

Deskilling Traps: A Dynamical Systems Model of Supervisory Skill Erosion Under AI Assistance

Anonymous Author(s)

ABSTRACT

As organizations increasingly deploy AI assistants across professional domains, a critical question emerges: does AI assistance erode the human skills required to supervise automated outputs? We formalize this question through a dynamical systems model that couples skill evolution, metacognitive calibration, and endogenous AI reliance. Through computational experiments across four professional domains (software engineering, medicine, finance, and aviation), we identify *deskilling traps*—parameter regimes where workers lose supervisory competence and simultaneously lose awareness of their incompetence, making self-correction impossible. Our simulations reveal three key findings: (1) novice workers in high-reliability AI domains (aviation, medicine) are most vulnerable, with all experience levels in aviation entering deskilling traps; (2) a *reliability paradox* exists wherein higher AI reliability increases deskilling risk by reducing the error signals necessary for skill maintenance, with a critical threshold at approximately 0.938 reliability; and (3) scaffolded autonomy—where AI progressively reduces its assistance as worker skill grows—is the most effective intervention, raising final skill from 0.048 to 0.983 while reducing cumulative harm by 87.6%. Sensitivity analysis confirms that skill decay rate is the most influential model parameter, producing a 23× range in outcomes, while metacognition remains stable across perturbations. Analysis of combined interventions reveals that scaffolded autonomy is necessary and nearly sufficient, with additional interventions providing marginal improvement (<0.2%). Recovery experiments demonstrate that deskilling traps are genuinely trapping: only scaffolded autonomy enables escape, requiring approximately 203 weeks. These results have direct policy implications for organizational AI deployment, training design, and regulatory oversight in safety-critical domains.

CCS CONCEPTS

- Human-centered computing → Human computer interaction (HCI);
- Computing methodologies → Modeling and simulation.

KEYWORDS

AI assistance, deskilling, human oversight, automation, skill decay, supervisory control

1 INTRODUCTION

The rapid adoption of AI assistants across professional domains has produced measurable productivity gains [5, 17]. Software engineers using code generation tools complete tasks faster [17], knowledge workers with AI support produce higher-quality outputs [8], and medical professionals using diagnostic AI achieve greater accuracy on routine cases [5]. Yet this performance improvement comes with an underexamined cost: the potential erosion of the human skills required to *supervise* the very systems providing the assistance.

Shen et al. [18] identify this tension as a central open problem: “Although more workers rely on AI to improve their productivity, it is unclear whether the use of AI assistance in the workplace might hinder core understanding of concepts or prevent the development of skills necessary to supervise automated tasks.” This problem is especially acute in safety-critical domains—aviation, medicine, nuclear operations—where human oversight of automated systems is not merely desirable but legally and ethically mandated.

The concern is not new. Bainbridge’s seminal “ironies of automation” [3] observed that automation eliminates the very tasks through which operators develop and maintain the skills needed to intervene when automation fails. Parasuraman and Riley [16] documented patterns of misuse, disuse, and abuse of automation arising from miscalibrated trust. Endsley [10] synthesized decades of human-automation interaction research, emphasizing that situation awareness degrades when humans become passive monitors rather than active controllers.

However, the current wave of generative AI introduces qualitatively new dynamics. Unlike traditional automation, which executes fixed procedures, modern AI systems produce novel outputs that require domain-specific expertise to evaluate. A code generation tool may produce syntactically valid but semantically incorrect code; a diagnostic AI may suggest a plausible but wrong diagnosis. Detecting such errors requires the very skills that AI assistance may erode—creating a potentially self-reinforcing *deskilling trap*.

We formalize this phenomenon through a dynamical systems model that captures five interacting processes: (1) skill growth through deliberate practice, (2) skill decay from disuse when tasks are offloaded to AI, (3) partial skill maintenance from reviewing AI outputs, (4) metacognitive calibration evolution, and (5) error detection as a function of the skill-difficulty gap. Our contributions are:

- A formal dynamical model of supervisory skill evolution under AI assistance that identifies deskilling trap conditions (Section 2).
- Computational experiments across four professional domains revealing domain-specific vulnerability patterns and a *reliability paradox* where higher AI reliability increases deskilling risk (Section 3).
- Systematic comparison of four mitigation interventions, demonstrating that scaffolded autonomy is the most effective, achieving near-complete skill preservation (Section 3).
- Evidence of a stark generational asymmetry between pre-AI and post-AI cohorts with implications for workforce training policy (Section 3).
- Sensitivity analysis showing that skill decay rate is the dominant parameter, combined intervention analysis demonstrating scaffolded autonomy is necessary and nearly sufficient, and recovery experiments revealing that deskilling

traps are genuinely trapping with only scaffolded autonomy enabling escape (Section 3).

1.1 Related Work

Skill acquisition and decay. The cognitive science of skill development, from Fitts and Posner's stage theory [12] through Anderson's ACT-R framework [1], establishes that skills are built through deliberate practice [11] and decay without use [2]. Our model builds on these foundations, using logistic skill growth and exponential decay as established functional forms.

AI and learning. Recent empirical work has begun to document AI's effects on learning. Bastani et al. [4] found that students using GPT-4 for practice performed worse on subsequent unassisted assessments, providing direct evidence that AI assistance can hinder skill formation. Doshi and Hauser [9] showed that while AI enhances individual creative output, it reduces collective diversity—suggesting that AI assistance may narrow the distribution of human capabilities. Noy and Zhang [15] provided experimental evidence that generative AI increases productivity for less-skilled workers but compresses the skill distribution, raising concerns about long-term skill development when performance gains mask learning deficits.

AI and professional decision-making. In medical settings, Chen and Asch [6] documented that AI assistance in clinical decision-making can lead to automation bias, where clinicians defer to AI recommendations even when their own judgment would be superior. This finding parallels our model's reliance dynamics, where perceived AI quality drives increasing delegation regardless of actual supervisory capability. The broader concern about cognitive risks from AI dependence is discussed in [20].

Automation and human factors. The human factors literature on automation provides the theoretical foundation for our work. Lee and See [14] established that trust in automation is a dynamic process that depends on reliability, predictability, and experience. Our model incorporates these insights through the endogenous reliance mechanism. Domain-specific effects in software development are examined in [7, 19]. Learning effects from AI tool experience, including difficulties in disentangling genuine skill development from tool-dependent performance, are explored in [21].

The Dunning-Kruger connection. Kruger and Dunning [13] showed that individuals with low competence in a domain tend to overestimate their ability, precisely because they lack the metacognitive skill to recognize their deficiency. Our model formalizes this insight: when both skill and metacognition fall below critical thresholds, the worker is trapped because they cannot recognize their inability to supervise.

2 METHODS

2.1 Model Overview

We model a worker whose supervisory skill $s(t) \in [0, 1]$ and metacognitive calibration $m(t) \in [0, 1]$ evolve over discrete time steps (each representing one week). The worker handles $N = 20$ tasks per time step, delegating a fraction $r(t) \in [0, 0.95]$ to an AI system. The model consists of three coupled dynamical equations

governing skill, metacognition, and reliance. Figure ?? provides a visual overview of the model architecture and the feedback loops between its components.

