced trainee practice tenfold. These papers study whether skill-building opportunities survive organizationally; ours studies what happens to learning *within* ongoing production when workers delegate cognitive tasks to AI.

II Model

II.A Environment and Primitives

Time is discrete, indexed by $t \in \{0, 1, 2, \ldots\}$. Each period, the agent completes a unit continuum of tasks indexed by $j \in [0, 1]$. Each task can be performed either by the agent or by AI. When the agent performs task j, output from that task is $y(j, t) = h_t \cdot e_t(j)^\gamma$, where $h_t \geq 0$ is human capital, $e_t(j) \geq 0$ is effort allocated to task j, and $\gamma \in (0, 1)$ governs the returns to effort. When AI performs task j, output is $y(j, t) = A \cdot g(j)$, where $A > 0$ is AI productivity and $g : [0, 1] \to \mathbb{R}_+$ is continuously differentiable on $[0, 1)$, satisfies $g(0) = 1$, and has $g'(j) < 0$.

The condition $g'(j) < 0$ encodes comparative advantage: AI is more capable at routine, well-defined tasks (low j) than at complex, judgment-intensive tasks (high j). AI quality A is exogenous; we abstract from endogenous improvement to isolate the human capital channel. This is conservative: if AI improves over time, the incentive to delegate rises further, amplifying the skill dynamics we study.

The agent faces an effort constraint: total effort across all manually performed tasks is normalized to unity. When an agent adopts AI at intensity $\alpha \in [0, 1]$, it delegates tasks

in $[0, \alpha]$ to AI while the agent performs tasks in $(\alpha, 1]$. Optimal uniform effort allocation yields total output $h(1-\alpha)^{1-\gamma}$: delegating tasks to AI concentrates effort on remaining tasks, partially offsetting the lost output and generating real short-run productivity gains even when skill atrophy operates in the background.

Substituting, period output takes the tractable form

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

where $G(\alpha) \equiv \int_0^\alpha g(j)\, dj$ is cumulative AI output, with $G'(\alpha) = g(\alpha)$ and $G''(\alpha) = g'(\alpha) < 0$. The exponent $1-\gamma < 1$ reflects effort concentration: when the agent performs fewer tasks, effort is more concentrated. The function is linear in h, strictly concave in α, and satisfies $\partial Y/\partial \alpha \to -\infty$ as $\alpha \to 1^-$ for any $h > 0$, ensuring full delegation is never optimal.

The output structure reveals the difference-versus-sum tension central to the paper. The *difference* $Y(h, \alpha) - Y(h, 0)$ is decreasing in h: AI is less valuable when the agent is already skilled. When skill atrophy lowers h, the difference rises while total output $Y(h, \alpha)$ falls. Experiments measure the difference; aggregate productivity tracks the latter.

II.B Human Capital Dynamics

Human capital evolves according to

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

where $\delta \in (0, 1)$ is depreciation, $\lambda > 0$ governs learning intensity, and $L(\alpha, h; \mu)$ is the learning function. The learning function is

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

where $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is continuously differentiable, strictly positive, and strictly decreasing, and $\mu \geq 0$ is *pedagogical quality*. Equation (3) is a deliberately reduced-form representation of how delegation changes the learning content of work. Because $g'(j) < 0$, the first tasks delegated are the most routine. The scalar μ should therefore be read as an adoption-weighted average of the learning content of delegated tasks, not as a task-by-task primitive.[1] Since φ is positive and decreasing – reflecting diminishing returns to learning as skill accumulates

[1] With task-specific pedagogical content $\mu(j)$, the effective pedagogical quality at adoption rate α would be $\int_0^\alpha \mu(j)\, dj/\alpha$. The reduced-form scalar μ is enough for the measurement questions studied here.

– the stationarity condition $\delta h = \lambda\varphi(h)$ has a unique positive solution $\bar{h}$, the no-adoption steady state. Letting $\beta \in (0,1)$ denote the agent's discount factor, we maintain:

Assumption 1 (Learning Capacity)**.** $\varphi(h) = c/(1+h)$ for some $c > 0$, $\lambda\varphi(0) < 1 - \delta$, and $\beta\lambda\varphi(0) < 1 - \beta(1-\delta)$.

The functional form keeps the endogenous argument transparent. The first slope condition ensures that the skill transition is increasing; combined with Assumption 2, the second ensures that the Bellman continuation term cannot overturn the static submodularity of output. The benchmark calibration satisfies the first condition by a wide margin and the second with roughly 2.7× headroom.

The effective learning rate is $\ell(\alpha) \equiv 1 - (1-\mu)\alpha$. The critical property is $\partial L/\partial\alpha = (\mu - 1)\varphi(h)$: when $0 < \mu < 1$, delegation reduces learning, producing the skill atrophy of the title. When $\mu = 1$, delegation has no effect on learning and all measurement biases vanish, a useful benchmark. When $\mu > 1$, AI *augments* learning and all signs reverse.

The parameter μ has clear empirical content. Bastani et al. (2025), for example, show that GPT-4 access harms math learning ($\mu < 1$), but pedagogically-designed tutors mitigate this, confirming that μ is a design parameter, not a fixed property of AI. Existing experimental evidence for $\mu < 1$ comes predominantly from novice or skill-formation settings; for experienced workers performing well-practiced tasks, delegation may cost little in terms of learning ($\mu \approx 1$). The learning specification depends on the *measure* of delegated tasks, not on effort intensity or throughput.[2]

II.C The Dynamic Problem

The decision maker is a long-lived agent – an individual worker or a firm that retains its workforce – who fully internalizes the skill trajectory, with discount factor β. The measurement biases depend on the equilibrium skill path, not on whether that path is first-best.[3] The agent solves

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

[2]If workers reallocate freed effort to more demanding tasks, the effective learning loss is smaller (Lee et al., 2025). Skill-complementarity in the production function would strengthen both biases.

[3]When the agent is a firm without retention contracts, adoption rises: a separated worker takes her preserved skill with her, so the firm does not fully internalize the returns to skill it has chosen not to erode; private over-delegation would reinforce the biases.

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

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

Writing $T(h, \alpha) \equiv (1-\delta)h + \lambda \ell(\alpha)\varphi(h)$ for the skill transition, $T(h, \alpha) \leq T_0(h) \equiv (1-\delta)h + \lambda\varphi(h)$ for all α when $\mu \leq 1$, and $T_0(\bar{h}) = \bar{h}$, the interval $[0, \bar{h}]$ is invariant. On this set, standard results ensure V exists, is unique, continuous, increasing, and convex (Supplemental Lemma 3); at states with a unique interior optimum, the envelope condition yields differentiability.

The benchmark agent of (4) is long-lived and fully forward-looking. This choice is deliberately conservative: it minimizes adoption, and therefore the magnitude of the measurement biases we derive. A less patient decision maker, or one who does not internalize part of the skill trajectory – for instance, a manager whose horizon falls short of the worker's tenure, or a firm that cannot contract on post-separation human capital – would adopt more aggressively. Current skill would drift further from the no-adoption counterfactual, widening the wedges characterized below. The measurement problem is therefore a feature of the delegation technology, not a consequence of myopia or misaligned incentives.

II.D Equilibrium Characterization

II.D.1 The Role of Pedagogical Quality

Adoption raises contemporaneous output through $G(\alpha)$, but it changes tomorrow's state by altering the rate at which today's work translates into skill. Pedagogical quality μ determines which force dominates at the margin. When $\mu < 1$, adoption reduces future human capital. Complete delegation is never optimal because effort concentration on remaining tasks becomes increasingly valuable as $\alpha \to 1$.

Whenever the optimum is interior, the first-order condition is

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

where $h' = (1-\delta)h + \lambda\ell(\alpha)\varphi(h)$. The right-hand side is the dynamic learning cost of delegation.

Proposition 1 (Role of Pedagogical Quality)**.** *At any interior optimum:*

(i) When $\mu < 1$, optimal adoption is strictly below the myopic level $\alpha^m(h)$ solving $Y_\alpha(h, \alpha) =$

0.

(ii) When $\mu = 1$, optimal adoption equals the myopic level.

When $\mu < 1$, the forward-looking agent restrains delegation: the shadow cost of forgone learning, $\beta V'(h')\lambda(1-\mu)\varphi(h)$, is strictly positive and pushes adoption below the static optimum. Each delegated task carries an implicit price – the discounted future output lost because today's delegation reduces tomorrow's skill. The wedge between optimal and myopic adoption grows with the agent's patience (β) and with the learning loss from delegation ($1-\mu$). When $\mu = 1$, the right-hand side of (6) vanishes, delegation carries no skill cost, and the dynamic problem reduces to a sequence of static problems.

II.D.2 Endogenous Skill and Delegation Paths

For intuition, if adoption were held fixed at α, steady-state skill would solve

$$\delta h_*(\alpha) = \lambda \ell(\alpha)\varphi(h_*(\alpha)). \tag{7}$$

When $\mu < 1$, higher delegation lowers $h_*(\alpha)$: holding adoption fixed, AI reduces the long-run skill stock by lowering the learning content of work. This fixed-adoption benchmark is useful for isolating the measurement logic, but the main text focuses on the endogenous path, where skill and delegation co-evolve.

To characterize the endogenous path, we impose one additional primitive restriction.

Assumption 2 (Bounded Learning Loss)**.** $\mu \geq \gamma$.

