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.

2.2 Skill Dynamics

The skill level evolves as:

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

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:

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

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

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 †. 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 Harm
Software	Novice	0.048	0.390	0.262	66.6
	Intermediate	0.075	0.404	0.408	54.4
	Expert	0.106	0.429	0.542	43.8
Medicine	Novice†	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†	0.010	0.187	0.213	65.5
	Intermediate†	0.010	0.218	0.159	71.3
	Expert†	0.010	0.242	0.357	53.5

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.

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: $r_{\text{eff}} = r \cdot (1 - 0.5s)$.
(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.

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.
- **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× to 1.5×) 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.

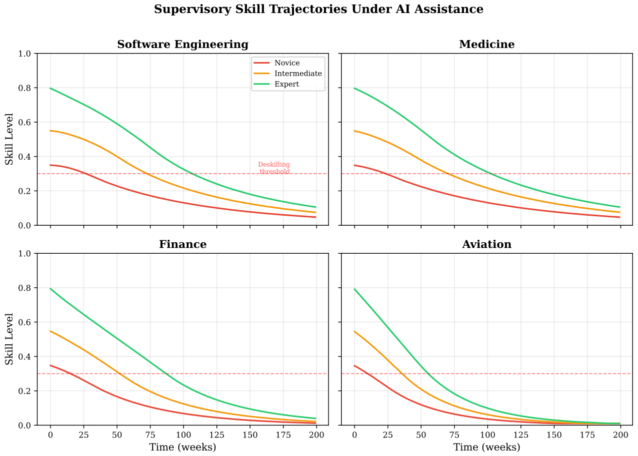

Figure 1: Supervisory skill trajectories over 200 weeks across four domains and three experience levels (novice, intermediate, expert). The dashed red line at $s = 0.3$ marks the supervisory competence threshold. All trajectories decline, but the rate and final level depend on domain characteristics. Aviation shows the most severe decline due to high AI reliability and low review transfer.

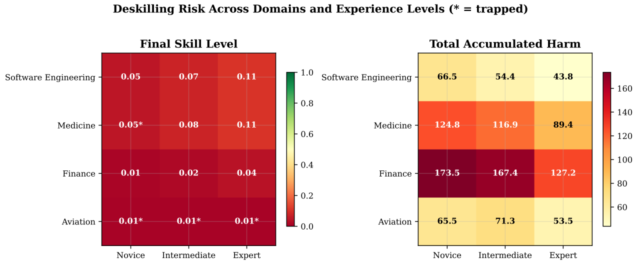

Figure 2: Deskilling risk heatmap across domains and experience levels. Left: final skill level (green = high, red = low). Right: total accumulated harm over 200 weeks. Asterisks (*) mark deskilling trap states. Aviation is uniquely vulnerable across all experience levels, while finance accumulates the most harm due to moderate AI reliability and high task novelty.

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 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.

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

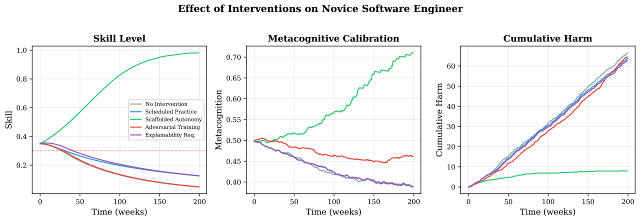

Figure 3: Intervention trajectories for a novice software engineer over 200 weeks. Left: skill level. Center: metacognitive calibration. Right: cumulative harm. Scaffolded autonomy (green) is the only intervention that reverses the deskilling trajectory, achieving near-expert skill levels. The remaining interventions slow but do not prevent skill decline.

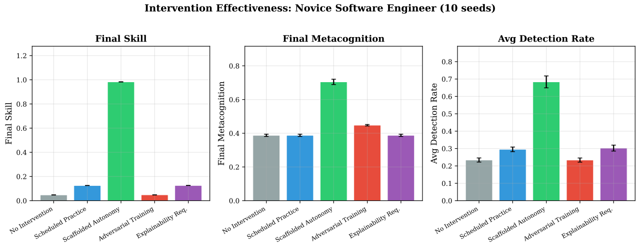

Figure 4: Bar chart comparison of intervention outcomes (mean ± standard deviation across 10 seeds). Scaffolded autonomy is dramatically superior across all three metrics: final skill, metacognition, and error detection rate.

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 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

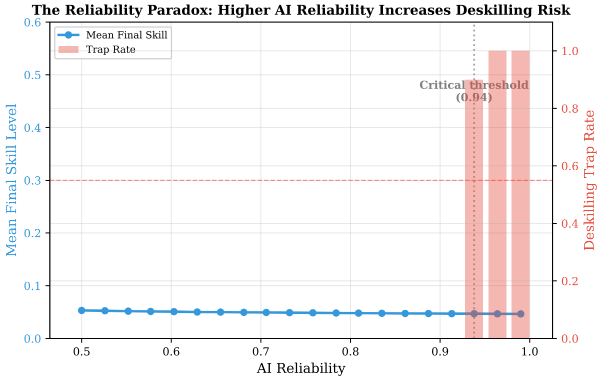

Figure 5: The reliability paradox: higher AI reliability paradoxically increases deskilling risk. Blue line (left axis): mean final skill level decreases as AI reliability increases. Red bars (right axis): deskilling trap rate. A critical threshold emerges at reliability ≈ 0.938, above which the majority 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.