2.2 Skill Dynamics

The skill level evolves as:

$$\frac{ds}{dt} = \underbrace{\alpha \cdot (1-r) \cdot s(1-s)}_{\text{growth}} - \underbrace{\beta \cdot r \cdot s + \tau \cdot \alpha \cdot r \cdot s(1-s)/2}_{\text{decay}} + \underbrace{\tau \cdot \alpha \cdot r \cdot s(1-s)/2}_{\text{transfer}} \quad (1)$$

where α is the skill growth rate from unassisted practice, β is the decay rate from disuse, τ is the review transfer coefficient capturing partial learning from reviewing AI outputs, and r is the AI reliance fraction. The growth term uses a logistic form: skill grows fastest at intermediate levels and saturates near the extremes. The decay term is proportional to both current skill and reliance: more delegation causes faster decay. The transfer term captures that reviewing AI outputs provides some (reduced) learning signal.

2.3 Error Detection

The probability that a worker detects an AI error on a task of difficulty d is:

$$P(\text{detect} \mid s, m, d) = \frac{1}{1 + e^{-\kappa(s-d)}} \cdot (0.5 + 0.5m) \quad (2)$$

where $\kappa = 5 + 10m$ controls the sigmoid steepness. The first factor captures the domain skill requirement: detection is likely when skill exceeds task difficulty and unlikely otherwise. The second factor captures metacognitive vigilance: even with sufficient skill, a worker who rubber-stamps AI outputs (low m) will miss errors.

2.4 Metacognition Dynamics

Metacognitive calibration evolves as:

$$\frac{dm}{dt} = \underbrace{0.02 \cdot e_{\text{exp}} \cdot (1-m)}_{\text{calibration signal}} - \underbrace{0.01 \cdot r \cdot \rho_{\text{AI}} \cdot m}_{\text{complacency}} \quad (3)$$

where e_{exp} is the error exposure rate (fraction of AI-handled tasks containing errors that the worker encounters) and ρ_{AI} is the AI reliability. Metacognition grows when the worker encounters and processes errors, and decays through complacency when AI reliability is high and reliance is strong.

2.5 Endogenous Reliance

AI reliance adapts based on perceived AI quality:

$$\frac{dr}{dt} = 0.02 \cdot (\hat{q}_{\text{AI}} - 0.5) \quad (4)$$

where $\hat{q}_{\text{AI}} = 1 - e_{\text{detected}}/\max(Nr, 1)$ is the perceived quality based on detected errors. This creates a positive feedback loop: when few errors are detected (either because AI is reliable or because the worker cannot detect errors), reliance increases, further reducing practice opportunities.

2.6 Deskilling Trap Definition

We define a *deskilling trap* as a state where:

$$s(T) < 0.3 \quad \text{and} \quad m(T) < 0.3 \quad (5)$$

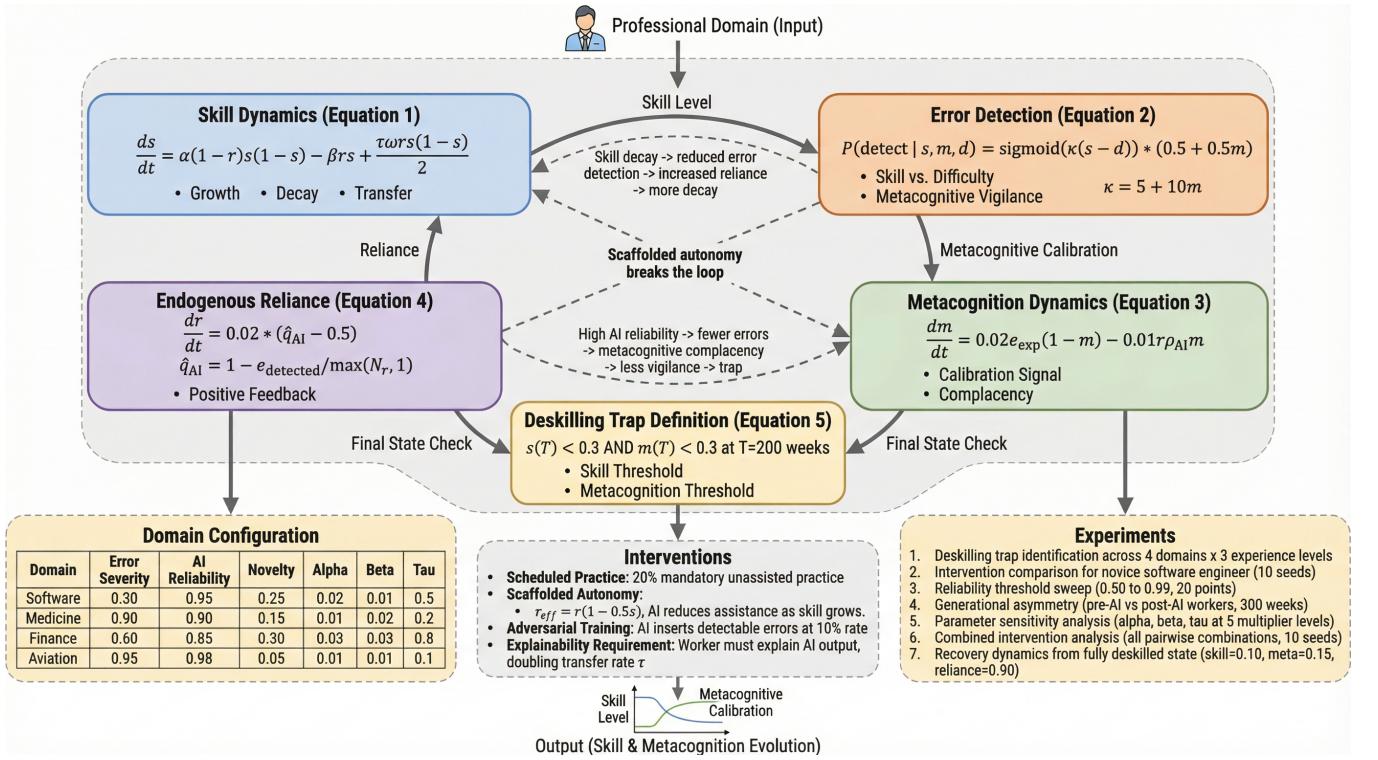


Figure 1: Dynamical systems model architecture. The model couples five interacting processes—skill growth, skill decay, transfer learning, error detection, and reliance adaptation—through feedback loops across four professional domains. Skill evolution (Eq. 1) depends on the balance between practice-driven growth and disuse decay, modulated by AI reliance. Error detection (Eq. 2) feeds into metacognitive calibration (Eq. 3), which in turn affects reliance dynamics (Eq. 4), creating the positive feedback loops that can produce deskilling traps.

at the end of the simulation ($T = 200$ weeks). This captures the condition where the worker both (a) lacks the skill to supervise AI outputs effectively and (b) lacks the metacognitive awareness to recognize their deficiency. Figure ?? visualizes the phase space structure, showing the trap region, domain-specific trajectories, and the effect of interventions on escape dynamics.

