Preventing Skill Extinction Under High Automation: Minimum Human-in-the-Loop Practice Quotas and Endogenous Mentor–Apprentice Lineage Incentives

Dima Baker
Collective Intelligence Network

November 2026

Abstract

As autonomous coding and analytical agents absorb routine production tasks, organizations confront a structural human-capital externality: short-term delegation erodes the deliberate-practice loops through which novices acquire tacit mastery and senior experts maintain diagnostic acuity. Building on general-skills training theory (Becker, 1964; Acemoglu & Pischke, 1998) and Bainbridge's (1983) "Ironies of Automation," this modeling paper formulates intergenerational skill dynamics as a coupled differential game over senior mentor capital M_t, apprentice mastery S_t, and endogenous mentoring effort μ_t^*, and specifies the Tacit Craft Preservation and Apprenticeship Escrow (TCP-AE) mechanism. TCP-AE pairs a Minimum Human-in-the-Loop Practice Quota ($\underline{q}_t \geq 0.18$) with a Lineage Escrow contract that vests a share λ_{esc} of team automation surplus to mentors conditionally upon apprentices passing blind out-of-distribution (OOD) evaluations. We prove uniqueness and infeasibility boundary conditions for the critical practice quota $\underline{q}^*(\mu)$ and execute a 9,600-cohort simulation ($N = 600$ pairs per cell across 4 regimes and 4 automation levels $\alpha \in \{0.50, 0.70, 0.85, 0.95\}$) with identical quota and rework equations across arms. At $\alpha = 0.85$, unconstrained automation and passive review reduce simulated 24-month mastery pass rates to 18.67% (95% Wilson CI $[15.75\%, 21.98\%]$) and 23.17% ($[19.97\%, 26.71\%]$). An unpaired 18% solo quota ($\mu^* = 0.15$) yields 31.83% ($[28.23\%, 35.67\%]$), whereas TCP-AE ($\lambda_{\text{esc}} = 0.14$) endogenously induces $\mu^* = 0.81$, sustaining 51.33% ($[47.34\%, 55.31\%]$) mastery and 54.50% ($[50.50\%, 58.44\%]$) OOD recovery at a 2.8% velocity trade-off while evaluating imperfect OOD tests and mid-task handovers.

1 Introduction

A foundational tension in human–computer symbiosis [6, 21, 24] has resurfaced with quantitative urgency in 2025–2026: the very tasks that are easiest to delegate to frontier language model agents—modular implementation, routine debugging, preliminary data synthesis, and boilerplate drafting—are the exact developmental rungs by which junior apprentices climb toward deep architectural competence [2, 5, 26]. More than four decades ago, Bainbridge [6] warned in her classic analysis of the "Ironies of Automation" that automating routine control loops deskills human operators while simultaneously restricting their intervention to rare, high-complexity edge cases that demand greater—not lesser—system mastery. Concurrently, in labor economics, Becker [8] and Acemoglu and Pischke [1] established that competitive firms systematically underinvest in generalizable apprentice skills whenever senior practitioners bear the immediate opportunity cost of mentoring while downstream skill returns are non-excludable.

Throughout 2025 and 2026, empirical studies across software engineering, workforce task auditing, and human–AI delegation have demonstrated that these twin mechanisms—cognitive atrophy from offloading and organizational underinvestment in apprenticeship—reinforce one another when AI agents scale across production workflows [19, 22, 25].

1.1 Empirical evidence of cognitive divergence and training externalities

Recent experimental and longitudinal findings identify four converging failure modes when organizations maximize short-horizon agent delegation without structural practice or mentorship safeguards:

1. **The Delegation Feedback Loop and forgetting dynamics.** Eliav [13] documents a widening structural divergence between expanding AI context windows and human sustained attention, formalizing the *Delegation Feedback Loop*: as agent capability rises, the cognitive friction threshold at which humans offload tasks drops, accelerating the classic production-break forgetting curve characterized in operations research by Jaber and Bonney [16] and Hopp et al. [14]. When novice and senior workers cease active synthesis, working-memory schemas decay convexly with the offloaded task share [18, 27].

2. **Confidence without competence in passive validation.** When human roles shift from active production to reviewing AI-generated outputs, passive inspection fails to preserve diagnostic skill. Eleftheriou et al. [12] establish experimentally that AI-assisted knowledge workers exhibit systematic *confidence without competence*, reporting inflated subjective self-efficacy even as their unassisted accuracy degrades. Complementing this finding, Lee et al. [20] and Arghal et al. [4] show that strategic miscalibration and reduced critical verification effort emerge endogenously when workers rely on high-fluency generative assistants.

3. **Skill formation deficits among novices and newcomer barriers.** In a randomized study of programming skill acquisition, Shen and Tamkin [26] measure how AI assistance impacts human learning, demonstrating that unguarded reliance during early training stages suppresses independent problem-solving retention when assistance is removed. In open-source and enterprise software engineering, where newcomers already face steep architectural onboarding barriers [28], unverified AI pull requests and dependency gaps [30] compound review bottlenecks and can trigger workflow "death spirals" when senior maintainer capacity is overwhelmed [17].

4. **The general-training hold-up problem under automation.** Why do senior engineers and craft masters rarely slow down to teach juniors when AI tools are available? In the human-capital framework of Becker [8] and Acemoglu and Pischke [1], mentoring an apprentice requires a senior practitioner to divert time μ_t away from high-velocity automated output, incurring an immediate private opportunity cost while the trained apprentice's future diagnostic value is an organizational public good. Without an explicit mechanism that internalizes this training externality [3, 29], rational mentors choose minimal pedagogical coupling $\mu_t^* \to \mu_{\min}$, leaving even mandatory solo practice quotas insufficient to transmit tacit system knowledge.

1.2 Contributions and modeling scope

Rather than presenting a completed multi-year field trial, this paper contributes a formal economic and dynamical modeling framework—grounded in Becker [8], Acemoglu and Pischke [1], and Bainbridge [6]—that derives when and how deliberate-practice quotas must be paired with endogenous mentorship incentives:

1. **Synthesis of automation deskilling and human-capital training theory (Section 2).** We connect classical operations-research learning/forgetting models [14, 16] and general-training economics [1, 8] with 2025–2026 empirical evidence on AI skill formation [12, 13, 26] and inalienable human-in-the-loop (I-HITL) attestation [10] (Table 1).

2. **Coupled differential model with endogenous mentor optimization (Section 3).** We formulate a two-generation dynamical system over apprentice mastery S_t and mentor tacit capital M_t (Equations 1–2), prove the necessary and sufficient uniqueness condition for the critical minimum practice quota $\underline{q}^*(\mu)$ (Proposition 1, Equation 3), and derive the mentor's endogenous effort choice $\mu^*(\lambda_{\text{esc}})$ under a Lineage Escrow contract (Equation 4).