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 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.

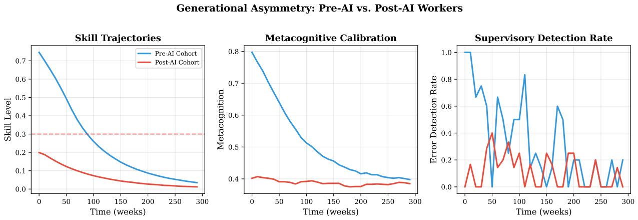

Figure 6: Generational asymmetry over 300 weeks. Pre-AI workers (blue, initial skill 0.75) maintain higher skill, metacognition, and detection rates than post-AI workers (red, initial skill 0.20) throughout the simulation. By week 90, the pre-AI cohort's skill drops below the supervision threshold (0.3), 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.

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.

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×, 0.75×, 1.0×, 1.25×, 1.5×) 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× range** in final skill (from 0.010 at 1.5× to 0.234 at 0.5×). In contrast, the skill growth rate produces a modest 4.7× range (0.021 to 0.098), and the review transfer rate produces a 3.0× 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.

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×) corresponds to the standard software engineering configuration. Skill decay rate (β) is the most influential parameter with a 23× range in outcomes.

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

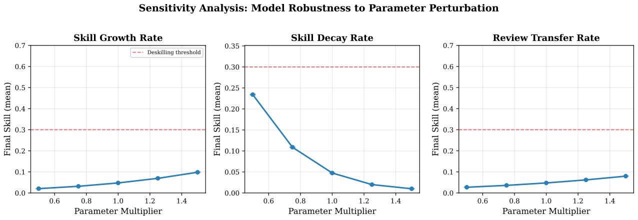

Figure 7: Parameter sensitivity analysis for a novice software engineer. Three panels show final skill as a function of the multiplier applied to (left) skill growth rate α, (center) skill decay rate β, and (right) review transfer rate τ. The decay rate panel shows the steepest response, confirming it as the dominant parameter. The dashed red line at $s = 0.3$ marks the supervisory competence threshold; all configurations remain below it.

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 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× 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.

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

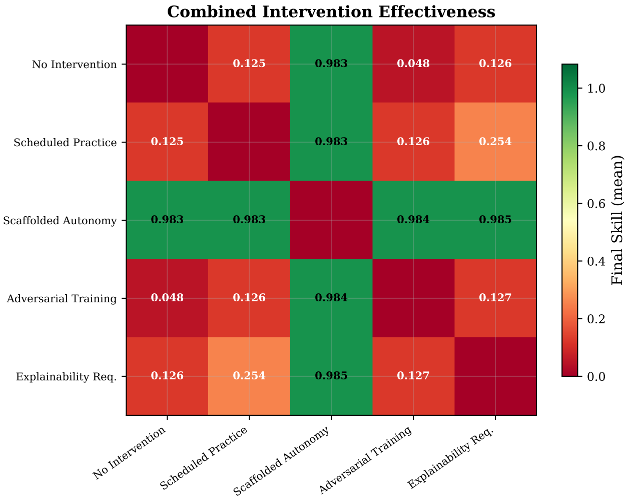

Figure 8: Heatmap of combined intervention effectiveness. Each cell shows the final skill level for a given pair of interventions (diagonal shows individual interventions). Scaffolded autonomy dominates: any combination including it achieves skill > 0.98 (dark green), while combinations without it remain below 0.26 (red/orange). The bimodal structure confirms that scaffolded autonomy is the critical factor.

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.

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	—

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 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 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.

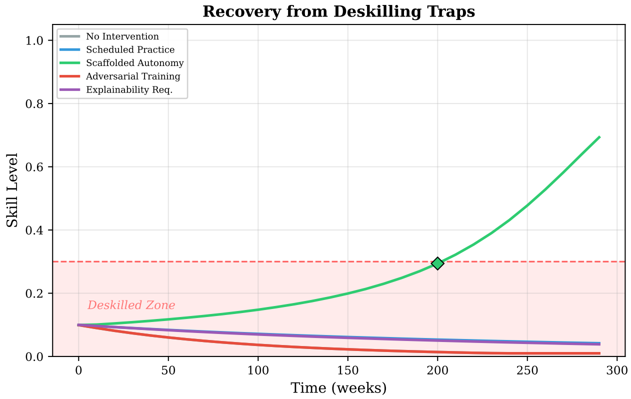

Figure 9: Recovery trajectories from a deskilled initial state (skill = 0.10, metacognition = 0.15, reliance = 0.90). Only scaffolded autonomy (green) achieves recovery above the 0.3 competence threshold (dashed red line), crossing it at week 203. All other interventions converge to near-zero skill. The trajectory confirms that deskilling traps are genuinely trapping under conventional interventions.