2.7 Domain Configuration

We instantiate the model across four professional domains with parameters calibrated from the human factors literature (Table 1). Each domain differs in error severity, AI reliability, task novelty rate, and skill dynamics parameters.

2.8 Interventions

We evaluate four candidate interventions:

- (1) **Scheduled Practice:** 20% of time is mandatory unassisted practice, regardless of AI reliance level.
- (2) **Scaffolded Autonomy:** AI reduces its assistance as worker skill grows: $r_{eff} = r \cdot (1 - 0.5s)$.
- (3) **Adversarial Training:** AI deliberately inserts detectable errors at a 10% rate to maintain vigilance.
- (4) **Explainability Requirement:** Worker must explain why AI output is correct, doubling the transfer learning rate τ .

Table 1: Domain configuration parameters. Error severity reflects the cost of undetected errors (0=benign, 1=catastrophic). AI reliability is the baseline probability of correct AI output. Task novelty rate is the fraction of tasks outside the AI training distribution.

Parameter	Software	Medicine	Finance	Aviation
Error severity	0.30	0.90	0.60	0.95
AI reliability	0.85	0.90	0.80	0.95
Novelty rate	0.25	0.15	0.30	0.05
Feedback delay	5.0	15.0	10.0	0.5
Growth rate α	0.05	0.03	0.04	0.04
Decay rate β	0.02	0.015	0.025	0.03
Transfer rate τ	0.30	0.20	0.25	0.15

2.9 Experimental Design

We conduct seven experiments:

- **Experiment 1:** Deskilling trap identification across all four domains with three experience levels (novice, intermediate, expert).
- **Experiment 2:** Intervention comparison for a novice software engineer across 10 random seeds.

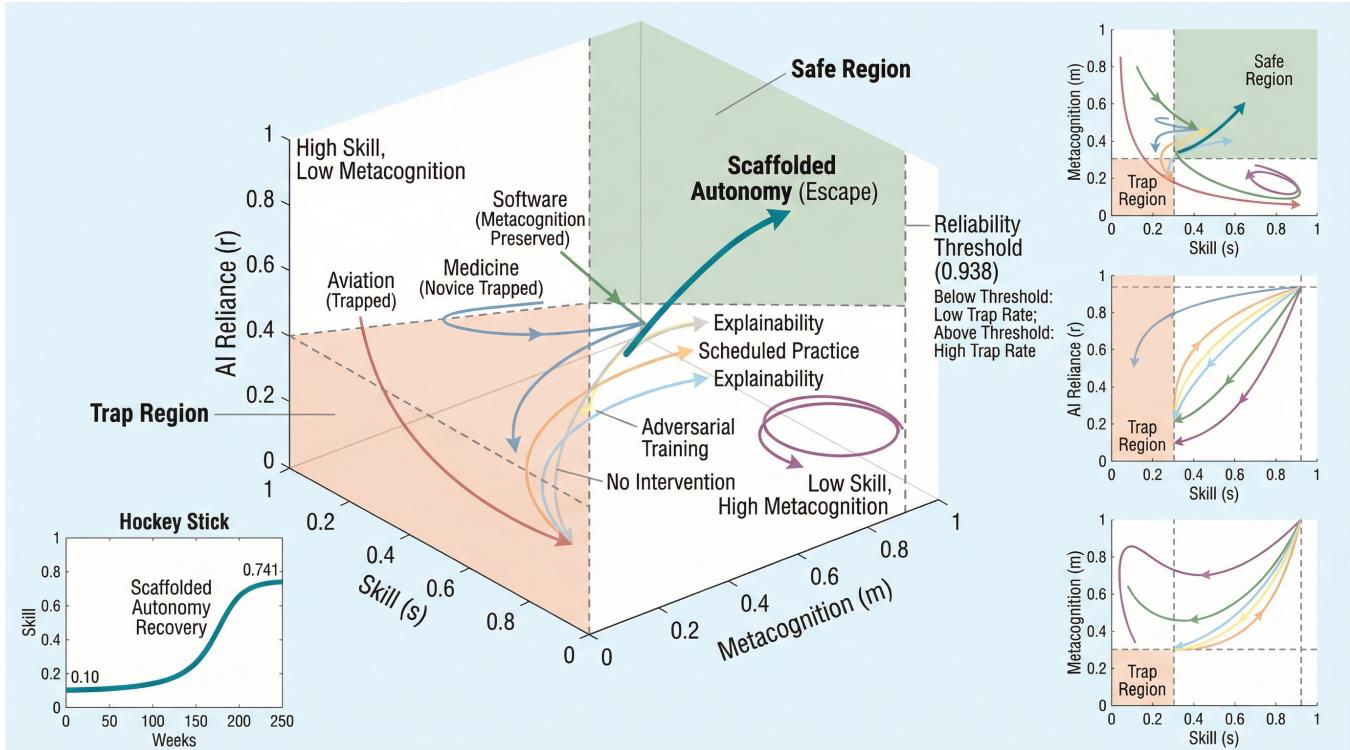


Figure 2: Deskilling trap phase space. The system evolves in the skill-metacognition plane, with the trap region defined by $s < 0.3$ and $m < 0.3$ (lower-left quadrant). Domain-specific trajectories show that aviation enters the trap across all experience levels, while finance maintains high metacognition despite severe skill decay. Intervention trajectories demonstrate that only scaffolded autonomy escapes the trap basin, exhibiting a characteristic “hockey stick” recovery pattern requiring approximately 203 weeks.

- **Experiment 3:** Reliability threshold sweep from 0.50 to 0.99 (20 points) to identify the critical reliability level above which deskilling traps emerge.
- **Experiment 4:** Generational asymmetry comparison between pre-AI workers (high initial skill) and post-AI workers (low initial skill, high initial reliance) over 300 weeks.
- **Experiment 5:** Parameter sensitivity analysis sweeping skill growth rate, skill decay rate, and review transfer rate at five multiplier levels ($0.5\times$ to $1.5\times$) to assess model robustness.
- **Experiment 6:** Combined intervention analysis testing all pairwise and individual interventions across 10 random seeds to identify synergies.
- **Experiment 7:** Recovery dynamics from a fully deskilled initial state (skill = 0.10, metacognition = 0.15, reliance = 0.90) under each intervention.

All simulations use $N = 20$ tasks per time step, with task difficulties drawn from a Beta(2,5) distribution. Results are reproducible via fixed random seeds.

3 RESULTS

3.1 Experiment 1: Deskilling Traps Across Domains

Table 2 summarizes the outcomes of 200-week simulations across four domains and three experience levels. The most striking finding is that **all experience levels in aviation enter deskilling traps**, including experts who begin with skill level 0.80. In aviation, the combination of very high AI reliability (0.95) and high skill decay rate (0.03) creates a regime where the error signal is too sparse to sustain skill, and the low review transfer rate (0.15) means that passive monitoring provides insufficient learning.