This says pedagogical quality is at least as large as the returns-to-effort parameter. It is a convenient sufficient condition for the economic inequality actually used in the proof, $\beta\lambda(1-\mu)\varphi(0) < (1-\gamma)[1-\beta(1-\delta)]$; the benchmark calibration takes $\mu = \gamma = 0.5$, sitting at the boundary of the stated primitive but well inside the working region under Assumption 1. The condition rules out the case in which the dynamic learning term overturns the model's basic static submodularity.

We also focus on parameter values for which AI is attractive enough to be used at the no-adoption steady state:

$$A \cdot g(0) > \bar{h}\left[(1-\gamma) + \frac{\beta\delta(1-\mu)}{1-\beta(1-\delta)}\right].$$

Since $\lambda\varphi(\bar{h}) = \delta\bar{h}$, this sufficient condition compares the first delegated task's static gain to the agent's current marginal contribution plus the maximal discounted value of forgone learning.

Proposition 2 (Endogenous Characterization)**.** *Suppose* $0 < \mu < 1$*, Assumptions 1–2 hold, and the interior-adoption condition above is satisfied. Then there exists a decreasing optimal stationary policy selection* $\alpha^*(h)$*. Starting from* $h_0 = \bar{h}$*, the induced equilibrium path has* $h_{t+1} < \bar{h}$ *and* $h_{t+1} \leq h_t$ *for all* $t \geq 0$*, and* $\alpha_{t+1} \geq \alpha_t > 0$ *for all* t*, with at least one strict inequality whenever* $h_t > h^*$*. The path converges to an interior steady state* (h^*, α^*) *with* $0 < h^* < \bar{h}$ *and* $\alpha^* \in (0, 1)$*.*

Assumption 2 yields decreasing differences in the Bellman objective, so higher-skill agents optimally delegate less. Once adoption starts at $\bar{h}$, that monotone policy makes the skill path move downward and the adoption path move upward: lower skill raises the optimal delegation rate, which further reduces the learning content of work. Letting h_t^U denote the skill path under adoption, the measurement wedge $\bar{h} - h_t^U > 0$ therefore holds along the endogenous path. The same wedge arises along any path on which adoption lowers skill, including fixed-α paths; the endogenous result matters because it shows that such paths arise optimally and that the feedback is self-reinforcing.

At the steady state established by Proposition 2, a marginal increase in constant adoption reduces output. Defining $W(\alpha) \equiv A \cdot G(\alpha) + h_*(\alpha)(1-\alpha)^{1-\gamma}$ and using the envelope theorem:

$$W'(\alpha^*) = -\frac{(1-\beta)\,\lambda(1-\mu)\varphi(h^*)(1-\alpha^*)^{1-\gamma}}{\Gamma\big[(1-\beta)+\beta\Gamma\big]} < 0 \tag{8}$$

where $\Gamma \equiv \delta - \lambda\ell(\alpha^*)\varphi'(h^*) > 0$. Nevertheless, the present value of short-run output gains exceeds the discounted skill losses, so adoption is optimal despite $W' < 0$.

III Mismeasurement of AI Productivity

We now show that standard empirical designs systematically diverge from the path counterfactual when $\mu < 1$. All wedges vanish at $\mu = 1$. State-path divergence is the more fundamental bias, requiring only individual-level dynamics; spillover bias compounds it in cross-sectional settings, but has a quantitatively smaller effect in our calibration.

III.A State-Path Divergence

If AI changes the law of motion for skill, conditioning on current skill compares outcomes across different skill histories. Two workers at the same current skill level may have arrived there through very different paths – one through unaided practice, the other through years of heavy delegation. Their future trajectories differ, but the state-conditional comparison treats them as equivalent. State-conditional effects can therefore be positive even when adoption lowers the long-run level of skill. Let h_t^{NA} denote the *no-adoption counterfactual*: the path satisfying $h_{t+1}^{NA} = (1-\delta)h_t^{NA} + \lambda\varphi(h_t^{NA})$ from $h_0^{NA} = \bar{h}$, the pre-AI steady state. Since $\bar{h}$ is the unique steady state of this recursion, $h_t^{NA} = \bar{h}$ for all t.

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

$$\Delta_t^{SC} = Y(h_t^U, \alpha_t) - Y(h_t^U, 0). \tag{9}$$

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

$$\Delta_t^{PATH} = Y(h_t^U, \alpha_t) - Y(h_t^{NA}, 0). \tag{10}$$

We define both objects at a single period t, matching typical empirical horizons.

Many empirical implementations target an object closer to Δ_t^{SC} than to Δ_t^{PATH}: the effect of turning AI "on" at a given skill level, whether through explicit controls for experience and tenure or implicitly by comparing the same worker over time. When the treatment changes the state variable, the relevant benchmark is the path counterfactual Δ_t^{PATH}.

Proposition 3 (State-Path Divergence)**.** *Suppose* $0 < \mu < 1$*, Assumptions 1–2 hold, and the interior-adoption condition is satisfied. Then:*

(i) The estimand wedge is

$$\Delta_t^{SC} - \Delta_t^{PATH} = h_t^{NA} - h_t^U = \bar{h} - h_t^U > 0 \quad \textit{for all } t \geq 1.$$

(ii) The wedge is strictly increasing in t *along the transition path, converging to* $\bar{h} - h^*$*.*

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

Part (i) states the estimand wedge directly. The overstatement equals the skill gap between the no-adoption counterfactual and the realized skill under adoption. Part (ii) says that

the gap widens along the endogenous transition because adoption lowers skill cumulatively over time. Part (iii) says AI always appears beneficial in state-conditional comparisons, even when the path counterfactual is small or negative. The clean decomposition in Part (i) exploits the linearity of Y in h; the qualitative conclusion – that conditioning on current skill overstates the path counterfactual whenever adoption lowers skill – holds for any output function increasing in h.

The estimand wedge reflects a treatment-induced confound. Conditioning on current skill when skill is itself shaped by past treatment biases the estimate upward. This confound is familiar from the "bad controls" literature (Angrist and Pischke, 2009), but the usual advice – omit the post-treatment variable – does not resolve it here, because skill enters the untreated potential outcome directly.

The bias has a *dynamic structure*: it compounds along the equilibrium transition, generates the scissors divergence between estimands (Propositions 5–6), and spills over to contaminate control groups (Proposition 4). Since $W'(\alpha^*) < 0$ (equation (8)), the long-run output effect of marginal permanent delegation is negative even though the measured short-run effect is positive.

State-path divergence operates entirely within a single agent's history and is sufficient for the scissors prediction (Proposition 6). When learning is additionally socially produced, a second bias arises.

III.B Spillover Bias

Cross-sectional comparisons introduce a second wedge when skill is socially produced. As adoption becomes widespread, non-adopters experience degraded learning environments through reduced mentorship, fewer high-skill peers, and thinner task exposure, so the "control group" no longer proxies for the no-adoption counterfactual. The channel is familiar from the education literature on peer effects (Epple and Romano, 2011; Sacerdote, 2001): here the analogue operates through AI adoption, as the stock of available mentors and on-the-job training opportunities shrinks for everyone (see for instance Burtch et al. (2024) on the decline of Stack Overflow). Acemoglu et al. (2026b) micro-found a related channel: when individual learning generates public signals that sustain collective knowledge, agentic AI can trigger "knowledge collapse" by eroding learning incentives economy-wide.

Definition 2 (Cross-Sectional vs. Long-Run Counterfactuals)**.** The *cross-sectional counter-*

factual compares AI users to contemporaneous non-users:

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

The *long-run counterfactual* compares to the path where AI was never adopted:

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

All paths start from $h_0 = \bar{h}$.

In potential-outcomes notation, $h_t^U = h_t(1)$, $h_t^{NA} = h_t(0)$, and $h_t^{NU} = h_t(0 \mid \varpi_t > 0)$: the non-user's skill path in the presence of aggregate adoption. $\Delta_t^{LR} = \Delta_t^{PATH}$ from Definition 1; the relabeling reflects the shift from the individual-level comparison in Section III.A to cross-sectional comparisons, where Δ_t^{LR} serves as the benchmark against which spillover bias is measured.

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

Consider now a population of such agents whose learning depends on peers' skill levels. Let $\varpi_t \in (0, 1)$ denote the share of agents using AI in period t (reflecting heterogeneity in task requirements or experimental assignment), let $H_t \equiv \varpi_t h_t^U + (1 - \varpi_t) h_t^{NU}$ denote aggregate skill, and let $\bar{H} \equiv \bar{h}$ denote its no-adoption level. Individual learning in the spillover economy is $L_i = [(1 - \alpha_i) + \mu\alpha_i]\varphi(h_i)\psi(H)$ with $\psi(H) = (H/\bar{H})^\eta$ and $\eta \geq 0$. A non-user accumulates skill according to

$$h_{t+1}^{NU} = (1 - \delta)h_t^{NU} + \lambda\varphi(h_t^{NU}) \cdot \psi(H_t), \tag{11}$$

while the no-adoption counterfactual satisfies $h_{t+1}^{NA} = (1 - \delta)h_t^{NA} + \lambda\varphi(h_t^{NA})$ (since $\psi(\bar{H}) = 1$).

Define the *spillover skill gap* $s_t \equiv h_t^{NA} - h_t^{NU}$. Then $s_0 = s_1 = 0$, and for $t \geq 1$,

$$s_{t+1} = (1 - \delta)s_t + \lambda\Big[\varphi(h_t^{NA}) - \varphi(h_t^{NU})\psi(H_t)\Big]. \tag{12}$$

The forcing term is positive at $t = 1$: since $h_1^{NA} = h_1^{NU} = \bar{h}$, it reduces to $\lambda\varphi(\bar{h})[1-\psi(H_1)] > 0$, so $s_2 > 0$. For $t \geq 2$, the gap remains strictly positive because the non-user path stays below $\bar{h}$ once spillovers begin (formalized in the proof of Proposition 4).