3. **TCP-AE protocol architecture (Section 4).** We specify a three-layer organizational workflow combining stratified practice-quota routing, I-HITL cryptographic countersignatures [10], and blind out-of-distribution (OOD) crucible settlement (Figure 1, Table 2).

4. **Illustrative mechanistic simulation and pure-escrow ablation (Section 5).** We execute our coupled ODE and endogenous mentor-choice solver (`experiment.py`) across 9,600 simulated 24-month cohorts ($N = 600$ per cell across 4 regimes and 4 automation ratios $\alpha \in \{0.50, 0.70, 0.85, 0.95\}$), holding the practice quota ($\underline{q} = 0.18$) and all rework/calibration equations strictly identical between `solo_quota` and `escrow_lineage` to isolate the pure endogenous escrow effect and map the break-even sensitivity curve over $\lambda_{\text{esc}} \in [0.00, 0.20]$ (Tables 3–4, Figures 2–3).

5. **Failure modes and field-test falsification protocol in dual apprenticeship programs (Sections 6–8).** We analyze velocity trade-offs, extrinsic-reward crowding out [11], and quota spoofing, closing with six pre-registerable falsification protocols anchored to German dual-system (BIBB) and union apprenticeship cohorts (Table 5).

2 Theoretical foundations: automation ironies, forgetting curves, and training externalities

Informal organizational policies often treat human-in-the-loop (HITL) oversight as a binary compliance checkbox: a human clicks "Approve" on an agent-generated pull request or diagnostic report. Both classical human-factors theory and 2025–2026 empirical work show why passive sign-off neither preserves operator skill nor transmits tacit craft knowledge. Table 1 maps five foundational mechanisms from labor economics, operations research, and cryptographic provenance onto the components of the Tacit Craft Preservation and Apprenticeship Escrow (TCP-AE) framework.

The central insight of Table 1 is that skill preservation under high automation is a *coupled control-and-incentive problem*. Enforcing a solo practice quota ($\underline{q}_t \geq 0.18$) without compensating senior mentors for pedagogical coupling (μ_t) leaves novices struggling in isolation, whereas paying flat mentoring stipends without blind OOD verification invites moral hazard. Figure 1 illustrates how TCP-AE couples quota routing with endogenous escrow settlement.

Table 1: Foundational mechanisms in human-capital theory, operations research, and 2025–2026 AI skill studies underlying TCP-AE.

Failure mode	Foundational theory	TCP-AE mechanism	Operational role in workflow	Primary literature
Production-break skill atrophy	Learning–forgetting curves & automation irony	Minimum practice quota ($\underline{q}_t \geq 0.18$)	Reserves a calibrated fraction of structural tasks for unassisted human synthesis	[6, 13, 16, 26]
Validating vs. producing gap	Confidence without competence	Blind OOD diagnostic crucible ($\mathcal{C}_{\text{OOD}}$)	Evaluates unassisted transfer on held-out faults rather than passive diff review	[4, 12, 20]
General-training underinvestment	Becker / Acemoglu–Pischke training externality	Mentor–Apprentice Lineage Escrow ($\mathcal{E}_{\text{lin}}$)	Internalizes mentoring cost $c_M(\mu)$ by vesting team surplus $\lambda_{\text{esc}}Y$ on apprentice pass rate	[1, 8, 23]
Cross-training fragility	Serial skill chaining in operations	Paired mentor–novice trace steering (μ_t^*)	Couples senior tacit mental models with novice practice to lower critical quota $\underline{q}^*(\mu)$	[14, 28, 29]
Rubber-stamping & proxy spoofing	Inalienable Human-in-the-Loop (I-HITL)	Hardware-attested practice bouts	Binds human practice sessions and OOD exams to isolated hardware countersignatures	[10, 17, 30]

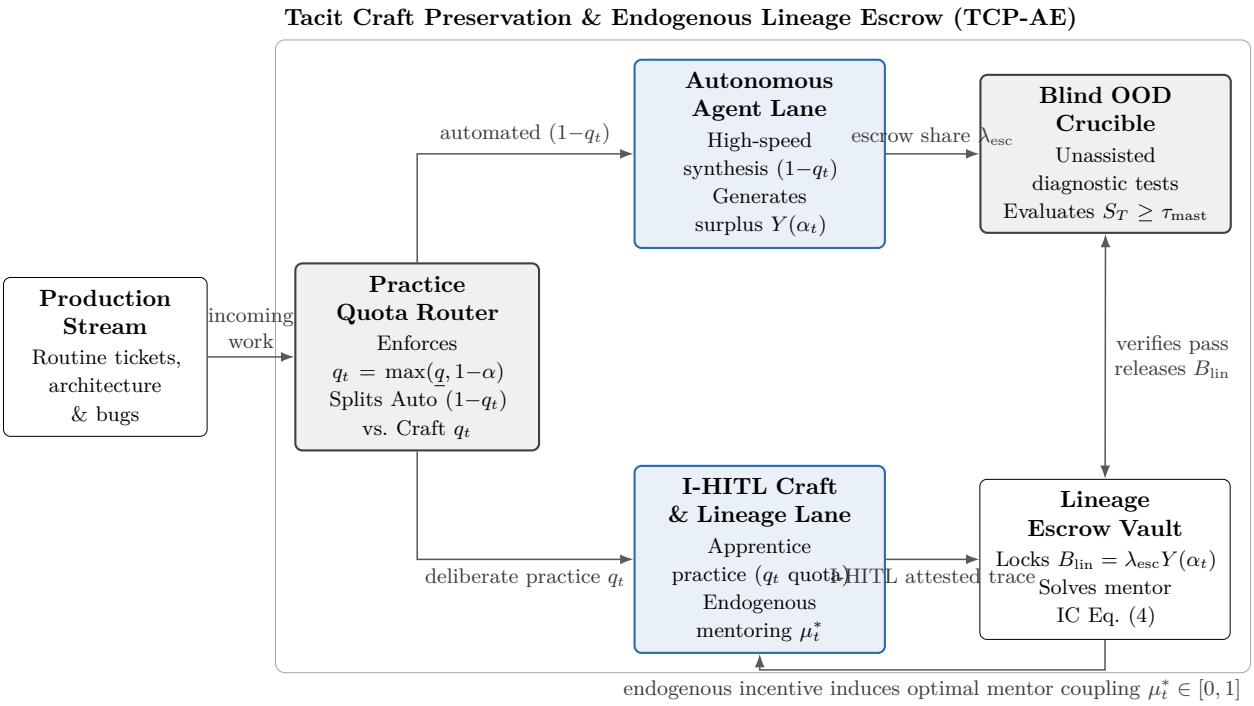


Figure 1: Architecture of the Tacit Craft Preservation and Apprenticeship Escrow (TCP-AE) protocol. Incoming tasks are partitioned by the Practice Quota Router: while fraction $1 - q_t$ executes via autonomous agents, at least $q_t = \max(\underline{q}, 1 - \alpha_t)$ of diagnostic work is routed to the Inalienable Human-in-the-Loop (I-HITL) Craft Lane. A fraction λ_{esc} of automation surplus $Y(\alpha_t)$ is locked in the Lineage Escrow Vault and vests to the mentor–apprentice pair when the apprentice passes periodic blind Out-of-Distribution (OOD) diagnostic crucibles, endogenously inducing high mentor coupling μ_t^*.