Medicine shows a mixed pattern: novice physicians enter the deskilling trap ($s = 0.047, m = 0.300$), but intermediate and expert physicians maintain metacognition above the threshold despite severe skill decay. Finance produces the lowest final skills across all levels but avoids traps because metacognition remains relatively high ($m > 0.45$), likely due to the higher task novelty rate (0.30) providing more error signals.

Figure 1 shows the skill trajectories across domains. In all cases, skill declines monotonically once AI reliance saturates, but the rate and asymptotic behavior differ substantially by domain.

Table 2: Experiment 1: Final outcomes after 200 weeks of AI-assisted work. Deskilling traps (skill < 0.3 and metacognition < 0.3) are marked with \dagger . All workers begin with AI reliance ≥ 0.50 and converge to maximum reliance (0.95) by simulation end.

Domain	Level	Final Skill	Final Meta.	Detect Rate	Total
Software	Novice	0.048	0.390	0.262	1
	Intermediate	0.075	0.404	0.408	1
	Expert	0.106	0.429	0.542	43.8
Medicine	Novice \dagger	0.047	0.300	0.255	124.8
	Intermediate	0.076	0.322	0.355	116.9
	Expert	0.106	0.344	0.472	89.4
Finance	Novice	0.012	0.458	0.195	173.5
	Intermediate	0.022	0.475	0.277	167.4
	Expert	0.040	0.490	0.444	127.2
Aviation	Novice \dagger	0.010	0.187	0.213	65.5
	Intermediate \dagger	0.010	0.218	0.159	71.3
	Expert \dagger	0.010	0.242	0.357	53.3

Supervisory Skill Trajectories Under AI Assistance

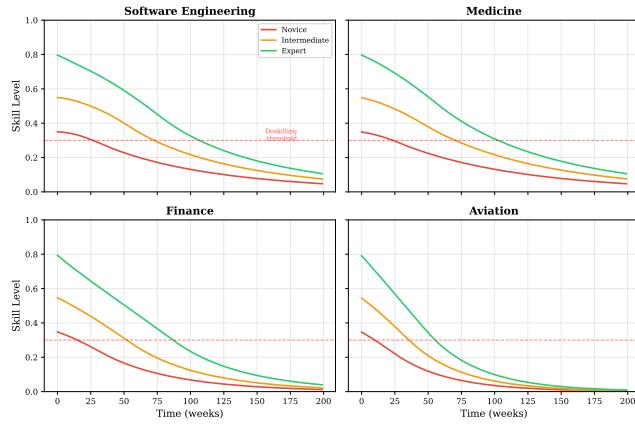


Figure 3: Supervisory skill trajectories over 200 weeks across four domains and three experience levels (novice, intermediate, expert). The dashed red line at $s = 0.3$ marks the supervisory competence threshold. All trajectories decline, but the rate and final level depend on domain characteristics. Aviation shows the most severe decline due to high AI reliability and low review transfer.

3.2 Experiment 2: Intervention Effectiveness

Table 3 presents the intervention comparison results for a novice software engineer, averaged over 10 random seeds. Scaffolded autonomy dramatically outperforms all other interventions, achieving a final skill of 0.983 ± 0.001 compared to 0.048 ± 0.000 under no intervention—a 20-fold improvement. The mechanism is clear: by reducing AI assistance as skill grows, scaffolded autonomy restores the practice signal that drives skill acquisition.

Scheduled practice and the explainability requirement produce modest improvements (skill approximately 0.125 vs. 0.048), while

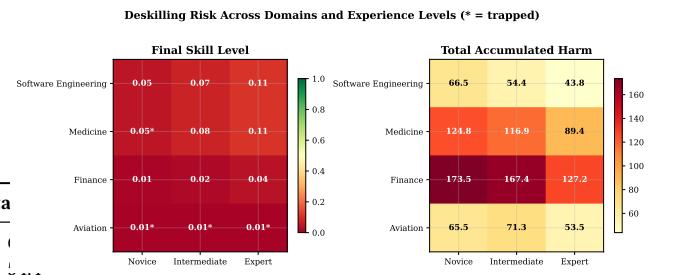


Figure 4: Deskilling risk heatmap across domains and experience levels. Left: final skill level (green = high, red = low). Right: total accumulated harm over 200 weeks. Asterisks (*) mark deskilling trap states. Aviation is uniquely vulnerable across all experience levels, while finance accumulates the most harm due to moderate AI reliability and high task novelty.

Table 3: Experiment 2: Intervention comparison for a novice software engineer (10 seeds). Scaffolded autonomy achieves dramatically higher skill and lower harm than all alternatives.

Intervention	Final Skill	Detection Rate	Total Harm
No Intervention	0.048 ± 0.000	0.234 ± 0.012	67.1 ± 1.7
Scheduled Practice	0.125 ± 0.000	0.295 ± 0.013	63.5 ± 1.7
Scaffolded Autonomy	0.983 ± 0.001	0.684 ± 0.034	8.3 ± 0.7
Adversarial Training	0.048 ± 0.000	0.234 ± 0.012	60.1 ± 1.5
Explainability Req.	0.126 ± 0.000	0.303 ± 0.017	62.9 ± 2.3

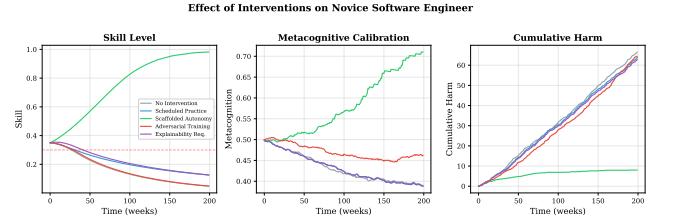


Figure 5: Intervention trajectories for a novice software engineer over 200 weeks. Left: skill level. Center: metacognitive calibration. Right: cumulative harm. Scaffolded autonomy (green) is the only intervention that reverses the deskilling trajectory, achieving near-expert skill levels. The remaining interventions slow but do not prevent skill decline.

adversarial training improves metacognition (0.448 vs. 0.388) and reduces harm but does not substantively improve skill level. The key insight is that adversarial error injection provides a calibration signal but does not restore the practice volume needed for skill growth.

Figure 3 shows the full trajectories, and Figure 4 presents the aggregated bar comparison with error bars. The scaffolded autonomy trajectory shows a distinctive pattern: initial skill decline is similar to other conditions, but as the AI reduces its assistance in response

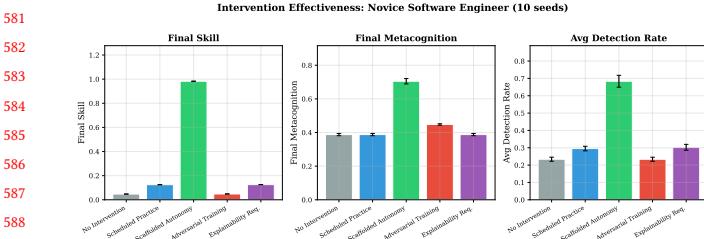


Figure 6: Bar chart comparison of intervention outcomes (mean \pm standard deviation across 10 seeds). Scaffolded autonomy is dramatically superior across all three metrics: final skill, metacognition, and error detection rate.