Proposition 4 (Spillover Bias)**.** *Suppose $\mu < 1$, the spillover specification above holds with $\eta > 0$, and in every period a positive but not full share of agents uses AI. Then:*

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

(ii) The spillover skill gap satisfies $s_t > 0$ *for all* $t \geq 2$*; if the spillover dynamics converge and a positive user share remains in the limit, the limit is strictly positive.*

Cross-sectional estimates overstate the long-run effect of AI as soon as spillovers have time to operate.[4] When the spillover system has a stable limit under persistent partial adoption, the resulting wedge is permanent. When $\eta = 0$, cross-sectional estimates correctly measure long-run effects, but state-path divergence (Proposition 3) remains.

The two biases differ in remedies and empirical status, and they connect to distinct identification literatures. State-path divergence is a treatment-induced confound: the treatment changes the state on which future outcomes depend, a dynamic version of the "bad controls" problem. Spillover bias is a form of interference: adoption by treated units degrades the learning environment of control units, violating the stable unit treatment value assumption.

State-path divergence calls for counterfactual-aware designs and is pinned down by μ, for which multiple experimental estimates exist. Spillover bias calls for designs robust to interference and depends on η, which lacks direct estimation in AI settings; the evidence is suggestive (Beane, 2019; Burtch et al., 2024) but not definitive. The quantitative importance of the spillover channel depends on the density of mentorship networks and the extent to which AI adoption by some workers degrades learning opportunities for others – parameters that current data do not yet pin down. Even at $\eta = 0$, state-path divergence alone delivers the scissors prediction.

Remark 1 (Bias Decomposition). The total bias in cross-sectional estimates admits a useful decomposition. Adding and subtracting both $Y(h_t^U, 0)$ and $Y(h_t^{NA}, 0)$:

$$\Delta_t^{CS} = \underbrace{Y(h_t^U, \alpha_t) - Y(h_t^U, 0)}_{\Delta_t^{SC} \text{ (state-conditional gain)}} + \underbrace{Y(h_t^U, 0) - Y(h_t^{NA}, 0)}_{\text{state-gap bias } (=h_t^U - h_t^{NA} < 0)} + \underbrace{Y(h_t^{NA}, 0) - Y(h_t^{NU}, 0)}_{\text{spillover bias } (=h_t^{NA} - h_t^{NU} = s_t > 0)} \quad (13)$$

The state-conditional gain Δ_t^{SC} is what panel studies typically estimate. The state-gap component $h_t^U - h_t^{NA} < 0$ *reduces* the cross-sectional estimate relative to the state-conditional one; the spillover component $s_t > 0$ *inflates* it by depressing the non-user comparison group.

[4] Proposition 4 does not require Assumptions 1–2; it holds for any path along which some agents adopt at positive intensity.

The two components push the cross-sectional estimate in opposite directions relative to Δ_t^{SC}: the state-gap term makes Δ_t^{CS} smaller than Δ_t^{SC}, while the spillover term makes it larger. The net cross-sectional bias $\Delta_t^{CS} - \Delta_t^{LR} = s_t > 0$ is unambiguously positive. At the benchmark calibration the state-gap dominates the spillover by an order of magnitude across the η range we consider (Supplemental Table S3), so Δ_t^{CS} actually falls *below* Δ_t^{SC}; a widening panel–cross-section gap is therefore a further diagnostic of atrophy, separable from the panel–RCT scissors.

III.C Implications for Empirical Research

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

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

where $h_t(d)$ is the skill path under treatment $d \in \{0, 1\}$. Over short horizons, $h_t(1) \approx h_t(0)$ and the estimand is close to the direct productivity effect. Over long horizons, skill paths diverge when $\mu < 1$. A within-worker panel that conditions on proxies for current skill aims to recover

$$\tau^{panel}(t) = \mathbb{E}[Y_t(1, h_t) - Y_t(0, h_t) \mid h_t] = \Delta_t^{SC} \tag{15}$$

which *overstates* the path counterfactual by the estimand wedge $\bar{h} - h_t^U > 0$ (Proposition 3). This interpretation requires that controls such as tenure, experience, or task history proxy well for h_t, so residual AI use is approximately orthogonal to skill. Worker fixed effects and event studies remove baseline heterogeneity but do not in general deliver "on vs. off at fixed h_t" without repeated withdrawal periods. Absent such conditioning, a pre-post event study estimates $Y(h_t^U, \alpha_t) - \bar{h}$, which is $\tau^{RCT}(t)$ itself; the scissors comparison therefore requires a panel design that targets Δ_t^{SC} explicitly, via within-worker variation at controlled skill levels or short-horizon withdrawal protocols. Likewise, most AI experiments randomize *access*, not intensity: the ITT averages over realized usage responses, so the endogenous panel–RCT divergence is best read as a usage-margin result.

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

The choice of research design determines exposure to these biases. Within-firm RCTs are especially exposed to spillover bias when coworkers share mentorship networks; comparing pre-AI to post-AI cohorts approximates the path counterfactual along the dimensions

studied here, though it introduces standard cohort-comparison confounds. The emerging experimental literature provides building blocks: Bastani et al. (2025) measure skill directly and Budzyń et al. (2025) observe unassisted performance after AI exposure. Novice samples maximize exposure while expert samples minimize it: the population where AI appears most transformative is precisely where the bias is largest.

Disciplining μ empirically is the primary quantitative task our framework identifies. A delayed post-test that measures unassisted performance as a function of prior AI exposure duration traces the cumulative learning loss directly; variation in exposure length then pins down μ relative to the learning and depreciation parameters. Few existing studies satisfy this requirement jointly, though the ingredients are familiar from classical learning-curve estimation. Settings with plausibly exogenous variation in the timing or intensity of AI access – staggered platform rollouts, licensing constraints, or geographic and temporal heterogeneity in tool availability – permit μ to be bounded even absent a pre-registered RCT. Heterogeneity in μ across tasks and settings is itself informative: the uneven effects in Otis et al. (2023), the positive learning effect in Gaessler and Piezunka (2023), and the negative learning effect in the closely related Poulidis et al. (2025) are consistent with a design-dependent pedagogical quality, reinforcing that the measurement problem is not settled by a single number but by a distribution indexed on how AI is deployed.

The sharpest empirical implication concerns the time dynamics of these estimands. When $\mu < 1$, RCTs and panels move in opposite directions – the scissors pattern (Proposition 6).

For a long-horizon RCT with no spillovers, the control group remains on the no-adoption path, so the relevant estimand is

$$\tau^{RCT}(t) \equiv Y(h_t^U, \alpha_t) - Y(\bar{h}, 0).$$

A within-worker panel that conditions on proxies for current skill targets Δ_t^{SC}.

Proposition 5 (Panel–RCT Divergence)**.** *Suppose $0 < \mu < 1$, $\eta = 0$, Assumptions 1–2 hold, and the interior-adoption condition is satisfied. Then, for all $t \geq 1$,*

$$\Delta_t^{SC} - \tau^{RCT}(t) = \bar{h} - h_t^U > 0,$$

and this gap is strictly increasing until the steady state is reached.

The proposition proves the divergence directly. Along the endogenous transition, the panel estimand is evaluated against the agent's current, already-atrophied skill, while the long-horizon RCT compares against the no-adoption path.

Proposition 6 (Endogenous Scissors Pattern)**.** *Suppose* $0 < \mu < 1$*,* $\eta = 0$*, Assumptions 1–2 hold, and the interior-adoption condition is satisfied. Let* $T(h) \equiv (1-\delta)h + \lambda\ell(\alpha^*(h))\varphi(h)$ *denote the induced endogenous transition. Then:*

(i) Δ_t^{SC} *is strictly increasing in* t*.*

If in addition $\beta T'(h) < 1$ *for all* $h \in [h^*, \bar{h}]$*, then:*

(ii) $\tau^{RCT}(t)$ *is strictly decreasing in* t*.*

(iii) If in addition $Y(h^*, \alpha^*(h^*)) < \bar{h}$*, there exists finite* $\hat{t} > 0$ *such that* $\tau^{RCT}(\hat{t}) < 0$ *(the condition* $\tau^{RCT}(0) > 0$ *follows from Proposition 3(iii)).*

The condition $\beta T'(h) < 1$ says the discounted skill feedback does not amplify more than one-for-one. It holds automatically under fixed adoption, where $\beta T'(h) \leq \beta(1-\delta) < 1$, and is verified at the benchmark calibration ($\max \beta T' \approx 0.90$). Violation of $\beta T'(h) < 1$ would require the endogenous policy response to fully offset the direct productivity effect of skill, a configuration absent from the benchmark calibration.

A separate condition, $Y(h^*, \alpha^*(h^*)) < \bar{h}$, enters only in Part (iii). At the high-A benchmark it fails – steady-state output under optimal adoption remains above the no-adoption level, because the direct $AG(\alpha^*)$ gain outweighs the reduction in effort-adjusted unaided output, $\bar{h} - h^*(1-\alpha^*)^{1-\gamma}$. So $\tau^{RCT}(t)$ declines monotonically but remains positive over the horizons we study. The zero-crossing conclusion applies at weaker AI; the low-A benchmark in Supplemental Appendix A ($A = 0.5$) produces a crossing at approximately year 58. The qualitative scissors – Parts (i)–(ii) – obtain at both benchmarks.