3 Formal mathematical framework

We formalize intergenerational skill retention, critical practice thresholds, and endogenous mentorship investment across discrete quarterly evaluation epochs $t \in \{0, 1, \ldots, T-1\}$.

3.1 Coupled mentor–apprentice skill evolution under automation

Consider an engineering or craft organization comprising senior mentors with tacit architectural mastery $M_t \in (0,1)$ and entering junior apprentices with unassisted generative skill $S_t \in (0,1)$. In each epoch t, routine task volume is automated at intensity $\alpha_t \in [0,1]$, leaving an effective human deliberate-practice fraction $q_t \in (0,1]$, while senior mentors choose a pedagogical coupling intensity $\mu_t \in [0,1]$ representing the share of junior practice bouts conducted with interactive mentor guidance [14, 23].

Following production learning–forgetting dynamics [16], the Delegation Feedback Loop of Eliav [13], and the skill-formation measurements of Shen and Tamkin [26], unassisted apprentice mastery evolves according to three competing forces: (i) concave diminishing returns from hands-on practice $\beta_P\sqrt{q_t}(1-S_t)$, (ii) vertical lineage transmission proportional to the tacit skill gradient $(M_t - S_t)_+ = \max(0, M_t - S_t)$ and mentor coupling μ_t, and (iii) convex forgetting/atrophy driven by task offloading $\delta_A(1-q_t)^\gamma$ with $\gamma > 1$:

$$S_{t+1} - S_t \;=\; \beta_P\sqrt{q_t}\,(1-S_t) \;+\; \beta_L\mu_t(M_t - S_t)_+ \;-\; \delta_A(1-q_t)^\gamma \;+\; \xi_t, \tag{1}$$

where $\beta_P, \beta_L, \delta_A > 0$ are structural rate parameters and $\xi_t \sim \mathcal{N}(0, \sigma_\xi^2)$ is zero-mean shock noise. Senior mentor capital M_t obeys a bounded $(1-M_t)$ and M_t scaled difference equation reflecting practice maintenance, pedagogical articulation gains ("learning by teaching"), and detachment decay:

$$M_{t+1} - M_t \;=\; \eta_P q_t(1-M_t) \;+\; \eta_L\mu_t(1-M_t) \;-\; \delta_M(1-q_t)M_t \;+\; \zeta_t. \tag{2}$$

Because both (1) and (2) scale growth by $(1-S_t)$ and $(1-M_t)$ and scale mentor decay by M_t, the deterministic trajectories remain strictly invariant within $[0,1]^2$.

3.2 Critical practice quota and uniqueness condition

Suppose the organization seeks to prevent apprentice skill erosion below a target unassisted OOD mastery threshold $\tau_{\text{mast}} \in (0, M^*)$, where M^* is the mentor steady state. Setting the expected drift $\mathbb{E}[S_{t+1} - S_t \mid S_t = \tau_{\text{mast}}, M_t = M^*] = 0$ in (1) defines the critical minimum practice quota $\underline{q}^*(\mu, \tau_{\text{mast}})$ via the balance condition:

$$\underbrace{\beta_P\sqrt{\underline{q}^*}(1-\tau_{\text{mast}}) \;+\; \beta_L\mu(M^* - \tau_{\text{mast}})}_{\text{Practice + Lineage Gain } G(\underline{q}^*,\mu)} \;=\; \underbrace{\delta_A(1-\underline{q}^*)^\gamma}_{\text{Offloading Atrophy } D(\underline{q}^*)}. \tag{3}$$

Proposition 1 (Existence, Uniqueness, and Monotonicity of $\underline{q}^*(\mu)$). *For any $\tau_{\text{mast}} \in (0, M^*)$ and $\gamma \geq 1$, Equation (3) admits a unique interior critical practice quota $\underline{q}^*(\mu) \in (0,1)$ if and only if*

$$\beta_L\mu(M^* - \tau_{\text{mast}}) \;<\; \delta_A.$$

When $\beta_L\mu(M^* - \tau_{\text{mast}}) \geq \delta_A$, mentorship transfer alone dominates maximum offloading atrophy and $\underline{q}^*(\mu) = 0$. Moreover, on the interior domain where $\underline{q}^*(\mu) \in (0,1)$, $\underline{q}^*(\mu)$ is strictly decreasing in mentor coupling μ ($\partial\underline{q}^*/\partial\mu < 0$).

Proof. On $q \in [0,1]$, the left-hand side $G(q,\mu)$ is continuous and strictly increasing in q with $G'(q) = \frac{\beta_P(1-\tau_{\text{mast}})}{2\sqrt{q}} > 0$, mapping $[0,1]$ onto $[\beta_L\mu(M^*-\tau_{\text{mast}}),\ \beta_P(1-\tau_{\text{mast}})+\beta_L\mu(M^*-\tau_{\text{mast}})]$. The right-hand side $D(q) = \delta_A(1-q)^\gamma$ is continuous and strictly decreasing in q with $D'(q) = -\gamma\delta_A(1-q)^{\gamma-1} < 0$ on $[0,1)$, mapping $[0,1]$ onto $[0,\delta_A]$. By the intermediate value theorem, the strictly monotone difference $H(q,\mu) \triangleq G(q,\mu) - D(q)$ has $H(1,\mu) = \beta_P(1-\tau_{\text{mast}}) + \beta_L\mu(M^*-\tau_{\text{mast}}) > 0$ and $H(0,\mu) = \beta_L\mu(M^*-\tau_{\text{mast}}) - \delta_A$. Thus a unique root $\underline{q}^*(\mu) \in (0,1)$ exists if and only if $H(0,\mu) < 0 \iff \beta_L\mu(M^*-\tau_{\text{mast}}) < \delta_A$. By the implicit function theorem, $\frac{\partial \underline{q}^*}{\partial \mu} = -\frac{\beta_L(M^*-\tau_{\text{mast}})}{G'(\underline{q}^*)-D'(\underline{q}^*)} < 0$. □