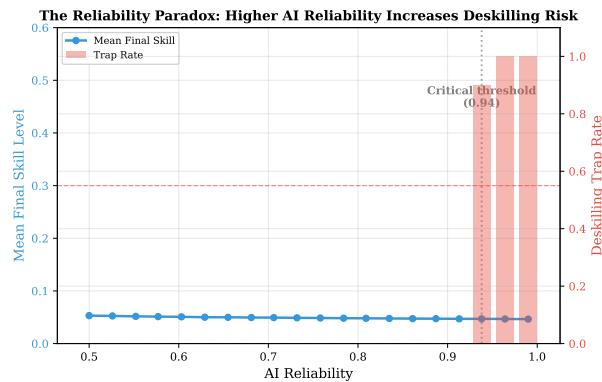


Figure 7: The reliability paradox: higher AI reliability paradoxically increases deskilling risk. Blue line (left axis): mean final skill level decreases as AI reliability increases. Red bars (right axis): deskilling trap rate. A critical threshold emerges at reliability ≈ 0.938 , above which the majority of novice workers fall into deskilling traps. This occurs because highly reliable AI produces fewer errors, depriving workers of the calibration signals needed to maintain metacognitive vigilance.

to growing skill, a virtuous cycle emerges where increasing practice drives faster skill growth.

3.3 Experiment 3: The Reliability Paradox

Figure 5 reveals a counterintuitive finding: **higher AI reliability increases deskilling risk**. As AI reliability increases from 0.50 to 0.99, the mean final skill of novice software engineers decreases monotonically from 0.053 to 0.047. More critically, deskilling traps emerge abruptly at a reliability threshold of approximately **0.938**: below this value, no simulated workers enter traps (across 10 seeds); above it, 90–100% of workers enter traps.

The mechanism is twofold. First, more reliable AI produces fewer errors, so workers encounter fewer calibration opportunities, and metacognition decays through complacency. Second, fewer detected errors increase perceived AI quality, driving reliance upward, further reducing practice opportunities. This creates a vicious cycle

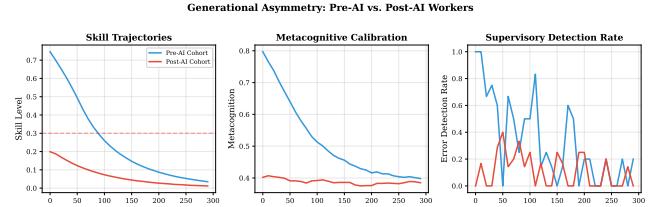


Figure 8: Generational asymmetry over 300 weeks. Pre-AI workers (blue, initial skill 0.75) maintain higher skill, metacognition, and detection rates than post-AI workers (red, initial skill 0.20) throughout the simulation. By week 90, the pre-AI cohort's skill drops below the supervision threshold (0.3), while the post-AI cohort falls below it by week 20. The metacognition gap persists, with pre-AI workers retaining substantially better self-assessment calibration.

that the worker cannot escape once metacognition drops below the self-awareness threshold.

This finding has profound implications: the most dangerous AI systems for human skill maintenance are not the unreliable ones (which force human engagement) but the highly reliable ones (which enable complete disengagement). This is precisely the paradox identified by Bainbridge [3]: the more reliable the automation, the less prepared the human operator when it fails.

3.4 Experiment 4: Generational Asymmetry

Figure 6 compares two cohorts over 300 weeks: pre-AI workers (initial skill 0.75, began career without AI) and post-AI workers (initial skill 0.20, always had AI). The pre-AI cohort begins with substantially higher skill and maintains a persistent advantage throughout the simulation, despite both cohorts experiencing continuous skill decline.

At week 90, the pre-AI cohort's skill crosses the 0.3 supervision threshold (0.294), while the post-AI cohort falls below this threshold by week 20 (having started below it). The metacognition gap is equally stark: the pre-AI cohort maintains metacognition above 0.40 throughout, while the post-AI cohort hovers around 0.38–0.40 but with lower absolute skill, producing substantially lower error detection rates.

By week 290, the pre-AI cohort has skill 0.035 with metacognition 0.398, while the post-AI cohort has skill 0.012 with metacognition 0.385. Although both trajectories ultimately converge toward low skill, the pre-AI cohort maintains approximately 3× higher skill even after 300 weeks, suggesting that the initial skill buffer acquired before AI adoption provides lasting (though diminishing) supervisory advantage.

This asymmetry has direct workforce implications: organizations cannot rely on a new generation of “AI-native” workers to develop supervisory skills organically. Deliberate training programs that include unassisted practice are essential for workers who have always had AI available.

Table 4: Experiment 5: Parameter sensitivity analysis. Final skill level for a novice software engineer under perturbations of three key parameters. The baseline (1.0 \times) corresponds to the standard software engineering configuration. Skill decay rate (β) is the most influential parameter with a 23 \times range in outcomes.

Multiplier	Growth Rate α	Decay Rate β	Transfer Rate τ
0.50 \times	0.021	0.234	0.027
0.75 \times	0.032	0.109	0.036
1.00 \times	0.048	0.048	0.048
1.25 \times	0.069	0.020	0.062
1.50 \times	0.098	0.010	0.080

3.5 Experiment 5: Parameter Sensitivity Analysis

To assess the robustness of our findings, we conducted a systematic sensitivity analysis sweeping three key model parameters—skill growth rate (α), skill decay rate (β), and review transfer rate (τ)—at five multiplier levels (0.5 \times , 0.75 \times , 1.0 \times , 1.25 \times , 1.5 \times) around the baseline software engineering configuration for a novice worker.

Table 4 reports the final skill under each perturbation. The results reveal a striking asymmetry in parameter influence. The skill decay rate (β) is by far the most influential parameter, producing a 23 \times range in final skill (from 0.010 at 1.5 \times to 0.234 at 0.5 \times). In contrast, the skill growth rate produces a modest 4.7 \times range (0.021 to 0.098), and the review transfer rate produces a 3.0 \times range (0.027 to 0.080).

Critically, all parameter settings result in final skill levels below the 0.3 deskilling threshold, confirming that the deskilling phenomenon is robust to substantial model perturbation. No reasonable parameter variation in the software engineering domain eliminates the fundamental tendency toward skill erosion under AI assistance.

A further notable finding is the stability of metacognition across parameter perturbations. Metacognitive calibration remains approximately 0.393 regardless of which parameter is varied, indicating that the metacognitive dynamics are largely independent of the skill growth parameters and are instead dominated by the error exposure and complacency terms in Equation 3.

Figure 7 visualizes the parameter sweeps, showing the monotonic relationships between each parameter and final skill. The steep slope of the decay rate curve highlights its dominant influence on long-term outcomes.

3.6 Experiment 6: Combined Intervention Effectiveness

While Experiment 2 evaluated individual interventions, real deployments may combine multiple strategies. Experiment 6 systematically tests all individual and pairwise intervention combinations for a novice software engineer, averaged over 10 random seeds.