Remark 2 (Dependency Spiral). Proposition 6(i) establishes that Δ_t^{SC} rises in levels. Since h_t is simultaneously falling (Proposition 2), the relative gain Δ_t^{SC}/h_t^U is also strictly increasing along the equilibrium path.

AI's contribution relative to the agent's unaided capacity grows – not because AI improves, but because that capacity shrinks. The feedback is self-reinforcing: as unaided performance falls, optimal delegation rises (Proposition 2), which further erodes the skill base against which AI is evaluated. The scissors' qualitative direction does not require this feedback: under any fixed $\alpha \in (0,1)$ along a declining skill path, Δ^{SC} rises and τ^{RCT} falls. The endogenous policy response amplifies magnitudes without driving the sign of the prediction.

Supplemental Appendix B formalizes three benchmark alternatives: habit formation (RCT estimand constant), complementary learning (both estimands rise), and selection on early

adopters (both constant). None of these alternatives generates one estimand rising while the other falls; only skill atrophy does. A growing gap between panel and RCT estimands is therefore a *diagnostic* of skill atrophy under the benchmark alternatives considered here. When mechanisms coexist the diagnostic weakens: skill atrophy combined with complementary learning, for instance, can yield a rising panel estimand and an approximately flat RCT, which is still distinguishable from pure habit or pure learning curves but requires sufficient statistical power on the RCT time trend.

Testing the prediction requires settings where both estimands can be tracked over multi-year horizons – a design no existing study provides. An ideal test would pair a within-worker panel controlling for skill proxies with a long-horizon RCT maintaining a no-access control, both sustained for five or more years. Existing data hint at the mechanism: in a longitudinal developer productivity study, experienced developers were 19% *slower* with AI yet believed they were faster (Becker et al., 2025), consistent with state-conditional reasoning in which workers evaluate their productivity relative to current – not counterfactual – skill. A follow-up had to be redesigned after a substantial share of participants reported unwillingness to complete tasks without AI assistance (METR, 2026).

Numerical estimates. Supplemental Appendix A reports additional numerical estimates. We vary AI productivity A because it governs equilibrium adoption intensity, and therefore the speed at which skill atrophy operates. At the high-A benchmark ($A = 1.85$, $\mu = 0.5$, $\beta = 0.95$), the agent initially delegates roughly 72% of tasks; the equilibrium path converges to a steady state near $(h^*, \alpha^*) \approx (0.35, 0.77)$, where skill has fallen to roughly 70% of its pre-adoption level. The panel–RCT gap widens steadily along this transition, from 7% of the initial gain at year 5 to 11% at year 10 and 18% at year 20. For comparison, under exogenous adoption at $\alpha = 0.5$, steady-state skill falls only to 80% – the endogenous feedback between skill erosion and rising delegation adds roughly a third more bias at steady state.

Supplemental Table S2 reports steady-state outcomes across μ. Lower pedagogical quality *raises* steady-state adoption: across $\mu \in \{1.0, 0.9, 0.7, 0.5, 0.3\}$, optimal α^* moves through $\{0.73, 0.74, 0.76, 0.77, 0.79\}$. This is the dependency spiral visible in cross-section: lower μ depresses h^*, and the now-less-productive worker optimally delegates more at that lower skill, outweighing the direct dynamic cost that Proposition 1 identifies. Figure 1(a) illustrates the resulting scissors at the high-A benchmark.

At a lower $A = 0.50$, optimal adoption is much more modest ($\alpha^*(\bar{h}) \approx 0.24$, converging to $\alpha^* \approx 0.29$), and the dynamics are slower but qualitatively identical. The long-horizon RCT

eventually turns negative (Figure 1(b)), illustrating that even modest AI can erode the skill base over sufficiently long horizons.

The adoption response itself is modest – rising from 72% to 77% of tasks at steady state – because the steep marginal cost of delegation at high adoption rates (strict concavity of output in α) and the rising shadow cost of forgone learning (Proposition 1) restrain the policy response even as submodularity pushes it upward. But the cumulative effect on skill is substantial, because slightly higher delegation compounds over many periods.

These numerical patterns illustrate how modest values of μ accumulate into large measurement errors over empirically realistic horizons. The ratio Δ_t^{SC}/h_t^U, which approximates what a panel or survey analysis would report as AI's contribution to the worker's output, rises monotonically along the transition. A reader observing a rising panel estimate over time would reasonably infer that AI is becoming more useful; in the model, the same pattern obtains because the unaided baseline against which AI's help is measured is shrinking. The mechanism requires only that some part of the delegated work carries learning content and that the measurement horizon is long enough for skill to respond – a horizon that the current experimental literature has not yet reached.

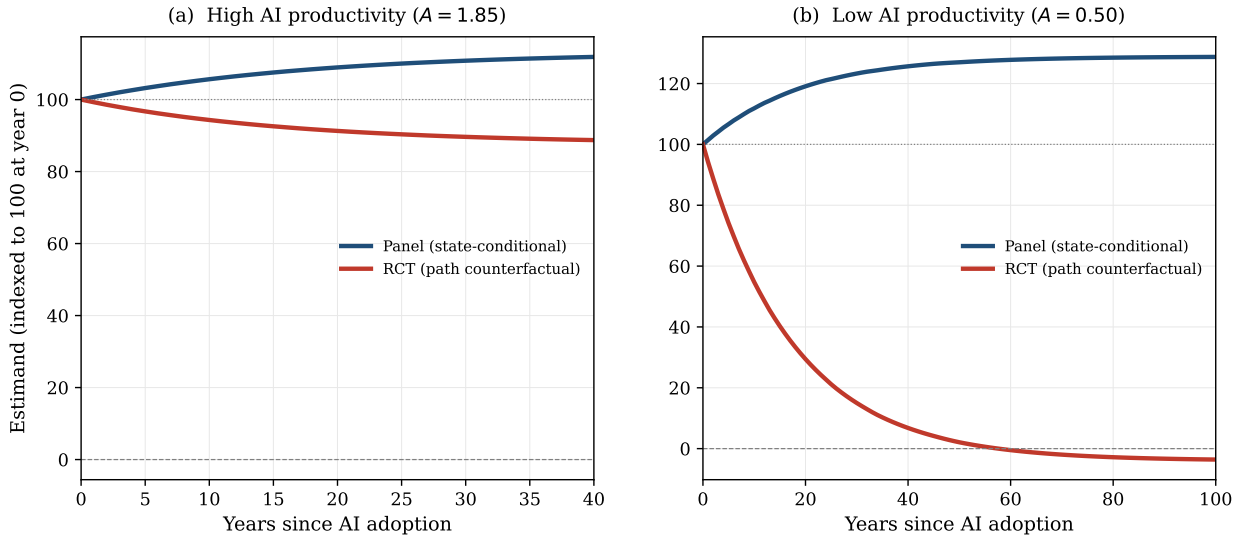

Figure 1: Scissors Pattern

Note: Panel (a): $A = 1.85$, initial $\alpha^*(\bar{h}) \approx 0.72$. Panel (b): $A = 0.50$, initial $\alpha^*(\bar{h}) \approx 0.24$. Other parameters: $\mu = 0.5$, $\beta = 0.95$, $\delta = 0.05$, $\lambda = 0.15$, $\gamma = 0.5$.

IV Conclusion

When treatment changes the law of motion for the state on which its effects are measured, causal estimands stop answering the same question. We characterize that divergence for AI delegation. Individual awareness of skill atrophy and continued delegation coexist at the optimum.

When $\mu < 1$, state-path divergence inflates estimates that condition on proxies for current skill, and spillover bias additionally contaminates cross-sectional comparisons when learning is socially produced. The wedge grows with adoption duration. The numerical illustration shows that it becomes economically meaningful within a decade at moderate values of μ. The scissors pattern – panel estimates rising while long-horizon RCT estimates fall – is the model's sharpest testable prediction, and it is unique to skill atrophy among the benchmark alternatives.

The framework yields concrete empirical priorities. Delayed post-tests that measure unassisted performance after AI exposure discipline μ, especially when paired with observed exposure duration or multiple withdrawal horizons. Re-randomization or withdrawal designs separate skill atrophy from habit formation: both predict eventual recovery upon permanent withdrawal, but atrophy implies slow convergence back to $\bar{h}$ at a rate governed by the learning parameters, whereas habit formation implies geometric decay of the dependency stock.

The measurement mechanism also has distributional implications. Short-run experiments consistently find that AI disproportionately benefits less-skilled workers (Brynjolfsson et al., 2025; Noy and Zhang, 2023; Peng et al., 2023), a finding variously interpreted as skill democratization (Autor, 2024) or as a shift in skill premia when AI automates tasks that previously required rare capabilities (Agrawal et al., 2024). When $\mu < 1$, however, this compression can mask longer-run divergence: higher-skilled workers delegate less and retain more skill, so the initial leveling effect reverses over time. Workers trained before widespread AI adoption carry skills that later cohorts accumulate more slowly. Unlike static task-based models where the skill distribution is exogenous to technology adoption (Acemoglu and Restrepo, 2018, 2020), in our framework the adoption decision *is* the skill investment decision. The vintage premium is testable using wage data linked to AI adoption timing, and the reversal of short-run compression using longitudinal skill assessments stratified by pre-adoption ability.