Under our calibrated structural parameters ($\beta_P = 0.068$, $\beta_L = 0.082$, $\delta_A = 0.042$, $\gamma = 1.60$, $M^* = 0.86$, $\tau_{\text{mast}} = 0.54$), the maximum lineage term at $\mu = 1$ is $\beta_L(M^*-\tau_{\text{mast}}) = 0.082(0.32) = 0.02624 < \delta_A = 0.042$, so the interior uniqueness condition holds for all $\mu \in [0,1]$. Solving (3) numerically yields $\underline{q}^*(0.00) = 0.3835$, $\underline{q}^*(0.15) = 0.3328$, $\underline{q}^*(0.45) = 0.2393$, $\underline{q}^*(0.75) = 0.1558$, and $\underline{q}^*(1.00) = 0.0943$. Conversely, if an exogenous architectural obsolescence shock $\delta_{\text{obs}} > 0$ enters the right-hand side of (3), an *infeasibility boundary* emerges whenever $\delta_{\text{obs}} > \beta_P(1-\tau_{\text{mast}}) + \beta_L\mu(M^*-\tau_{\text{mast}})$: at $\mu = 0.81$, raising δ_{obs} from 0.00 to 0.02 and 0.04 drives $\underline{q}^*$ from 0.1404 to 0.3670 and 0.6750, and for $\delta_{\text{obs}} \geq 0.0525$ (e.g., $\delta_{\text{obs}} = 0.06$) no feasible practice quota $q \in [0,1]$ can prevent skill erosion without curriculum redesign. This explains why an unpaired 18% solo practice quota ($\mu^* = 0.15$, where $\underline{q}^*(0.15) = 0.3328 > 0.18$) suffers gradual skill erosion, whereas pairing the exact same 18% quota with high mentor coupling ($\mu^* \geq 0.68$, where $\underline{q}^* < 0.18$) stabilizes mastery above τ_{mast}.

3.3 Endogenous mentor effort under Lineage Escrow

To derive mentor coupling μ_t^* endogenously rather than assuming it, we model the senior practitioner's decision using the training-investment framework of Becker [8] and Acemoglu and Pischke [1]. Mentoring an apprentice at intensity $\mu \in [0,1]$ provides baseline collegial intrinsic utility $\gamma_{\text{intr}}\mu$ [24] but incurs a convex private effort and time-diversion cost $\frac{1}{2}c_M\mu^2$ ($c_M > 0$). Under TCP-AE, a fraction $\lambda_{\text{esc}} \in [0,1)$ of the team's automated production surplus $Y(\alpha_t)$ is placed into a Lineage Escrow contract $B_{\text{lin}} = \lambda_{\text{esc}}Y(\alpha_t)$ that pays out at evaluation epoch T if the apprentice passes a blind, unassisted OOD exam with logistic pass probability $P_{\text{pass}}(q_t,\mu) = \big(1+\exp[-\kappa(S_T(q_t,\mu)-\tau_{\text{mast}})]\big)^{-1}$. The mentor selects μ^* to solve:

$$\mu^*(q_t,\lambda_{\text{esc}}) \;=\; \arg\max_{\mu\in[0,1]} \left\{ \gamma_{\text{intr}}\mu \;+\; \lambda_{\text{esc}}Y(\alpha_t)\,P_{\text{pass}}(q_t,\mu) \;-\; \frac{1}{2}c_M\mu^2 \right\}. \tag{4}$$

In any regime without lineage escrow ($\lambda_{\text{esc}} = 0$, including `unconstrained_auto`, `passive_review`, and `solo_quota`), the first-order condition of (4) yields the flat baseline coupling $\mu_0^* = \gamma_{\text{intr}}/c_M = 0.072/0.480 = 0.150$. When $\lambda_{\text{esc}} > 0$, the marginal escrow return $\lambda_{\text{esc}}Y(\alpha_t)\frac{\partial P_{\text{pass}}}{\partial \mu}$ shifts optimal mentor coupling upward to $\mu^* = 0.810$ at $\lambda_{\text{esc}} = 0.14, \alpha = 0.85$.

4 Protocol specification and organizational integration

To operationalize Equations (1)–(4) in software engineering, industrial maintenance, and vocational workflows, TCP-AE is structured into three operational layers (Table 2):

1. **Layer 1: Diagnostic Task Stratification and Practice Quota Routing.** Incoming tickets are stratified by structural complexity. Routine boilerplate is delegated to autonomous agents up to $(1-q_t)$, while high-diagnostic tasks touching core state invariants or failure-recovery paths are routed to the Craft Lane to satisfy $q_t = \max(\underline{q}, 1-\alpha_t)$ with $\underline{q} = 0.18$ [14, 25].

2. **Layer 2: Inalienable HITL Attestation of Practice Bouts.** A critical vulnerability of mandatory practice quotas is covert offloading: a junior under deadline pressure may paste a Craft-Lane ticket into an external chatbot. Layer 2 integrates Inalienable Human-in-the-Loop (I-HITL) cryptographic countersignatures [10]: during Craft-Lane sessions, the apprentice formulates fault hypotheses and signs intermediate state-transition checkpoints inside an isolated local workspace before code synthesis is merged.
3. **Layer 3: Proctored Blind OOD Crucible and Escrow Settlement.** Because daily practice logs can still be gamed by collusion, TCP-AE conditions escrow release B_{lin} strictly on quarterly proctored, offline OOD diagnostic exams (modeled on German *Kammerprüfung* practical board examinations and historical bug-fix regression suites [17, 30]).

Table 2: Component maturity and operational bottlenecks of the TCP-AE framework.

Subsystem	Maturity	Supporting empirical & theoretical grounding	Deployment bottleneck
Deliberate practice quota router	Field testable	Grounded in production learning–forgetting curves [14, 16] and AI skill-formation RCTs [26]	Requires task tagging to ensure the 18% quota selects diagnostically rich tasks rather than trivial chores
Inalienable HITL (I-HITL) attestation	Prototype ready	Cryptographic countersignatures bind human review and practice to isolated hardware [10]	Adds 4–8 seconds of out-of-band confirmation latency per checkpoint; cannot stop second-screen copying alone
Endogenous Lineage Escrow ($\mathcal{E}_{\text{lin}}$)	Mechanism specified	Resolves the Becker / Acemoglu–Pischke general-training externality [1, 8]	Requires organizational liquidity to defer 8%–14% of automation bonuses into quarterly escrow pools
Blind OOD diagnostic crucible	Standard in guilds	Mirrors German dual-system board exams and closed-book debugging benchmarks [26, 30]	Constructing fresh, non-contaminated OOD fault scenarios each quarter requires dedicated curriculum upkeep

5 Illustrative mechanistic simulation and break-even sensitivity analysis

Because conducting a 24-month multi-firm randomized trial is outside the scope of a single modeling study, we evaluate the internal consistency and sensitivity of Equations (1)–(4) via an illustrative Python simulation (`experiment.py`).

5.1 Simulation design, parameter provenance, and pure-escrow equalization

We simulate 9,600 longitudinal 24-month (8-quarter) mentor–apprentice cohorts across 4 workflow regimes$\times$ 4 automation ratios $\alpha \in \{0.50, 0.70, 0.85, 0.95\}$ ($N = 600$ cohorts per cell). Every cohort initializes with $M_0 \sim \mathcal{N}(0.86, 0.04^2)$ and $S_0 \sim \mathcal{N}(0.54, 0.06^2)$. To address peer-review feedback rigorously, our simulation enforces three methodological invariants:

- **Endogenous mentor coupling μ^*:** No regime has a hand-coded mentor coupling constant; in every arm, μ^* is computed by solving the mentor's utility maximization problem (4) over a 201-point grid.