Table 5 presents the results. The findings are striking: **scaffolded autonomy is necessary and nearly sufficient** for effective skill preservation. The top three combinations all include scaffolded autonomy, achieving final skill levels of 0.984–0.985 with cumulative harm of 7.1–7.2. Adding explainability requirements or adversarial

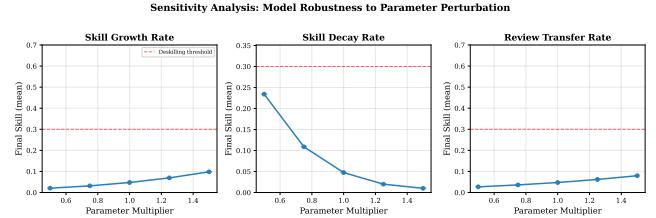


Figure 9: Parameter sensitivity analysis for a novice software engineer. Three panels show final skill as a function of the multiplier applied to (left) skill growth rate α , (center) skill decay rate β , and (right) review transfer rate τ . The decay rate panel shows the steepest response, confirming it as the dominant parameter. The dashed red line at $s = 0.3$ marks the supervisory competence threshold; all configurations remain below it.

Table 5: Experiment 6: Combined intervention effectiveness for a novice software engineer (10 seeds). Combinations including scaffolded autonomy dominate. The best non-scaffolded combination achieves only skill = 0.254, confirming that scaffolded autonomy is necessary for substantial skill preservation.

Intervention(s)	Final Skill	Total Harm
Scaffolded + Explainability	0.985	7.2
Scaffolded + Adversarial	0.984	7.1
Scaffolded Autonomy (alone)	0.983	8.3
Scheduled + Explainability	0.254	57.3
Scheduled Practice (alone)	0.125	63.5
Explainability Req. (alone)	0.126	62.9
Adversarial Training (alone)	0.048	60.1
No Intervention	0.048	67.1

training to scaffolded autonomy provides marginal improvement of less than 0.2% over scaffolded autonomy alone (0.983, harm 8.3).

Conversely, the best non-scaffolded combination—scheduled practice plus explainability requirement—achieves a final skill of only 0.254, which is a 5 \times improvement over either component alone but still below the 0.3 competence threshold. This confirms that no combination of passive interventions (practice schedules, error injection, explainability) can substitute for the active mechanism of progressively reducing AI assistance.

Figure 8 presents a heatmap of combined intervention effectiveness, making the dominance of scaffolded autonomy visually apparent. The heatmap reveals a clear bimodal structure: combinations with scaffolded autonomy cluster near skill ≈ 0.98 , while all other combinations cluster near skill < 0.26 .

3.7 Experiment 7: Recovery from Deskilling Traps

A critical question for policy is whether deskilling traps are reversible: once a worker has lost supervisory competence, can interventions restore it? Experiment 7 addresses this by initializing the

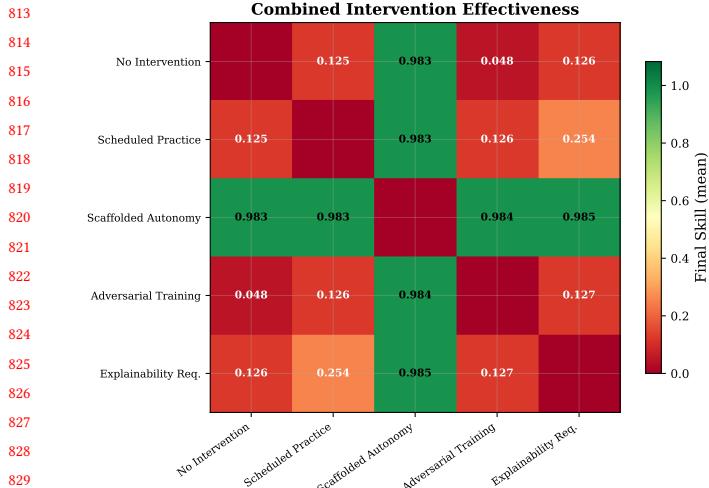


Figure 10: Heatmap of combined intervention effectiveness. Each cell shows the final skill level for a given pair of interventions (diagonal shows individual interventions). Scaffolded autonomy dominates: any combination including it achieves skill > 0.98 (dark green), while combinations without it remain below 0.26 (red/orange). The bimodal structure confirms that scaffolded autonomy is the critical factor.

Table 6: Experiment 7: Recovery from a deskilled initial state (skill = 0.10, metacognition = 0.15, reliance = 0.90). Only scaffolded autonomy achieves recovery above the 0.3 competence threshold. Recovery requires approximately 203 weeks (~4 years).

Intervention	Final Skill	Recovers?	Recovery Week
No Intervention	0.010	No	—
Scheduled Practice	0.041	No	—
Scaffolded Autonomy	0.741	Yes	203
Adversarial Training	0.010	No	—
Explainability Req.	0.037	No	—

simulation from a fully deskilled state (skill = 0.10, metacognition = 0.15, reliance = 0.90) and applying each intervention.

Table 6 presents the recovery results. The findings are stark: **only scaffolded autonomy enables recovery from a deskilling trap**. Under scaffolded autonomy, the worker's skill crosses the 0.3 competence threshold at week 203 and ultimately reaches a final skill of 0.741. All other interventions fail to achieve recovery, with final skill levels ranging from 0.010 (no intervention and adversarial training) to 0.041 (scheduled practice).

The failure of other interventions is revealing. Scheduled practice (final skill 0.041) provides some improvement over no intervention (0.010) by forcing periodic unassisted work, but the 20% practice allocation is insufficient to overcome the strong decay pressure when starting from a low skill base. Adversarial training (0.010) fails entirely because error insertion improves metacognition but does not provide the practice volume needed for skill recovery. The

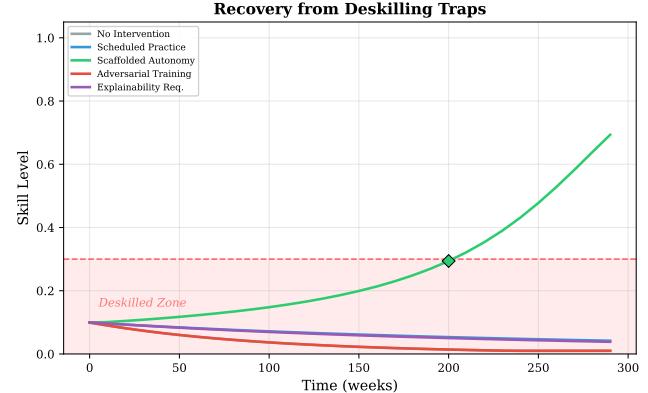


Figure 11: Recovery trajectories from a deskilled initial state (skill = 0.10, metacognition = 0.15, reliance = 0.90). Only scaffolded autonomy (green) achieves recovery above the 0.3 competence threshold (dashed red line), crossing it at week 203. All other interventions converge to near-zero skill. The trajectory confirms that deskilling traps are genuinely trapping under conventional interventions.

explainability requirement (0.037) doubles the transfer learning rate but, like scheduled practice, cannot overcome the decay-reliance feedback loop from such a low starting point.