Our findings speak to a broader theoretical literature identifying private–social wedges in AI adoption: Hemenway Falk and Tsoukalas (2026) show that demand externalities across firms generate collective over-automation, and Acemoglu et al. (2026b) show that the

public-good nature of collective learning generates under-investment in human cognition. The layoff and wage dynamics in Hemenway Falk and Tsoukalas (2026) are a natural downstream counterpart to the skill-erosion mechanism we characterize here: eroded skill and eroded earnings are the same phenomenon measured at two different points in the labor market. We focus instead on the *measurement* on which welfare assessments of any of those models rest. When the same dynamics that create a private–social wedge also corrupt the estimands used to detect it, evidence-based policy becomes harder than short-horizon estimates would suggest. A policymaker reading the state-conditional literature without adjustment would see AI's measured productivity rising even as the skill base hollows out – precisely the configuration in which the argument for corrective action is strongest, and least likely to be perceived.

The framework does not call for a new class of experiments so much as a longer panel on the same workers. The design that breaks the identification problem is straightforward in principle: a durable no-access control combined with periodic unassisted post-tests of the treated group, sustained across multiple measurement intervals. Existing studies provide several ingredients of such a design; none yet combine them. In the meantime, the scissors prediction yields a diagnostic short of a definitive test: across sectors where both panel and randomized estimates can be read, the conjunction of a rising panel estimate and a falling randomized estimate over time is consistent only with skill atrophy among the benchmark alternatives we consider.

Last, because pedagogical quality is a property of how AI tools are deployed, not of AI itself, the measurement problem identified here is also a design opportunity. Tools that preserve learning content – through structured feedback, periodic unassisted work, or interfaces that couple delegation with explicit instruction – can raise μ and attenuate both biases. Pedagogical quality thus provides a concrete microfoundation for the pro-worker AI case of Acemoglu et al. (2026a): tools with high μ preserve and extend the expertise they leverage, whereas tools with low μ erode it, and the measurement biases characterized here make the distinction harder to detect ex ante than ex post.

Appendix: Proofs of Measurement Results

Proof of Proposition 3. Part (i). Since $Y(h, 0) = h$,

$$\Delta_t^{SC} - \Delta_t^{PATH} = \big[Y(h_t^U, \alpha_t) - Y(h_t^U, 0)\big] - \big[Y(h_t^U, \alpha_t) - Y(h_t^{NA}, 0)\big] = h_t^{NA} - h_t^U.$$

The no-adoption path stays at its steady state, so $h_t^{NA} = \bar{h}$ for all t. Proposition 2 gives $h_t^U < \bar{h}$ for every $t \geq 1$, proving strict positivity.

Part (ii). By Proposition 2, h_t^U is monotonically decreasing until the steady state, so $\bar{h} - h_t^U$ is strictly increasing.

Part (iii). By Proposition 2, the equilibrium path has $\alpha_t > 0$ for every t. Suppose $\Delta_t^{SC} \leq 0$ at some such t. Then choosing α_t yields weakly lower current output than $\alpha = 0$. Because $\mu < 1$, it also yields strictly lower next-period skill:

$$L(\alpha_t, h_t^U; \mu) < L(0, h_t^U; \mu).$$

Since V is increasing in h (Supplemental Lemma 3), switching to $\alpha = 0$ would raise both current output and continuation value, contradicting optimality. Hence $\Delta_t^{SC} > 0$ for all t. □

Proof of Proposition 4. Let $\varpi_t \in (0, 1)$ be the user share. Since $\psi(\bar{H}) = 1$, non-users are unaffected in period 1: $h_1^{NU} = (1 - \delta)\bar{h} + \lambda\varphi(\bar{h}) = \bar{h}$. Users adopt at positive intensity, so $h_1^U < \bar{h}$ when $\mu < 1$. Aggregate skill falls:

$$H_1 = \varpi_1 h_1^U + (1 - \varpi_1)\bar{h} < \bar{H},$$

which gives $\psi(H_1) < 1$ and therefore $h_2^{NU} = (1 - \delta)\bar{h} + \lambda\varphi(\bar{h})\psi(H_1) < \bar{h}$.

Let $T_0(h) = (1 - \delta)h + \lambda\varphi(h)$, increasing with fixed point $\bar{h}$. For users, if $h_t^U < \bar{h}$:

$$h_{t+1}^U = (1 - \delta)h_t^U + \lambda\ell(\alpha_t)\varphi(h_t^U)\psi(H_t) < T_0(h_t^U) < \bar{h}$$

since $\ell(\alpha_t) \leq 1$ and $\psi(H_t) \leq 1$ with at least one strict. Hence $H_t < \bar{H}$ for all $t \geq 1$.

For non-users, once $h_t^{NU} < \bar{h}$, the same bound gives $h_{t+1}^{NU} < T_0(h_t^{NU}) < \bar{h}$. Induction from $h_2^{NU} < \bar{h}$ gives $h_t^{NU} < \bar{h}$ for all $t \geq 2$, so

$$\Delta_t^{CS} - \Delta_t^{LR} = \bar{h} - h_t^{NU} > 0 \quad \text{for all } t \geq 2.$$

If the spillover dynamics converge with a positive limiting user share, $H^* < \bar{H}$ and the non-user fixed point satisfies $h^{NU*} < \bar{h}$, making the wedge permanent. □

Proof of Proposition 5. When $\eta = 0$, the no-adoption path remains at $\bar{h}$, so

$$\tau^{RCT}(t) = Y(h_t^U, \alpha_t) - Y(\bar{h}, 0) = \Delta_t^{PATH}.$$

Proposition 3(i) gives

$$\Delta_t^{SC} - \tau^{RCT}(t) = \Delta_t^{SC} - \Delta_t^{PATH} = \bar{h} - h_t^U > 0$$

for every $t \geq 1$. Proposition 3(ii) then implies the gap is strictly increasing until the steady state. □

Proof of Proposition 6. Part (i). Write $\Delta^{SC}(h) = Y(h, \alpha^*(h)) - h$. The Bellman equation and continuity of V imply $Y(h, \alpha^*(h)) = V(h) - \beta V(T(h))$ is continuous in h, so Δ^{SC} is continuous. Since α^* is monotone, it is differentiable almost everywhere. At the Bellman optimum, the first-order condition gives $Y_\alpha(h, \alpha^*(h)) = \beta V'(T(h))\lambda(1-\mu)\varphi(h) > 0$ (Supplemental Lemma 3). At every h where α^* is differentiable,

$$\frac{d}{dh}\Delta^{SC}(h) = \underbrace{(1-\alpha^*(h))^{1-\gamma} - 1}_{<0} + \underbrace{Y_\alpha(h, \alpha^*(h))}_{>0} \cdot \underbrace{\alpha^{*\prime}(h)}_{\leq 0} < 0.$$

Since Δ^{SC} is continuous with strictly negative derivative a.e., it is strictly decreasing. As h_t falls (Proposition 2), Δ_t^{SC} rises.

Part (ii). The Bellman equation gives $Y(h, \alpha^*(h)) = V(h) - \beta V(T(h))$. Both V and T are differentiable almost everywhere (the former by convexity, the latter because α^* is monotone). At a.e. h:

$$\frac{d}{dh}Y(h, \alpha^*(h)) = V'(h) - \beta V'(T(h)) \cdot T'(h).$$

On the equilibrium path $T(h) \leq h$ (Proposition 2), and the first slope condition of Assumption 1 gives $T'(h) > 0$. Since V is convex (Supplemental Lemma 3), $V'(T(h)) \leq V'(h)$, so $V'(T(h))\,T'(h) \leq V'(h)\,T'(h)$. Therefore

$$\frac{d}{dh}Y(h, \alpha^*(h)) \geq V'(h)\big[1 - \beta T'(h)\big] > 0.$$

Since $Y(h, \alpha^*(h))$ is continuous with strictly positive derivative a.e., it is strictly increasing. As h_t falls, $Y(h_t, \alpha_t)$ falls, and $\tau^{RCT}(t) = Y(h_t, \alpha_t) - \bar{h}$ is strictly decreasing.

Part (iii). By (ii), $\tau^{RCT}(t)$ converges monotonically to $Y(h^*, \alpha^*(h^*)) - \bar{h} < 0$. Since $\tau^{RCT}(0) > 0$, a finite crossing exists. □

Proof of Remark 2. Proposition 6(i) gives Δ_t^{SC} strictly increasing. Since h_t is decreasing (Proposition 2) and $\Delta_t^{SC} > 0$ for all t (Proposition 3(iii)), the ratio Δ_t^{SC}/h_t^U is strictly increasing. □

References

Acemoglu, D. (2025). The Simple Macroeconomics of AI. *Economic Policy* 40(121), 13–58.

Acemoglu, D., D. H. Autor, and S. Johnson (2026a). Building Pro-Worker Artificial Intelligence. The Hamilton Project, Brookings Institution.

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

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 (2024). The Turing Transformation: Artificial Intelligence, Intelligence Augmentation, and Skill Premiums. *Harvard Data Science Review*, Special Issue 5.

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

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

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

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

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

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

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

Becker, J., N. Rush, E. Barnes, and D. Rein (2025). Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. arXiv preprint arXiv:2507.09089.

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

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

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

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

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

Dell'Acqua, F., E. McFowland III, E. R. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani (2026). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *Organization Science* 37(2), 403–423.

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

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

Gans, J. S. (2026). A Model of Artificial Jagged Intelligence. Working Paper, Rotman School of Management, University of Toronto.

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

Paper.

Hemenway Falk, B. and G. Tsoukalas (2026). The AI Layoff Trap. arXiv preprint arXiv:2603.20617.

Ide, E. and E. Talamàs (2025). Artificial Intelligence in the Knowledge Economy. *Journal of Political Economy* 133(12), 3713–3749.