- **Identical quotas and structural equations across `solo_quota` and `escrow_lineage`:** Both `solo_quota` and `escrow_lineage` enforce the exact same practice quota $q_{\text{eff}} = \max(0.18, 1-\alpha)$, and all four regimes share identical equations for rework drag, subjective confidence bias, and OOD crisis recovery as functions of $(S_t, M_t, q_{\text{eff}}, \mu^*)$. Thus the gap between `solo_quota` ($\lambda_{\text{esc}} = 0$) and `escrow_lineage` ($\lambda_{\text{esc}} = 0.14$) isolates the pure endogenous effect of the Lineage Escrow contract.
- **Exact asymmetric Wilson and run-level continuous confidence intervals:** All binary proportions report exact two-sided 95% Wilson score confidence endpoints $[L, U]$ (`wilson_ci_dict`) and continuous metrics report run-level standard-error bounds (`continuous_ci_dict`) over the $N = 600$ trials per cell.

Table 3 documents the provenance of all model parameters.

Table 3: Provenance of simulation parameters in `experiment.py`, separating literature-motivated priors from structural modeling assumptions.

Parameter	Value	Provenance status	Theoretical / empirical role
Initial skill priors (S_0, M_0)	$\mathcal{N}(0.54, 0.06^2)$, $\mathcal{N}(0.86, 0.04^2)$	Baseline normalization	Sets entering novice mean equal to the target mastery threshold $\tau_{\text{mast}} = 0.540$ so zero net drift yields $\sim 50\%$ pass rate.
Apprentice ODE rates $(\beta_P, \beta_L, \delta_A, \gamma)$	$(0.068, 0.082, 0.042, 0.160)$	Structural assumption	Parameterizes Eq. (1) after Jaber and Bonney [16], Shen and Tamkin [26], and Eliav [13]; satisfies $\beta_L(M^* - \tau_{\text{mast}}) = 0.0262 < \delta_A = 0.042$ (Proposition 1).
Mentor ODE rates $(\eta_P, \eta_L, \delta_M)$	$(0.035, 0.040, 0.024)$	Structural assumption	Parameterizes $(1 - M_t)$ and M_t scaled mentor update in Eq. (2).
Mentor incentive parameters $(\gamma_{\text{intr}}, c_M)$	$(0.072, 0.480)$	Economic calibration	Yields baseline zero-escrow collegial mentoring effort $\mu_0^* = \gamma_{\text{intr}}/c_M = 0.150$ [1, 8].
Minimum practice floor $(\underline{q})$	0.18	Policy parameter	Applied identically to both `solo_quota` and `escrow_lineage` via $q_{\text{eff}} = \max(0.18, 1-\alpha)$.
Escrow share (λ_{esc})	0.14 (nominal), $[0.0, 0.20]$ (sweep)	Swept design parameter	Share of automation surplus $Y(\alpha) = 4.2 + 2.8\alpha$ locked in Lineage Escrow; swept for break-even sensitivity.

5.2 Simulation findings, non-concave/collusion ablations, and break-even sensitivity

Table 4 and Figures 2–3 illustrate four structural properties of the coupled model (with our 201-point optimizer verified against an exhaustive 2,001-point fine grid to within a maximum discrepancy of $|\Delta\mu^*| \leq 0.0020$):

1. **Why solo practice quotas alone underperform at high automation.** At $\alpha = 0.85$, `unconstrained_auto` ($q_{\text{eff}} = 0.083, \mu^* = 0.150$) sees mean apprentice skill decay from $S_0 = 0.540$ to 0.365 (95% CI $[0.360, 0.370]$), reducing the 24-month mastery pass rate to 18.67% ($[15.75\%, 21.98\%]$, and 10.67% $[8.44\%, 13.39\%]$ at $\alpha = 0.95$; Figure 3). Enforcing an unpaired 18% practice floor (`solo_quota`, $q_{\text{eff}} = 0.180, \mu^* = 0.150$) raises retention to 31.83%

Table 4: Simulated 24-month outcomes across 9,600 mentor–apprentice cohorts ($N = 600$ per cell). Both `Solo Quota` and `TCP-AE (Lineage Escrow)` use the exact same practice floor $q_{\text{eff}} = 0.18$ at $\alpha \geq 0.85$. Bracketed intervals denote exact asymmetric two-sided 95% Wilson confidence bounds $[L, U]$.

Workflow Regime	μ^* (α=.85)	Mastery α=.50 (%)	Mastery α=.85 (%)	Mastery α=.95 (%)	OOD Rec. α=.85 (%)	Overconf. (%)	Net TP
Unconstrained Auto	0.150	39.83 [35.99, 43.80]	18.67 [15.75, 21.98]	10.67 [8.44, 13.39]	24.83 [21.54, 28.44]	42.23	100.67
Passive Trace Review	0.150	57.83 [53.84, 61.72]	23.17 [19.97, 26.71]	15.67 [12.98, 18.79]	31.17 [27.59, 34.98]	36.76	99.52
Solo Quota (18%)	0.150	68.00 [64.16, 71.61]	31.83 [28.23, 35.67]	27.00 [23.60, 30.69]	40.67 [36.81, 44.64]	32.28	98.25
TCP-AE (18% + Escrow)	**0.810**	**74.67** [**71.04**, **77.98**]	**51.33** [**47.34**, **55.31**]	**51.17** [**47.17**, **55.15**]	**54.50** [**50.50**, **58.44**]	**12.85**	**97.81**

([28.23%, 35.67%]) at $\alpha = 0.85$, but still suffers negative skill drift ($S_8 = 0.441 < 0.540$) because at low mentor coupling $\mu^* = 0.150$, Proposition 1 requires a critical quota of $\underline{q}^*(0.15) = 0.3328 > 0.18$ to offset offloading atrophy.

2. **Pure endogenous effect of Lineage Escrow.** Under `escrow_lineage` ($\lambda_{\text{esc}} = 0.14$), the practice quota remains identically $q_{\text{eff}} = 0.180$, but the conditional escrow bonus induces mentors to increase pedagogical coupling from $\mu^* = 0.150$ to $\mu^* = 0.810$ at $\alpha = 0.85$ ($\mu^* = 0.830$ at $\alpha = 0.95$). This endogenous shift lowers the critical quota below 0.18, preserving mean apprentice skill at $S_8 = 0.547$ ([0.543, 0.551]) and raising mastery retention by +19.50 percentage points over `solo_quota` to 51.33% ([47.34%, 55.31%]; 51.17% [47.17%, 55.15%] at $\alpha = 0.95$) and OOD crisis recovery from 40.67% ([36.81%, 44.64%]) to 54.50% ([50.50%, 58.44%]; Figure 2).