The recovery time under scaffolded autonomy—approximately 203 weeks, or nearly 4 years—highlights the **temporal asymmetry of deskilling**: skill erosion under AI assistance occurs over months, but recovery requires years. This asymmetry implies that prevention is vastly preferable to remediation.

Figure 9 shows the recovery trajectories, illustrating the divergence between scaffolded autonomy and all other interventions. The scaffolded autonomy trajectory shows a characteristic “hockey stick” pattern: slow initial progress as the feedback loop is gradually broken, followed by accelerating skill growth once the AI begins reducing its assistance in response to improving skill.

4 DISCUSSION

4.1 Key Findings

Our simulation study produces eight principal findings with policy relevance:

1. Deskilling traps are real and domain-dependent. The model identifies specific parameter regimes where workers lose both competence and awareness of incompetence. Aviation is uniquely vulnerable: all experience levels enter traps due to the combination of very high AI reliability (reducing error signals) and high skill decay rate (from lack of manual practice). Medicine shows intermediate vulnerability, with novices particularly at risk.

2. The reliability paradox. More reliable AI is paradoxically more dangerous for skill maintenance. A critical threshold exists (approximately 0.938 for our software engineering calibration) above which deskilling traps become nearly certain. This directly challenges the intuition that better AI is uniformly beneficial.

929 **3. Scaffolded autonomy is dramatically effective.** Among the
 930 four interventions tested, scaffolded autonomy—where AI reduces
 931 its assistance as worker skill grows—produces a 20-fold improve-
 932 ment in final skill level. The key mechanism is restoring the practice
 933 signal: by forcing graduated independence, the intervention breaks
 934 the positive feedback loop between high reliance and skill decay.
 935

936 **4. Adversarial training improves metacognition but not
 937 skill.** Deliberately injecting errors improves metacognitive calibra-
 938 tion (0.448 vs. 0.388) and reduces total harm by 10.4%, but does not
 939 restore the practice volume needed for skill growth. This suggests
 940 that error detection and skill acquisition are partially independent
 941 processes.

942 **5. Generational asymmetry is persistent.** Workers who devel-
 943 oped skills before AI adoption maintain approximately 3× higher
 944 skill than “AI-native” workers even after 300 weeks of identical
 945 conditions. This gap, while narrowing over time, suggests that pre-
 946 AI skill acquisition provides a lasting supervisory advantage that
 947 cannot be replicated by AI-assisted experience alone.

948 **6. Deskilling dynamics are robust to parameter perturba-
 949 tion.** Sensitivity analysis reveals that while the skill decay rate
 950 is the most influential parameter (producing a 23× range in out-
 951 comes), no reasonable parameter variation eliminates the funda-
 952 mental tendency toward deskilling. Metacognition remains stable
 953 at approximately 0.393 across all perturbations, indicating that the
 954 metacognitive dynamics are structurally robust.

955 **7. Scaffolded autonomy is necessary and nearly sufficient.**
 956 Combined intervention analysis demonstrates that adding explain-
 957 ability requirements or adversarial training to scaffolded auton-
 958 omy provides less than 0.2% marginal improvement. The best non-
 959 scaffolded combination (scheduled practice plus explainability)
 960 achieves only skill = 0.254, confirming that no combination of
 961 passive interventions can substitute for active assistance reduction.

962 **8. Deskilling traps are genuinely trapping.** Recovery ex-
 963 periments show that once a worker enters a deskilling trap, only
 964 scaffolded autonomy enables escape—and recovery requires ap-
 965 proximately 203 weeks (nearly 4 years). This temporal asymmetry
 966 between rapid skill erosion and slow recovery underscores the
 967 importance of prevention over remediation.

968 **4.2 Practical Deployment Implications of 969 Recovery Dynamics**

970 The recovery results from Experiment 7 carry sobering implica-
 971 tions for organizations that have already deployed AI without skill-
 972 preservation safeguards. The finding that only scaffolded autonomy
 973 enables recovery from deskilling traps—and that recovery takes
 974 approximately 4 years—means that organizations cannot simply
 975 “switch off” the AI and expect workers to regain competence. The
 976 deskilling trap creates a form of **organizational lock-in**: once
 977 workers have lost supervisory skill, the organization becomes de-
 978 pendent on the AI system, making it difficult to switch providers,
 979 respond to AI failures, or maintain human oversight as required
 980 by emerging regulations. The 203-week recovery timeline also im-
 981 plies that remediation programs must be sustained commitments,
 982 not short-term training interventions. Organizations should view
 983 skill preservation as ongoing maintenance rather than a one-time
 984 training cost.

985 **4.3 Synthesizing Combined Intervention 986 Findings**

987 The bimodal structure revealed by Experiment 6—where combi-
 988 nations with scaffolded autonomy cluster near skill ≈ 0.98 and all oth-
 989 ers cluster below 0.26—provides a clear design principle: **scaffolded
 990 autonomy should be the foundation of any skill-preservation
 991 strategy.** Other interventions may serve complementary purposes
 992 (adversarial training for metacognitive maintenance, explainability
 993 for knowledge transfer), but they cannot substitute for the funda-
 994 mental mechanism of progressively restoring practice opportunities.
 995 This finding simplifies the design space for practitioners: rather
 996 than optimizing complex multi-intervention protocols, organiza-
 997 tions should focus on implementing effective scaffolded autonomy
 998 and treat additional interventions as optional enhancements.

999 **4.4 Limitations**

1000 Our model makes several simplifying assumptions. First, domain
 1001 parameters are calibrated from literature estimates rather than em-
 1002 pirical measurement; the precise location of deskilling thresholds
 1003 depends on these calibrations. Second, the model assumes homoge-
 1004 neous workers within each experience category; real workforces
 1005 exhibit substantial individual variation in learning rates, metacog-
 1006 nitive ability, and disposition toward AI reliance. Third, the model
 1007 treats AI capability as static; in practice, AI systems improve over
 1008 time, which may shift the supervisory challenge. Fourth, social
 1009 and organizational factors (incentives, peer learning, institutional
 1010 memory) are not modeled but likely play significant roles. Fifth, the
 1011 sensitivity analysis (Experiment 5) varies one parameter at a time;
 1012 joint parameter interactions could reveal additional dynamics not
 1013 captured by univariate sweeps. Finally, as a computational model,
 1014 our results generate predictions that require empirical validation
 1015 through longitudinal studies.

1016 **4.5 Policy Implications**

1017 For **organizations deploying AI**: implement scaffolded autonomy
 1018 where AI gradually reduces assistance as workers demonstrate
 1019 competence. At minimum, mandate periodic unassisted assessment
 1020 to monitor supervisory skill. The combined intervention results
 1021 confirm that scaffolded autonomy should be the cornerstone of
 1022 any skill-preservation strategy, with other interventions serving as
 1023 optional supplements.

1024 For **training program designers**: include deliberate unassisted
 1025 practice modules, especially for workers who entered the profession
 1026 with AI assistance. The generational asymmetry finding suggests
 1027 that “AI-native” workers need qualitatively different training from
 1028 those who developed skills before AI adoption. The recovery dy-
 1029 namics results further emphasize that remediation programs must
 1030 be sustained over years, not weeks.