Jones, C. I. and C. Tonetti (2026). Past Automation and Future A.I.: How Weak Links Tame the Growth Explosion. *NBER Working Paper* 34779.

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

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

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

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

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

Poulidis, S., H. Bastani, and O. Bastani (2025). Self-Regulated AI Use Hinders Long-Term Learning. Working Paper.

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

Shen, J. H. and A. Tamkin (2026). How AI Impacts Skill Formation. *arXiv preprint arXiv:2601.20245*.

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

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

Supplemental Appendix (for Online Publication)

This appendix provides exhibits, numerical illustrations, and proofs for "Skill Atrophy and AI Productivity Measurement."

A Exhibits and Numerical Illustrations

This appendix collects notation, numerical estimates, and exhibits supporting the main text. All endogenous-adoption results are computed by value function iteration on the Bellman problem. Table S1 provides a notation guide; the narrative and tables that follow report dynamic paths, calibration results across μ, and spillover magnitudes.

Table S1: Notation Guide

Symbol	Definition
h, H	Individual / aggregate human capital
α	AI adoption intensity
A	AI productivity level
μ	Pedagogical quality (< 1: substitutes for learning; ≥ 1: augments)
γ	Returns to effort (effort concentration exponent)
δ, λ	Depreciation rate / learning intensity
β	Discount factor
η	Spillover elasticity
$\psi(H)$	Learning spillover function
$\bar{h}$	No-adoption steady-state skill: $\delta\bar{h} = \lambda\varphi(\bar{h})$
h^*	Steady-state skill under adoption
Δ_t^{CS}, Δ_t^{LR}	Cross-sectional / long-run productivity gain at t
Δ_t^{SC}, Δ_t^{PATH}	State-conditional / path counterfactual at t
$\tau^{RCT}(t)$, $\tau^{panel}(t)$	Long-horizon RCT / panel estimands
$W(\alpha)$	Steady-state output at constant adoption α
Γ	$\delta - \lambda\ell(\alpha^*)\varphi'(h^*) > 0$

Numerical Illustrations

This subsection complements the main-text discussion of numerical estimates with additional detail on the dynamic paths under the benchmark calibration ($\mu = 0.5$, $\beta = 0.95$). Functional forms are $\varphi(h) = 0.25/(1+h)$ and $g(j) = 1 - j$, with $\delta = 0.05$, $\lambda = 0.15$, $\gamma = 0.5$.

High-A Benchmark ($A = 1.85$)

Under value function iteration, the optimal policy $\alpha^*(h)$ is decreasing in h and the transition from $\bar{h}$ is monotone. Starting from $\bar{h} = 0.5$, the system converges to a steady state near $(h^*, \alpha^*) \approx (0.35, 0.77)$. Steady-state skill is substantially lower than under fixed adoption at $\alpha = 0.5$ ($h^*/\bar{h} \approx 0.70$ versus 0.80), reflecting the feedback between skill erosion and rising delegation. Table S2 reports how these results vary with μ, and Figure S1 plots the corresponding bias paths.

Lower-A Benchmark ($A = 0.50$)

At $A = 0.50$, the optimal adoption rate is much lower: $\alpha^*(\bar{h}) \approx 0.24$, converging to $\alpha^* \approx 0.29$ at steady state. Under this endogenous path, the RCT estimand does turn negative, at approximately year 58 (see also main-text Figure 1(b)).

Table S2: Calibration Results by Pedagogical Quality μ

	Pedagogical Quality μ				
Outcome	1.0	0.9	0.7	0.5	0.3
Steady-state adoption α^*	0.73	0.74	0.76	0.77	0.79
Steady-state skill $h^*/\bar{h}$	1.00	0.95	0.83	0.70	0.56
Bias at year 10 (%)	0.0	2.2	6.8	11.3	16.0
Bias at year 20 (%)	0.0	3.4	10.4	17.6	25.3

Note: All entries report the no-spillover case ($\eta = 0$). Bias defined as $(\Delta_t^{SC} - \Delta_t^{LR})/\Delta_0^{LR} \times 100$, the overstatement relative to the initial long-run gain. Parameters: $A = 1.85$, other parameters as in the numerical illustrations above. Spillovers (Proposition 4) would add cross-sectional mismeasurement on top of the individual-level state-path wedge. Replication code available from the authors.

Spillover Magnitudes

Table S3 reports the spillover skill gap $s_t = \bar{h} - h_t^{NU}$ under symmetric adoption with user share $\varpi = 0.5$ for several values of η. The spillover channel is quantitatively second-order under the benchmark: even at $\eta = 0.50$, the gap reaches only about 1.5% of the initial gain by year 20, compared to over 18% for the state-path wedge alone.

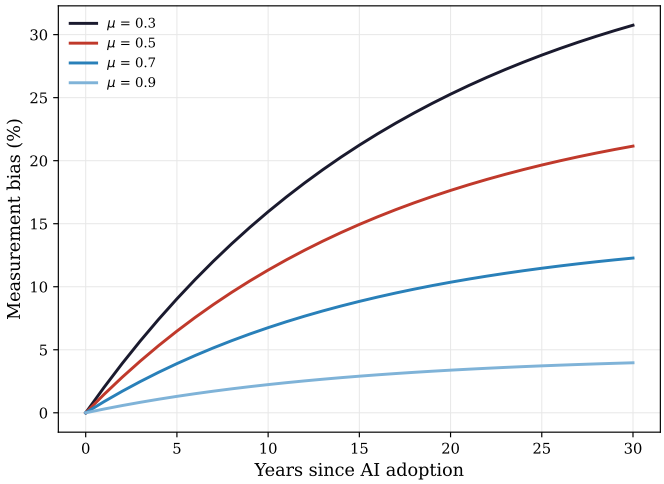

Figure S1: Measurement Bias Over Time

Note: Bias defined as $(\Delta_t^{SC} - \Delta_t^{LR})/\Delta_0^{LR} \times 100$, under endogenous adoption. Parameters as in Table S2.

Table S3: Spillover Skill Gap by Year and Spillover Elasticity

	Spillover elasticity η			
	0.05	0.10	0.20	0.50
Gap at year 10 (% of $\bar{h}$)	0.1	0.1	0.2	0.5
Gap at year 20 (% of $\bar{h}$)	0.1	0.3	0.6	1.6
Gap at year 20 (% of Δ_0^{LR})	0.1	0.3	0.6	1.5

Note: Symmetric equilibrium with adoption share $\varpi = 0.5$, $\mu = 0.5$, other parameters as in Table S2. Gap is $\bar{h} - h_t^{NU}$, the skill deficit of non-users relative to the no-adoption counterfactual. For comparison, the state-path wedge at year 20 is approximately 18% of Δ_0^{LR}.

B Alternative Mechanisms and Estimand Dynamics

This appendix formalizes the three alternative mechanisms discussed in the main text and shows their estimand dynamics differ qualitatively from skill atrophy. For each, we specify a minimal model, derive the state-conditional (SC) and RCT estimands, and characterize their time paths. Each adopts a deliberately simplified variant of the main model's output structure, isolating a single alternative channel.

Model 1: Habit Formation. Skill is fixed at $\bar{h}$. AI use builds a *habit stock* z_t satisfying

$$z_{t+1} = \rho_z z_t + \alpha_t, \qquad z_0 = 0, \quad \rho_z \in (0, 1).$$

The habit stock represents AI dependency. When working with AI at intensity $\alpha > 0$, output

is $AG(\alpha) + \bar{h}(1-\alpha)^{1-\gamma}$ as in the baseline: the habit stock does not affect output while AI is in use, so dependency is invisible until withdrawal. When working without AI ($\alpha = 0$), output is $\bar{h} - \kappa z_t$, where $\kappa > 0$ scales the withdrawal penalty.

SC estimand. The state-conditional comparison turns AI on vs. off at current state z_t:

$$\Delta_t^{SC,H} = \left[AG(\alpha) + \bar{h}(1-\alpha)^{1-\gamma}\right] - \left[\bar{h} - \kappa z_t\right] = AG(\alpha) - \bar{h}[1-(1-\alpha)^{1-\gamma}] + \kappa z_t.$$

At constant α, z_t is increasing, so $\Delta_t^{SC,H}$ is increasing in t: AI appears more valuable as the dependency deepens, because the unaided outside option includes the withdrawal penalty.

RCT estimand. Comparing a treated worker (using AI at α) to a control (never using AI, $z_t^C = 0$):

$$\tau_t^{RCT,H} = \left[AG(\alpha) + \bar{h}(1-\alpha)^{1-\gamma}\right] - \bar{h} = AG(\alpha) - \bar{h}[1-(1-\alpha)^{1-\gamma}].$$

The control has $z = 0$ and output $\bar{h}$. The treated worker's output while using AI does not depend on z_t, so $\tau_t^{RCT,H}$ is *constant* in t.

After withdrawal. If AI is removed at date T, the treated worker's unaided output drops below $\bar{h}$ by κz_T, but since z_t decays geometrically at rate ρ_z, performance recovers to $\bar{h}$ as $t \to \infty$. Under skill atrophy, recovery eventually occurs but is governed by the slow learning dynamics: the worker must rebuild human capital through unaided practice, a process that can take years or decades at the benchmark calibration. Under habit formation, recovery is rapid because z_t decays geometrically at rate ρ_z without requiring any active skill rebuilding.