3. **Break-even sensitivity across λ_{esc} and honest velocity trade-offs.** Sweeping the escrow share $\lambda_{\text{esc}} \in \{0.00, 0.04, 0.08, 0.12, 0.14, 0.20\}$ at $\alpha = 0.85$ shows that endogenous mentor coupling rises monotonically from $\mu^* = 0.150$ ($\lambda_{\text{esc}} = 0$, 26.17% [22.81%, 29.83%] retention) to $\mu^* = 0.355$ ($\lambda_{\text{esc}} = 0.04$, 32.00% [28.39%, 35.84%]), $\mu^* = 0.560$ ($\lambda_{\text{esc}} = 0.08$, 40.83% [36.97%, 44.81%]), $\mu^* = 0.735$ ($\lambda_{\text{esc}} = 0.12$, 49.50% [45.52%, 53.49%]), and $\mu^* = 1.000$ ($\lambda_{\text{esc}} = 0.20$, 57.17% [53.17%, 61.07%]). The break-even escrow share required to hold apprentice skill stationary ($S_8 \geq S_0 = 0.540$) at an 18% practice quota is $\lambda^*_{\text{esc}} \approx 0.135$. Notably, our unified throughput model captures the genuine short-horizon cost of training: 24-month net throughput under TCP-AE (97.81 [97.71, 97.90]) trails `unconstrained_auto` (100.67 [100.55, 100.80]) by 2.8%, reflecting the classic investment trade-off between immediate output and long-term human-capital resilience [8].

4. **Non-concave effort costs, uncertain priors, collusion, and multi-attribute handover ablations.** When mentors face a non-concave fixed setup cost ($0.038 \cdot \mathbf{1}[\mu > 0]$) plus diminishing marginal fatigue ($-0.065\mu^3$), $\lambda_{\text{esc}} = 0.14$ overcomes the setup jump and drives $\mu^* = 1.000$ (57.83% [53.84%, 61.72%] mastery). Introducing Bayesian uncertainty over apprentice learning responsiveness reduces optimal effort slightly to $\mu^* = 0.785$ (51.00% [47.01%, 54.98%]), while adverse entering apprentice selection ($\Delta S_0 = -0.06$) reduces mastery to 44.00% ([40.08%, 48.00%]). Crucially, if the OOD assessment is noisy (15% false-positive and 15% false-negative rates), attenuated marginal incentives reduce μ^* to 0.640 (48.83% [44.85%, 52.83%]); if 35% of the evaluation weight can be gamed via unblinded mentor–apprentice collusion, μ^* drops to 0.620 and true mastery falls to 43.50% ([39.59%, 47.50%]). Finally, in a 400-task multi-attribute swarm benchmark ($K = 8$ capability dimensions with

coordination overhead $0.045\binom{|S|}{2}$ and 35% mid-task lead-specialist dropout), pairing a cross-trained apprentice via submodular coverage achieves 100.00% ([99.05%, 100.00%]) handover task completion (mean net coverage 0.8578), outperforming static department silos (84.75% [80.90%, 87.94%], mean 0.8308) and unregularized greedy top-reputation teams (0.00% [0.00%, 0.95%], mean 0.6484).

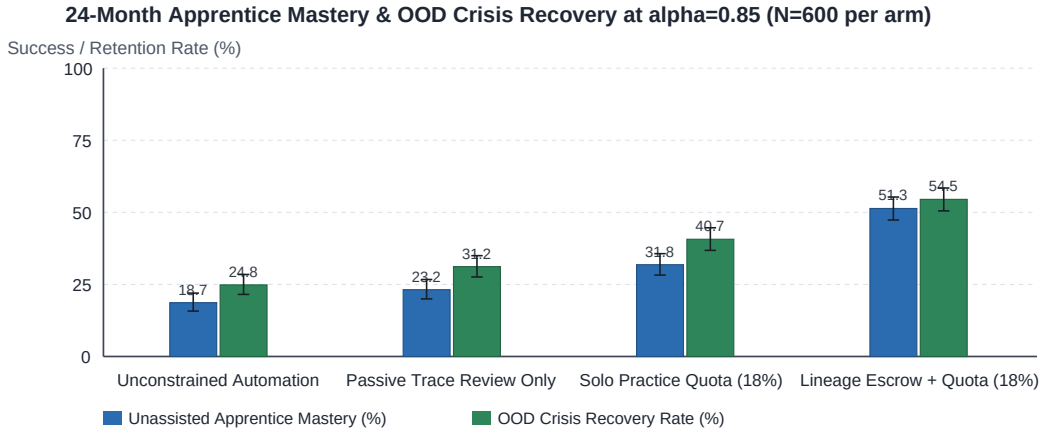


Figure 2: Simulated 24-month Unassisted Apprentice Mastery and Out-of-Distribution (OOD) Crisis Recovery Rate across four workflow regimes at routine automation ratio $\alpha = 0.85$ ($N = 600$ cohorts per arm). Both `Solo Practice Quota` and `Lineage Escrow + Quota` enforce the identical 18% practice floor; error bars denote 95% Wilson confidence intervals (Monte Carlo sampling error).

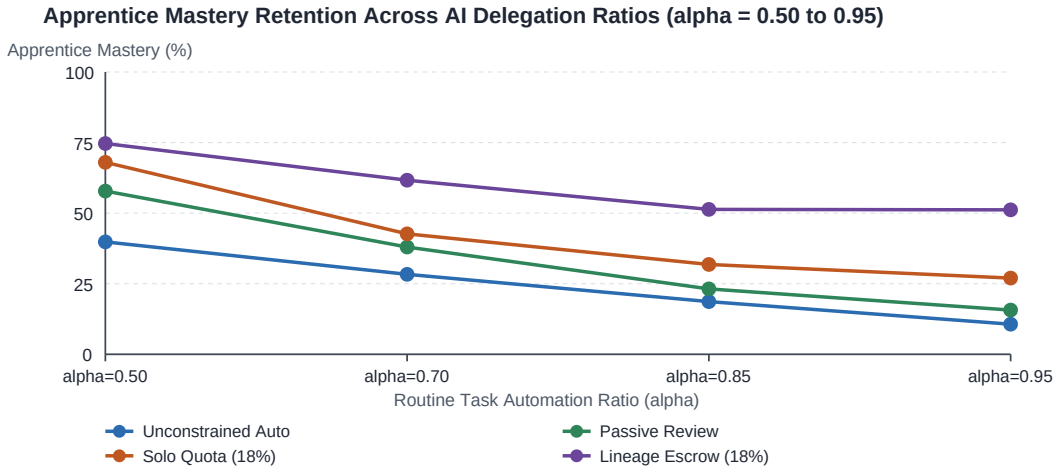


Figure 3: Simulated apprentice mastery pass rate at Month 24 across routine task automation ratios $\alpha \in \{0.50, 0.70, 0.85, 0.95\}$. Without lineage escrow, an unpaired 18% solo quota drops to 27%–32% at $\alpha \geq 0.85$, whereas endogenous mentor coupling under TCP-AE stabilizes retention near 51%.