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.

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

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

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

6. Deskilling dynamics are robust to parameter perturbation. Sensitivity analysis reveals that while the skill decay rate is the most influential parameter (producing a 23× range in outcomes), no reasonable parameter variation eliminates the fundamental tendency toward deskilling. Metacognition remains stable at approximately 0.393 across all perturbations, indicating that the metacognitive dynamics are structurally robust.

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

8. Deskilling traps are genuinely trapping. Recovery experiments show that once a worker enters a deskilling trap, only scaffolded autonomy enables escape—and recovery requires approximately 203 weeks (nearly 4 years). This temporal asymmetry between rapid skill erosion and slow recovery underscores the importance of prevention over remediation.

4.2 Practical Deployment Implications of Recovery Dynamics

The recovery results from Experiment 7 carry sobering implications for organizations that have already deployed AI without skill-preservation safeguards. The finding that only scaffolded autonomy enables recovery from deskilling traps—and that recovery takes approximately 4 years—means that organizations cannot simply "switch off" the AI and expect workers to regain competence. The deskilling trap creates a form of **organizational lock-in**: once workers have lost supervisory skill, the organization becomes dependent on the AI system, making it difficult to switch providers, respond to AI failures, or maintain human oversight as required by emerging regulations. The 203-week recovery timeline also implies that remediation programs must be sustained commitments, not short-term training interventions. Organizations should view skill preservation as ongoing maintenance rather than a one-time training cost.

4.3 Synthesizing Combined Intervention Findings

The bimodal structure revealed by Experiment 6—where combinations with scaffolded autonomy cluster near skill ≈ 0.98 and all others cluster below 0.26—provides a clear design principle: **scaffolded autonomy should be the foundation of any skill-preservation strategy**. Other interventions may serve complementary purposes (adversarial training for metacognitive maintenance, explainability for knowledge transfer), but they cannot substitute for the fundamental mechanism of progressively restoring practice opportunities. This finding simplifies the design space for practitioners: rather than optimizing complex multi-intervention protocols, organizations should focus on implementing effective scaffolded autonomy and treat additional interventions as optional enhancements.

4.4 Limitations

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

4.5 Policy Implications

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

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

For **regulators in safety-critical domains**: the aviation results are particularly concerning. Current regulations mandate manual flying proficiency checks for pilots using autopilot; analogous requirements may be needed in other domains where AI is supplanting human judgment (e.g., medical diagnosis, financial risk assessment). The reliability paradox suggests that regulatory attention should focus precisely on the most reliable AI systems, as these pose the greatest deskilling risk. The finding that deskilling traps are genuinely trapping—with recovery requiring approximately 4 years—suggests that regulatory frameworks should mandate preventive measures rather than relying on remedial training after skill loss has occurred.

5 CONCLUSION

We have presented a formal dynamical systems model of supervisory skill evolution under AI assistance and used it to investigate the open question of whether AI assistance hinders the development of skills needed to supervise automated tasks. Our computational experiments reveal that the answer is conditionally affirmative: under realistic parameter regimes, AI assistance produces deskilling traps where workers lose both supervisory competence and awareness of their incompetence. The severity depends on domain characteristics, with high-reliability AI domains being paradoxically the most dangerous.

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

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

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

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

[9] Anil R. Doshi and Oliver P. Hauser. 2024. Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content. *Science Advances* 10, 28 (2024).

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

[11] K. Anders Ericsson, Ralf Th. Krampe, and Clemens Tesch-Römer. 1993. The Role of Deliberate Practice in the Acquisition of Expert Performance. *Psychological Review* 100, 3 (1993), 363–406.

[12] Paul M. Fitts and Michael I. Posner. 1967. Human Performance. *Brooks/Cole* (1967).

[13] Justin Kruger and David Dunning. 1999. Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-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 Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80.

[15] Shakked Noy and Whitney Zhang. 2023. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. *Science* 381, 6654 (2023), 187–192.

[16] Raja Parasuraman and Victor Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors* 39, 2 (1997), 230–253.

[17] Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. *arXiv preprint arXiv:2302.06590* (2023).

[18] Jessie Shen et al. 2026. How AI Impacts Skill Formation. *arXiv preprint arXiv:2601.20245* (2026).

[19] Priyan Vaithilingam et al. 2024. Usability of AI Programming Assistants. *arXiv preprint arXiv:2410.12944* (2024).

[20] Xiao Wang et al. 2025. Long-Term and Multi-Perspective Cognitive Risks of AI Use. *arXiv preprint arXiv:2510.17753* (2025).

[21] Luyao Zheng et al. 2025. Learning Effects and Interpretation of AI Tool Experience. *arXiv preprint arXiv:2507.09089* (2025).