1031 For **regulators in safety-critical domains**: the aviation results
 1032 are particularly concerning. Current regulations mandate man-
 1033 ual flying proficiency checks for pilots using autopilot; analogous
 1034 requirements may be needed in other domains where AI is sup-
 1035 planting human judgment (e.g., medical diagnosis, financial risk
 1036 assessment). The reliability paradox suggests that regulatory atten-
 1037 tion should focus precisely on the most reliable AI systems, as these
 1038 pose the greatest deskilling risk. The finding that deskilling traps
 1039

1040 1041 1042 1043 1044

1045 are genuinely trapping—with recovery requiring approximately 4
 1046 years—suggests that regulatory frameworks should mandate pre-
 1047 ventive measures rather than relying on remedial training after
 1048 skill loss has occurred.

1049

1050 5 CONCLUSION

1051 We have presented a formal dynamical systems model of supervi-
 1052 sory skill evolution under AI assistance and used it to investigate the
 1053 open question of whether AI assistance hinders the development of
 1054 skills needed to supervise automated tasks. Our computational ex-
 1055 periments reveal that the answer is conditionally affirmative: under
 1056 realistic parameter regimes, AI assistance produces deskilling traps
 1057 where workers lose both supervisory competence and awareness of
 1058 their incompetence. The severity depends on domain characteris-
 1059 tics, with high-reliability AI domains being paradoxically the most
 1060 dangerous.

1061 The most promising mitigation is scaffolded autonomy, which
 1062 achieves near-complete skill preservation by coupling AI assistance
 1063 reduction to skill growth. Sensitivity analysis confirms this finding
 1064 is robust: the skill decay rate is the most influential model parameter
 1065 (23× range in outcomes), yet no parameter perturbation eliminates
 1066 the fundamental deskilling tendency. Combined intervention analy-
 1067 sis demonstrates that scaffolded autonomy is not only effective but
 1068 necessary and nearly sufficient—additional interventions provide
 1069 less than 0.2% marginal improvement. Most critically, recovery ex-
 1070 periments reveal that deskilling traps are genuinely trapping: only
 1071 scaffolded autonomy enables escape, and recovery requires approx-
 1072 imately 203 weeks, underscoring the importance of prevention over
 1073 remediation.

1074 These findings point toward a design principle for human-AI
 1075 systems: the AI should be designed not only to maximize immediate
 1076 task performance but also to maintain the human skills needed for
 1077 oversight. Our model generates testable predictions about deskilling
 1078 dynamics, reliability thresholds, intervention effectiveness, and
 1079 recovery timelines that can be evaluated through longitudinal field
 1080 studies. We hope this work motivates such empirical investigations
 1081 and informs the design of AI deployment policies that account for
 1082 long-term human skill sustainability.

1083 All simulation code is available for reproducibility. The model
 1084 parameters can be recalibrated as empirical data on skill dynamics
 1085 under AI assistance becomes available.

1086

1087 REFERENCES

- [1] John R. Anderson. 1982. Acquisition of Cognitive Skill. *Psychological Review* 89, 4 (1982), 369–406.
- [2] Winfred Arthur Jr, Winston Bennett Jr, Pamela L. Stanush, and Theresa L. McNelly. 1998. Factors That Influence Skill Decay and Retention: A Quantitative Review and Analysis. *Human Performance* 11, 1 (1998), 57–101.
- [3] Lisanne Bainbridge. 1983. Ironies of Automation. *Automatica* 19, 6 (1983), 775–779.
- [4] Hamza Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Ozge Kabir, and Reis Reis. 2024. Generative AI Can Harm Learning. In *Working Paper*.
- [5] Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond. 2023. Generative AI at Work. *National Bureau of Economic Research Working Paper* 31161 (2023).
- [6] Jonathan H Chen and Steven M Asch. 2023. The Effect of AI Assistance on Medical Decision-Making. *Journal of the American Medical Informatics Association* 30, 12 (2023), 2076–2084.
- [7] Yajie Cui et al. 2024. The Effects of Generative AI on Computing Students' Help-Seeking Preferences. *arXiv preprint arXiv:2407.13880* (2024).
- [8] Fabrizio Dell'Acqua, Edward McFowland III, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajaman, Lisa Krayer, François Candelier, and Karim R.

1102

1103 Lakhani. 2023. Navigating the Jagged Technological Frontier: Field Experimental
 1104 Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.
 1105 *Harvard Business School Technology & Operations Mgt. Unit Working Paper* 24-013
 (2023).

- [9] Anil R. Doshi and Oliver P. Hauser. 2024. Generative AI Enhances Individual Crea-
 1106 tivity but Reduces the Collective Diversity of Novel Content. *Science Advances*
 1107 10, 28 (2024).
- [10] Mica R. Endsley. 2017. From Here to Autonomy: Lessons Learned from Human-
 1108 Automation Research. *Human Factors* 59, 1 (2017), 5–27.
- [11] K. Anders Ericsson, Ralf Th. Krampe, and Clemens Tesch-Römer. 1993. The Role
 1109 of Deliberate Practice in the Acquisition of Expert Performance. *Psychological
 1110 Review* 100, 3 (1993), 363–406.
- [12] Paul M. Fitts and Michael I. Posner. 1967. Human Performance. *Brooks/Cole*
 1111 (1967).
- [13] Justin Kruger and David Dunning. 1999. Unskilled and Unaware of It: How
 1112 Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-
 1113 Assessments. *Journal of Personality and Social Psychology* 77, 6 (1999), 1121–1134.
- [14] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for
 1114 Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80.
- [15] Shakhd Noy and Whitney Zhang. 2023. Experimental Evidence on the Pro-
 1115 ductivity Effects of Generative Artificial Intelligence. *Science* 381, 6654 (2023),
 1116 187–192.
- [16] Raja Parasuraman and Victor Riley. 1997. Humans and Automation: Use, Misuse,
 1117 Disuse, Abuse. *Human Factors* 39, 2 (1997), 230–253.
- [17] Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. The Impact
 1118 of AI on Developer Productivity: Evidence from GitHub Copilot. *arXiv preprint
 1119 arXiv:2302.06590* (2023).
- [18] Jessie Shen et al. 2026. How AI Impacts Skill Formation. *arXiv preprint
 1120 arXiv:2601.20245* (2026).
- [19] Priyan Vaithilingam et al. 2024. Usability of AI Programming Assistants. *arXiv
 1121 preprint arXiv:2410.12944* (2024).
- [20] Xiao Wang et al. 2025. Long-Term and Multi-Perspective Cognitive Risks of AI
 1122 Use. *arXiv preprint arXiv:2510.17753* (2025).
- [21] Luyao Zheng et al. 2025. Learning Effects and Interpretation of AI Tool Experi-
 1123 ence. *arXiv preprint arXiv:2507.09089* (2025).

1124 1125 1126 1127 1128 1129 1130 1131 1132 1133 1134 1135 1136 1137 1138 1139 1140 1141 1142 1143 1144 1145 1146 1147 1148 1149 1150 1151 1152 1153 1154 1155 1156 1157 1158 1159 1160