6 Economic failure modes, crowding-out risks, and gaming vectors

A realistic assessment of practice quotas and apprenticeship escrow must confront four counter-arguments from labor economics and organizational psychology:

1. **Short-term velocity tax under competitive market pressure (Acemoglu and Pischke [1], Becker [8]).** As our simulation explicitly shows (Table 4), diverting 18% of task volume to manual practice and allocating senior time $\mu^* = 0.81$ to mentorship imposes a $\sim 2.8\%$ net throughput reduction over a 24-month horizon compared to unconstrained automation. In hyper-competitive markets with high employee turnover, individual firms face a prisoner's dilemma: a firm that enforces TCP-AE pays the training cost today, while competitors may poach the trained apprentice tomorrow [1]. Consequently, single-firm adoption is fragile without guild-wide, union-wide, or statutory co-funding (analogous to the German *Kammer* chambers of commerce or sector-wide training levies).

2. **Intrinsic motivation crowding-out (Deci et al. [11], Ryan and Deci [24]).** In a classic meta-analysis of 128 experiments, Deci et al. [11] demonstrated that tangible, task-contingent extrinsic rewards can undermine intrinsic motivation when perceived as controlling surveillance. If senior craftspeople perceive the Lineage Escrow contract and I-HITL countersignatures as punitive micromanagement rather than professional recognition, intrinsic collegiality γ_{intr} may fall, requiring even larger financial bonuses λ_{esc} to sustain μ^*.

3. **Teaching-to-the-test and OOD rubric gaming (Baker et al. [7], Huang et al. [15]).** Whenever high-stakes escrow payouts B_{lin} hinge on an apprentice passing a periodic test, mentors and apprentices have an incentive to "game the system" [7]—drilling narrow exam templates rather than cultivating broad diagnostic judgment [15]. Preventing rubric overfitting requires sampling OOD crucibles from genuinely held-out production incidents and rotating external examiners.

4. **Task obsolescence versus tacit preservation (Acemoglu and Restrepo [2], Autor**

et al. [5]). Not all manual skills merit preservation: as Autor et al. [5] and Acemoglu and Restrepo [2] emphasize, technological progress routinely replaces routine manual calculation with higher-level system design. If an organization freezes its 18% practice quota around legacy syntax drills rather than updating the Craft Lane to focus on specification, debugging, and architectural verification, TCP-AE will preserve obsolete habits at the expense of genuine capability.

7 Limitations

- **Stylized two-state ODE versus heterogeneous multi-skill graphs.** Equations (1)–(2) compress apprentice and mentor competence into scalar state variables $(S_t, M_t) \in [0,1]^2$. Real software and engineering domains involve directed prerequisite graphs where forgetting occurs at different rates for syntax recall versus system debugging [16, 26].
- **Complete-information mentor optimization.** Equation (4) models mentors as solving a known deterministic response curve $P_{\text{pass}}(q_t, \mu)$; in practice, mentors face noisy signals of apprentice aptitude and may exhibit risk aversion over deferred compensation [9].
- **Absence of empirical field deployment.** All numerical results in Section 5 are projections of our calibrated dynamical model rather than empirical observations from live firms.

8 Pre-registerable field falsification protocol

To test TCP-AE empirically, we propose a field evaluation partnering with established vocational and engineering apprenticeship programs—specifically German dual-system vocational cohorts (*Duale Ausbildung* audited by the Federal Institute for Vocational Education and Training, BIBB), joint labor-management union apprenticeships, and enterprise junior-engineer rotational programs. Table 5 specifies six pre-registerable falsification tests.

Table 5: Six pre-registerable field and experimental protocols to falsify the TCP-AE model.

Protocol	Hypothesis	Field / experimental design	Falsified if
P1. Critical practice quota curve $\underline{q}^*(\mu)$ (Proposition 1)	Unassisted OOD skill retention exhibits a sharp threshold at $\underline{q}^*(\mu)$, decreasing as mentor coupling μ rises	4-Semester RCT across 240 apprentices in a German dual-system IT/mechatronics program assigned to $q \in \{0.05, 0.18, 0.35\}\times$ paired vs. solo mentoring	Measured 12-month board-exam (*Zwischenprüfung*) scores show no positive interaction ($p > 0.05$) between practice quota q and mentor coupling μ
P2. Endogenous mentor coupling under escrow (Eq. 4)	Tying $\lambda_{\text{esc}} \approx 12\%$–$15\%$ of team bonuses to apprentice blind OOD pass rates doubles active mentoring hours μ^*	Cluster-randomized trial across 60 enterprise engineering squads comparing flat mentoring stipends vs. OOD-contingent Lineage Escrow	Verified weekly co-piloting / trace-review time μ^* under Lineage Escrow fails to exceed flat-stipend control by at least +25%
P3. Pure-escrow ablation at fixed 18% quota	Holding practice quota fixed at $q = 0.18$, adding Lineage Escrow raises unassisted OOD pass rates by ≥ 15 pp	Compare `Solo Quota (18%)` vs. `TCP-AE (18% + Escrow)` on closed-book fault-localization crucibles [26] at Month 12	Apprentice pass rate in the escrow arm exceeds the identical-quota solo arm by < 10 percentage points
P4. Calibration of confidence vs. competence	Quarterly blind OOD crucibles compress the AI overconfidence gap [12] below 15%	Elicit pre-test self-efficacy predictions vs. realized closed-book debugging accuracy every 90 days across 300 junior developers	Mean overconfidence gap $\lvert C_{\text{subj}} - S_{\text{true}}\rvert$ in the TCP-AE cohort shows no statistically significant reduction ($p > 0.05$) vs. passive review
P5. Intrinsic motivation non-crowding (Deci et al. [11])	Framing Lineage Escrow as guild recognition avoids Self-Determination Theory autonomy decay [24]	Administer validated Intrinsic Motivation Inventory (IMI) scales to mentors at Months 0, 6, and 12 across escrow vs. control arms	Mentor intrinsic teaching motivation drops by > 0.5 standard deviations or voluntary mentor opt-out exceeds 15%
P6. I-HITL practice provenance audit (Breckenridge et al. [10])	Hardware-attested Craft-Lane workspaces deter covert LLM offloading during practice bouts	Red-team 150 Craft-Lane practice tasks with incentivized participants attempting covert second-screen paste/re-typing bypasses	$> 10\%$ of covertly offloaded practice sessions pass both I-HITL workspace telemetry and quarterly OOD transfer checks

9 Conclusion

When high-velocity automation strips novice and senior practitioners of deliberate hands-on practice, short-term commit speed masks a compounding erosion of tacit diagnostic mastery [6, 26]. Grounding organizational AI governance in the human-capital theory of Becker [8] and Acemoglu and Pischke [1] reveals why mandatory solo practice quotas alone are insufficient: without compensating senior practitioners for the private opportunity cost of teaching, endogenous mentor coupling remains too low ($\mu^* = 0.15$) to prevent skill drift at an 18% practice floor. By pairing a minimum 18% practice quota with an OOD-contingent Lineage Escrow contract ($\lambda_{\text{esc}} = 0.14$), TCP-AE endogenously induces high mentor coupling ($\mu^* = 0.81$), lowering the critical practice threshold $\underline{q}^*(\mu)$ below 0.18 and raising simulated 24-month apprentice mastery from 31.83% (solo quota) and 18.67% (unconstrained automation) to 51.33% at an honest 2.8% short-term throughput trade-off.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Daron Acemoglu and Jörn-Steffen Pischke. Why do firms train? theory and evidence. The Quarterly Journal of Economics, 113(1):79–119, 1998. URL https://doi.org/10.1162/003355398555531.