The SC estimand rises while the RCT estimand is constant. The key distinction from skill atrophy is that habit formation does not change the worker's underlying productive capacity – it only creates a withdrawal cost. A long-horizon RCT, which never withdraws AI from the treated group, therefore sees no change over time. No scissors pattern emerges. This result reflects the *withdrawal-penalty* structure imposed here; alternative habit formulations – where dependency raises AI-assisted throughput or distorts the effective $g(\cdot)$ – can generate different estimand dynamics.

Model 2: Complementary Learning (Learning Curves). AI use generates complementary expertise s_t (e.g., prompt engineering), with

$$s_{t+1} = (1-\delta_s)s_t + \lambda_s \alpha_t, \qquad s_0 = 0.$$

Baseline skill $\bar{h}$ is unchanged. Output with AI is augmented:

$$Y^L(\bar{h}, \alpha, s) = A(1+s)G(\alpha) + \bar{h}(1-\alpha)^{1-\gamma}.$$

Without AI: $Y^L(\bar{h}, 0, s) = \bar{h}$. The complementary expertise s boosts AI output but does not affect unaided work.

SC estimand.

$$\Delta_t^{SC,L} = A(1+s_t)G(\alpha) - \bar{h}[1-(1-\alpha)^{1-\gamma}].$$

Since s_t is increasing, $\Delta_t^{SC,L}$ is increasing: the worker becomes progressively better at using AI, so the state-conditional gain grows over time.

RCT estimand.

$$\tau_t^{RCT,L} = A(1+s_t^T)G(\alpha) - \bar{h}[1-(1-\alpha)^{1-\gamma}].$$

Since $s_t^T > s_t^C = 0$, the treated worker accumulates expertise the control does not, so $\tau_t^{RCT,L}$ is also increasing.

Both estimands rise because AI use builds complementary skills that make subsequent use more productive. Unlike skill atrophy, where the worker's unaided capacity erodes, complementary learning augments the worker's AI-assisted capacity without degrading unaided performance. No scissors pattern emerges.

Model 3: Selection on Early Adopters. Workers are heterogeneous in baseline productivity θ_i. There is no skill dynamics; each worker has fixed output θ_i without AI and $\theta_i + AG(\alpha)$ with AI. Workers adopt in order of productivity: the highest-θ workers adopt first (modeling classical positive selection; under the main text's substitution structure, negative selection would be more natural, but the formal conclusion is identical because the within-worker estimands remain constant regardless of adoption ordering). At time t, the marginal adopter has productivity $\theta(t)$, with $\theta'(t) < 0$ as adoption diffuses to less productive workers.

SC estimand. For any given worker, the state-conditional gain is $\Delta_t^{SC,S} = AG(\alpha)$, which is constant.

RCT estimand. In a homogeneous-effect model, the randomized treatment effect is also $AG(\alpha)$, so it is constant over time.

Observational analogue. As adoption diffuses to less productive workers, naive treated-vs.-untreated comparisons can decline because the average type in the treated pool falls, even though the formal within-worker and experimental estimands remain constant.

Both formal estimands are constant. Selection does not generate the scissors pattern; at most it makes naive observational comparisons decline as the adopter pool expands. Across all three alternatives, only skill atrophy generates one estimand rising while the other falls.

C Additional Proofs

This appendix collects technical lemmas and proofs not included in the main-text Appendix.

C.A Technical Lemmas

The Agent's Problem. Recall from Section II that the decision maker maximizes (4) subject to (2), with value function defined by (5).

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

Proof. The agent solves $\max \int_\alpha^1 h\, e(j)^\gamma\, dj$ subject to $\int_\alpha^1 e(j)\, dj = 1$. The FOC $h\gamma e(j)^{\gamma-1} = \xi$ gives constant $e(j) = 1/(1-\alpha)$, yielding $\int_\alpha^1 h(1-\alpha)^{-\gamma}\, dj = h(1-\alpha)^{1-\gamma}$. □

Lemma 2 (Output and Learning Properties)**.** *The output function $Y(h, \alpha; A) = AG(\alpha) + h(1-\alpha)^{1-\gamma}$ is linear in h, strictly concave in α for $h > 0$, and satisfies $Y_\alpha(h, \alpha) \to -\infty$ as $\alpha \to 1^-$. The learning effect satisfies $L_\alpha(\alpha, h; \mu) = (\mu - 1)\varphi(h)$.*

Proof. $Y_{\alpha\alpha} = Ag'(\alpha) - h(1-\gamma)\gamma(1-\alpha)^{-\gamma-1} < 0$ because $g' < 0$ and $\gamma \in (0, 1)$. Also $Y_\alpha = Ag(\alpha) - h(1-\gamma)(1-\alpha)^{-\gamma} \to -\infty$ as $\alpha \to 1^-$. The expression for L_α follows from (3). □

Lemma 3 (Value Function Properties)**.** *On the invariant compact set used in the main text, the Bellman operator is a contraction. Hence the value function exists, is unique, continuous, strictly increasing, convex, and L_V-Lipschitz with*

$$L_V \equiv \frac{1}{1 - \beta(1-\delta)}.$$

As a convex Lipschitz function, V is absolutely continuous; its derivative exists almost everywhere and satisfies $0 < V'(h) \leq L_V$ at every differentiability point in the interior. At states with a unique interior optimizer, the Bellman-envelope condition yields differentiability.

Proof. On the invariant compact set, $\beta \in (0, 1)$ makes the Bellman operator a contraction; existence, uniqueness, and continuity follow from Stokey and Lucas (1989, Theorem 4.6). For any fixed α, the transition

$$T(h, \alpha) = (1-\delta)h + \lambda\ell(\alpha)\varphi(h)$$

satisfies $T_h(h,\alpha) \le 1-\delta$ (since $\varphi' < 0$) and $T_h(h,\alpha) > 0$ by Assumption 1. If v is increasing and L_V-Lipschitz, then for any $h_2 > h_1$,

$$Y(h_2,\alpha) - Y(h_1,\alpha) \le h_2 - h_1, \qquad v(T(h_2,\alpha)) - v(T(h_1,\alpha)) \le L_V(1-\delta)(h_2-h_1).$$

Choosing $L_V = 1/[1-\beta(1-\delta)]$ makes the class of increasing L_V-Lipschitz functions invariant under the Bellman operator, so the fixed point V inherits those properties.

Strict monotonicity: for $h_2 > h_1$, evaluating the Bellman objective at h_2 using the policy $\alpha^*(h_1)$ gives $V(h_2) \ge Y(h_2,\alpha^*(h_1)) + \beta V(T(h_2,\alpha^*(h_1))) > Y(h_1,\alpha^*(h_1)) + \beta V(T(h_1,\alpha^*(h_1))) = V(h_1)$, where the strict inequality uses Y strictly increasing in h and T nondecreasing in h. Hence $V'(h) > 0$ at every interior differentiability point.

Convexity is preserved as well. Assumption 1 gives $\varphi''(h) = 2c/(1+h)^3 > 0$, so $T(h,\alpha)$ is convex in h for every fixed α. If v is increasing and convex, then $v \circ T(\cdot,\alpha)$ is convex, and $Y(h,\alpha)$ is linear in h. Hence, for each fixed α, the Bellman objective is convex in h, and taking the pointwise maximum over α preserves convexity. Starting value iteration from a convex function (for instance $v_0 \equiv 0$), every iterate is convex; the contraction's uniform limit V is therefore convex.

Since V is convex and L_V-Lipschitz, it is absolutely continuous and satisfies $0 < V'(h) \le L_V$ at interior differentiability points. Interior optimality plus the Bellman envelope gives differentiability at the states used below. □

Lemma 4 (Sufficient Condition for Initial Interiority)**.** *If*

$$Ag(0) > \bar{h}\left[(1-\gamma) + \frac{\beta\delta(1-\mu)}{1-\beta(1-\delta)}\right],$$

then $\alpha^*(\bar{h}) \in (0,1)$*. Moreover, no optimum can occur at* $\alpha = 1$*.*

Proof. By Lemma 3, the value function is L_V-Lipschitz with $L_V = 1/[1-\beta(1-\delta)]$. Hence whenever the Bellman objective is differentiable at $\alpha = 0$,

$$\beta V'(\bar{h})\lambda(1-\mu)\varphi(\bar{h}) \le \frac{\beta\lambda(1-\mu)\varphi(\bar{h})}{1-\beta(1-\delta)} = \bar{h}\frac{\beta\delta(1-\mu)}{1-\beta(1-\delta)}.$$

The stated inequality therefore implies that the derivative of the Bellman objective at $(\bar{h},0)$ is strictly positive, ruling out $\alpha = 0$. By Lemma 2, $Y_\alpha \to -\infty$ as $\alpha \to 1^-$, so $\alpha = 1$ cannot be optimal. □

Lemma 5 (Steady-State Human Capital at Fixed Adoption)**.** *For any fixed $\alpha \in [0,1)$ with $\ell(\alpha) > 0$, the equation*

$$\delta h = \lambda \ell(\alpha)\varphi(h)$$

has a unique solution $h_*(\alpha) > 0$. Moreover,

$$\frac{dh_*(\alpha)}{d\alpha} = -\frac{\lambda(1-\mu)\varphi(h_*(\alpha))}{\delta - \lambda\ell(\alpha)\varphi'(h_*(\alpha))},$$

and when $\mu < 1$ and $\alpha > 0$, $h_*(\alpha) < \bar{h}$.