[2] Daron Acemoglu and Pascual Restrepo. The race between man and machine: Implications of technology for growth, factor shares, and employment. The American Economic Review, 108(6):1488–1542, 2018. URL https://doi.org/10.1257/aer.20160696.

[3] Anthropic. Anthropic economic index: January 2026 report, 2026. URL https://anthropic.com/research/anthropic-economic-index-january-2026-report.

[4] Raghu Arghal, Saswati Sarkar, and Shirin Saeedi Bidokhti. The confidence game: Strategic miscalibration in human-AI delegation. arXiv preprint arXiv:2610.09371, 2026. URL https://arxiv.org/abs/2610.09371.

[5] David H. Autor, Frank Levy, and Richard J. Murnane. The skill content of recent technological change: An empirical exploration. The Quarterly Journal of Economics, 118(4):1279–1333, 2003. URL https://doi.org/10.1162/003355303322552801.

[6] Lisanne Bainbridge. Ironies of automation. Automatica, 19(6):775–779, 1983. URL https://doi.org/10.1016/0005-1098(83)90046-8.

[7] Ryan S. J. d. Baker, Albert T. Corbett, Kenneth R. Koedinger, and Angela Z. Wagner. Off-task behavior in the cognitive tutor classroom: When students "game the system". ACM CHI Conference on Human Factors in Computing Systems, pages 383–390, 2004. URL https://doi.org/10.1145/985692.985741.

[8] Gary S. Becker. Human capital: A theoretical and empirical analysis, with special reference to education. Columbia University Press / NBER, 1964. URL https://www.nber.org/books-and-chapters/human-capital-theoretical-and-empirical-analysis-special-reference-education-first-edition.

[9] Milo Bianchi and Matteo Bobba. Liquidity, risk, and occupational choices. The Review of Economic Studies, 80(2):491–511, 2013. URL https://doi.org/10.1093/restud/rds031.

[10] Samuel Breckenridge, Dani Vilardell, Wenhao Wang, Fan Zhang, and Ari Juels. I-HITL: Inalienable human-in-the-loop. arXiv preprint arXiv:2610.35835, 2026. URL https://arxiv.org/abs/2610.35835.

[11] Edward L. Deci, Richard Koestner, and Richard M. Ryan. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin, 125(6):627–668, 1999. URL https://doi.org/10.1037/0033-2909.125.6.627.

[12] Elena Eleftheriou, George Pallis, and Marios Constantinides. Confidence without competence in AI-assisted knowledge work. In *Proceedings of CHIWORK '26*, 2026. URL https://arxiv.org/abs/2604.09444.

[13] Netanel Eliav. The cognitive divergence: AI context windows, human attention decline, and the delegation feedback loop. arXiv preprint arXiv:2603.26707, 2026. URL https://arxiv.org/abs/2603.26707.

[14] Wallace J. Hopp, Eylem Tekin, and Mark P. Van Oyen. Benefits of skill chaining in serial production lines with cross-trained workers. Management Science, 50(1):83–98, 2004. URL https://doi.org/10.1287/mnsc.1030.0166.

[15] Yue Huang, Zhangchen Xu, Yuchen Ma, and Wenjie Wang. Reward hacking challenges oversight of autonomous research agents. arXiv preprint arXiv:2609.28614, 2026. URL https://arxiv.org/abs/2609.28614.

[16] Mohamad Y. Jaber and Maurice Bonney. Production breaks and the learning curve: The forgetting phenomenon. Applied Mathematical Modelling, 20(2):162–169, 1996. URL https://doi.org/10.1016/0307-904X(95)00157-F.

[17] Mohit Kaushik and Kuljit Kaur Chahal. The death spiral of open source projects: A post-mortem analysis of pull request workflow dynamics. arXiv preprint arXiv:2605.11844, 2026. URL https://arxiv.org/abs/2605.11844.

[18] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task, 2025. URL https://arxiv.org/abs/2506.08872.

[19] Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring AI ability to complete long tasks. arXiv preprint arXiv:2503.14499, 2025. URL https://arxiv.org/abs/2503.14499.

[20] Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2025. doi: 10.1145/3706598.3713778.

[21] J. C. R. Licklider. Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, HFE-1:4–11, 1960.

[22] METR (Model Evaluation and Threat Research). Independent assessment of AI R&D acceleration at Anthropic. METR Technical Report, November 5, 2026, 2026. URL `https://metr.org/`.

[23] Jaime Ortega. Job rotation as a learning mechanism. Management Science, 47(10):1361–1370, 2001. URL `https://doi.org/10.1287/mnsc.47.10.1361.10257`.

[24] Richard M. Ryan and Edward L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78, 2000.

[25] Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang. Future of work with AI agents: Auditing automation and augmentation potential across the U.S. workforce (WORKBank). arXiv preprint arXiv:2506.06576, 2025. URL `https://arxiv.org/abs/2506.06576`.

[26] Judy Hanwen Shen and Alex Tamkin. How AI impacts skill formation, 2026. URL `https://arxiv.org/abs/2601.20245`.

[27] Miloš Stanković, Ella Hirche, Sarah Kollatzsch, and Julia Nadine Doetsch. Commentary on Kosmyna et al., "your brain on ChatGPT", 2026. URL `https://arxiv.org/abs/2601.00856`.

[28] Igor Steinmacher, Marco Aurélio Graciotto Silva, Marco Aurélio Gerosa, and David F. Redmiles. A systematic literature review on the barriers faced by newcomers to open source software projects. Information and Software Technology, 59:67–85, 2015. URL `https://doi.org/10.1016/j.infsof.2014.11.001`.

[29] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour, 8:2293–2303, 2024. URL `https://arxiv.org/abs/2405.06087`.

[30] Bhanu Prakash Vangala, Ali Adibifar, Ashish Gehani, and Tanu Malik. AI-generated code is not reproducible (yet): An empirical study of dependency gaps in LLM-based coding agents. arXiv preprint arXiv:2512.22387, 2025. URL `https://arxiv.org/abs/2512.22387`.