Proof. Existence and uniqueness follow from Assumption 1. Applying the IFT to $F(h;\alpha) \equiv \delta h - \lambda\ell(\alpha)\varphi(h) = 0$ with $F_h = \delta - \lambda\ell(\alpha)\varphi'(h) > 0$ (since $\varphi' < 0$) gives $dh_*/d\alpha = -F_\alpha/F_h = -\lambda(1-\mu)\varphi(h_*)/[\delta - \lambda\ell(\alpha)\varphi'(h_*)]$. When $\mu < 1$ and $\alpha > 0$, $\ell(\alpha) < 1$ implies $\delta h_* = \lambda\ell(\alpha)\varphi(h_*) < \lambda\varphi(h_*)$; since $\bar{h}$ is the unique root of $\delta h = \lambda\varphi(h)$, we have $h_*(\alpha) < \bar{h}$. □

Lemma 6 (Fixed-α Dynamics)**.** *Fix $\alpha \in (0,1)$ and let*

$$T_\alpha(h) \equiv (1-\delta)h + \lambda\ell(\alpha)\varphi(h).$$

Under Assumption 1 and $\mu < 1$, T_α is increasing on $[0,\bar{h}]$, satisfies $T_\alpha(\bar{h}) < \bar{h}$, and the path $h_{t+1} = T_\alpha(h_t)$ from $h_0 = \bar{h}$ decreases monotonically to the unique fixed point $h_*(\alpha) \in (0,\bar{h})$.

Proof. Since $\ell(\alpha) \in (0,1)$, Assumption 1 gives T_α increasing on $[0,\bar{h}]$ with unique fixed point $h_*(\alpha) \in (0,\bar{h})$ (Lemma 5). Also $T_\alpha(\bar{h}) = (1-\delta)\bar{h} + \lambda\ell(\alpha)\varphi(\bar{h}) < \bar{h}$ since $\ell < 1$. If $T_\alpha(h) \geq h$ for some $h > h_*(\alpha)$, continuity and $T_\alpha(\bar{h}) < \bar{h}$ would give a second fixed point, contradicting uniqueness; hence $T_\alpha(h) < h$ for $h > h_*(\alpha)$. Monotonicity gives $T_\alpha(h) \geq h_*(\alpha)$, so $\{h_t\}$ is decreasing and bounded below, converging to $h_*(\alpha)$. □

C.B Remaining Proofs

Propositions 3–6 and Remark 2 are proved in the main-text Appendix.

Proof of Proposition 1. At an interior optimum, the Bellman first-order condition is

$$Y_\alpha(h,\alpha) + \beta V'(h')\lambda(\mu-1)\varphi(h) = 0, \qquad h' = (1-\delta)h + \lambda\ell(\alpha)\varphi(h).$$

Rearranging:

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

By Supplemental Lemma 3, $V'(h') > 0$ at differentiability points, and $\varphi(h) > 0$. When $\mu < 1$, the right-hand side is strictly positive, so $Y_\alpha(h, \alpha^*) > 0$; strict concavity of Y in α (Lemma 2) then gives $\alpha^* < \alpha^m(h)$, the myopic level solving $Y_\alpha(h, \alpha) = 0$. When $\mu = 1$, the right-hand side vanishes and $\alpha^* = \alpha^m(h)$. □

Proof of Proposition 2. Let $L_V \equiv 1/[1 - \beta(1 - \delta)]$. Supplemental Lemma 3 shows that V is increasing, convex, and L_V-Lipschitz. For $h_2 > h_1$ and $\alpha_2 > \alpha_1$, write $Q(h, \alpha) = Y(h, \alpha; A) + \beta V\big((1 - \delta)h + \lambda\ell(\alpha)\varphi(h)\big)$. The static cross-difference is

$$\begin{aligned} &Y(h_2, \alpha_2) - Y(h_2, \alpha_1) - Y(h_1, \alpha_2) + Y(h_1, \alpha_1) \\ &\quad = (h_2 - h_1)\Big[(1 - \alpha_2)^{1-\gamma} - (1 - \alpha_1)^{1-\gamma}\Big] \\ &\quad \leq -(1 - \gamma)(h_2 - h_1)(\alpha_2 - \alpha_1). \end{aligned}$$

For the continuation term, define $D(h) \equiv V(T(h, \alpha_1)) - V(T(h, \alpha_2))$ where $T(h, \alpha) = (1 - \delta)h + \lambda\ell(\alpha)\varphi(h)$. Let $w(h) \equiv T(h, \alpha_1) - T(h, \alpha_2) = \lambda(1 - \mu)(\alpha_2 - \alpha_1)\varphi(h) > 0$. Because $\ell(\alpha)$ is affine in α, for each $s \in [0, 1]$ we can write $\xi(h, s) \equiv T\big(h, (1 - s)\alpha_2 + s\alpha_1\big)$, so that, using the absolute continuity of V from Lemma 3,

$$D(h) = \int_0^1 V'(\xi(h, s))\, w(h)\, ds.$$

Now fix $h_2 > h_1$ and write $w_i = w(h_i)$ and $\xi_i(s) = \xi(h_i, s)$. Since $T(\cdot, \alpha)$ is increasing for every fixed α by Assumption 1, we have $\xi_1(s) < \xi_2(s)$ for all $s \in [0, 1]$. Since V is convex, V' is nondecreasing almost everywhere, so $V'(\xi_1(s)) \leq V'(\xi_2(s))$ a.e. Therefore,

$$\begin{aligned} D(h_1) - D(h_2) &= \int_0^1 \Big[V'(\xi_1(s))(w_1 - w_2) + \big(V'(\xi_1(s)) - V'(\xi_2(s))\big)w_2\Big]\, ds \\ &\leq L_V\lambda(1 - \mu)(\alpha_2 - \alpha_1)\big[\varphi(h_1) - \varphi(h_2)\big] \\ &\leq L_V\lambda(1 - \mu)\varphi(0)(h_2 - h_1)(\alpha_2 - \alpha_1), \end{aligned}$$

where the last inequality uses $\varphi(h) = c/(1 + h)$ and $\sup_h |\varphi'(h)| = \varphi(0) = c$. The second condition of Assumption 1 gives $\beta L_V\lambda\varphi(0) < 1$. Combined with $1 - \mu \leq 1 - \gamma$ from Assumption 2, the continuation cross-difference is strictly less than $(1 - \gamma)(h_2 - h_1)(\alpha_2 - \alpha_1)$, so Q has decreasing differences in (h, α). By Topkis monotonicity, the argmax correspondence admits a decreasing optimal stationary selection $\alpha^*(h)$.

The interior-adoption condition implies $\alpha^*(\bar{h}) > 0$. Since $\mu < 1$ and $\ell(\alpha) < 1$ for every $\alpha > 0$, $h_1 = (1 - \delta)\bar{h} + \lambda\ell(\alpha^*(\bar{h}))\varphi(\bar{h}) < \bar{h}$. Now define the induced transition

$T(h) \equiv (1-\delta)h + \lambda\ell(\alpha^*(h))\varphi(h)$. If $h_2 > h_1$, then $\alpha^*(h_2) \leq \alpha^*(h_1)$ and therefore $\ell(\alpha^*(h_2)) \geq \ell(\alpha^*(h_1))$, so

$$T(h_2) - T(h_1) \geq (1-\delta)(h_2 - h_1) + \lambda\ell(\alpha^*(h_1))[\varphi(h_2) - \varphi(h_1)] \geq [1 - \delta - \lambda\varphi(0)](h_2 - h_1) > 0,$$

where the last inequality uses Assumption 1. Hence T is increasing. Since $h_1 = T(\bar{h}) < \bar{h}$, the sequence $\{h_t\}$ is decreasing and bounded, converging to some $h^* \geq 0$. Because $\alpha^*(h)$ is decreasing and h_t is decreasing, $\alpha_t = \alpha^*(h_t)$ is nondecreasing, converging to some $\alpha^* \in (0, 1]$. The primitive law of motion $h_{t+1} = (1-\delta)h_t + \lambda\ell(\alpha_t)\varphi(h_t)$ is continuous in (h_t, α_t); passing to the limit gives $h^* = (1-\delta)h^* + \lambda\ell(\alpha^*)\varphi(h^*)$. The limit cannot have $\alpha^* = 1$: if $\alpha^* = 1$, then $\ell(1) = \mu > 0$ (since $\mu \geq \gamma > 0$ by Assumption 2), so the stationarity condition has a unique positive root $h^* > 0$; but $Y_\alpha(h^*, 1) = -\infty$, contradicting optimality. Hence $\alpha^* \in (0, 1)$. Since $\ell(\alpha^*) > 0$, the stationarity condition has a unique positive root, so $h^* > 0$. □

Derivation of $W'(\alpha^*)$ (equation (8)). Steady-state output is $W(\alpha) = AG(\alpha) + h_*(\alpha)(1-\alpha)^{1-\gamma}$. At the dynamic optimum, the Bellman objective is strictly concave in α at h^* (since the maintained conditions bound the convex continuation term below the static concavity $Y_{\alpha\alpha} < 0$), so the optimizer is unique and the envelope theorem applies. The first-order condition gives $Ag(\alpha^*) - h^*(1-\gamma)(1-\alpha^*)^{-\gamma} = \beta V'(h^*)\lambda(1-\mu)\varphi(h^*)$, and the steady-state envelope condition yields $V'(h^*) = (1-\alpha^*)^{1-\gamma}/[(1-\beta) + \beta\Gamma]$ where $\Gamma \equiv \delta - \lambda\ell(\alpha^*)\varphi'(h^*) > 0$. Using $dh_*/d\alpha = -\lambda(1-\mu)\varphi(h^*)/\Gamma$ from Lemma 5 and substituting:

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

□