Advancing Expertise Development Through Adaptive Human-AI Training

Jessica M. Johnson
Old Dominion University
Norfolk, VA
J17johnso@odu.edu

ABSTRACT

Adaptive automation in complex training environments continuously adjusts to varying human expertise levels, yet there is limited synthesis of how these systems scale support, extract knowledge, and enhance performance as learners progress from novice to expert. As training environments grow increasingly complex and reliant on human-AI teaming, there is a critical need to understand how automation can not only support performance but actively foster expertise development through intelligent knowledge capture. This meta-analysis examined two decades of empirical research on adaptive automation across high-stakes domains, identifying how AI-driven training systems adapt based on expertise, utilize knowledge elicitation techniques, and formalize knowledge extraction to refine system intelligence. This research provided the first cross-domain synthesis of adaptive automation's role in AI-enabled knowledge elicitation of expertise for complex training environments (medical, military, maritime, defense, etc.). Findings revealed three major gaps: 1) an overemphasis on novice-level support with minimal longitudinal tracking of expertise development, 2) inconsistent integration of real-time knowledge elicitation methods, and 3) underdeveloped strategies for converting elicited knowledge into reusable models for adaptive systems. Cross-study themes highlighted importance of dynamically balancing automation control with human input, the need for scalable feedback tailored to cognitive and psychomotor growth, and the critical role of transparent, explainable AI adaptation to build trust across expertise levels. These insights informed the design of an Adaptive Expertise Mutual Learning Loop conceptual framework that aligns adaptive automation with continuous knowledge capture, supporting both learner growth and system evolution. By embedding expertise-driven scaffolding and iterative knowledge elicitation, the conceptual framework closes the loop between human insight and AI adaptation. It provides the foundation for future research, prototyping and design of next-generation human-AI systems that advance knowledge elicitation, extraction, and expertise development in complex training environments.

ABOUT THE AUTHORS

Jessica M. Johnson Ph.D. is a Research Assistant Professor and Director of the Applied Cognitive Engineering and Simulation Lab at Old Dominion University's Office of Enterprise Research and Innovation. She specializes in cognitive engineering and joint cognitive systems within complex, collaborative environments. Her expertise includes designing and analyzing cognitive work in sociotechnical systems, with emphasis on cognitive work analysis, human-system integration, adaptive decision support, and the modeling of macrocognitive functions. She develops intelligent training solutions that incorporate cognitive modeling techniques to support and optimize human performance in dynamic, high-stakes scenarios. Her scholarly pursuits focus on decision-making under uncertainty, macrocognitive processes, and human-AI collaboration in immersive learning environments.

Advancing Expertise Development Through Adaptive Human-AI Training

Jessica M. Johnson
Old Dominion University
Norfolk, VA
J17johnso@odu.edu

INTRODUCTION

Background and Significance

Training for high-stakes domains, such as defense, aviation, and maritime operations, demands increasingly sophisticated instructional systems capable of preparing learners for dynamic, ambiguous environments. Adaptive automation, which modifies task conditions, feedback, or support in response to real-time user states, has emerged as a promising mechanism for scaling high-fidelity training across diverse learner populations. However, despite this promise, adaptive systems often fall short in cultivating deep, transferable expertise. Instead, many prioritize short-term performance improvements, optimizing for early task success rather than longitudinal cognitive growth. This gap underscores a fundamental limitation in the current generation of training systems: a lack of integration between adaptive system logic and the developmental trajectory of expertise.

In this paper, we use the term *adaptive automation* to describe training system functions that dynamically adjust instructional support, task difficulty, or autonomy allocation in response to learner states. This concept overlaps with *adaptive training* and *adaptive learning* in education literature; however, our focus is on the automation logic that underpins adaptive instructional decisions. To reduce ambiguity, we treat these terms as interconnected, emphasizing how adaptive automation in training contexts supports learner expertise trajectories.

Problem Statement

Despite extensive interest in adaptive automation, research to date remains fragmented in its approach to supporting skill progression across the novice-to-expert continuum. Many systems are optimized for novices, with little consideration of how support should evolve as user expertise increases. Similarly, while knowledge elicitation is central to intelligent system design, current training solutions often fail to incorporate mechanisms that capture and reuse expert knowledge decision-making systematically. This results in underutilized opportunities for both learner development and system refinement.

Research Purpose and Goals

Unlike static one-size-fits-all training, adaptive systems tailor support to a learner's current expertise, providing more guidance to novices and scaling back as skill improves. They can also form human–Artificial Intelligence (AI) teams where the AI not only supports performance but also learns from the human, capturing expert knowledge over time. However, despite two decades of research, there has been limited synthesis of how these systems truly foster expertise development (i.e. the progression from novice to expert) through intelligent adaptation and knowledge capture. The purpose of this study is to synthesize empirical research on adaptive automation in training and to uncover how these systems support expertise development and intelligent knowledge capture for future research and development. This meta-analysis therefore examines empirical studies (2005–2025) on adaptive automation in training contexts, focusing on how support is adjusted by expertise level, what knowledge elicitation methods are used, and what gaps remain. We included 12 peer-reviewed studies filtered from thousands of research literature meeting our inclusion criteria. We excluded studies on static automation (no adaptation), non-training contexts, or opinion papers lacking empirical data. The resulting analysis informs the design of a conceptual framework that aligns expertise development via adaptive automation with iterative knowledge elicitation and learner progression for future research.

METHODS

Meta-Analysis Methodology

A comprehensive systematic review and meta-analysis was conducted to identify empirical studies published between 2005 and 2025 that examined adaptive automation in high-stakes training environments. The databases searched included IEEE Xplore, PubMed, Scopus, Web of Science, and ProQuest Dissertations and Theses Global. Search terms included combinations of "adaptive automation," "training systems," "expertise progression," "knowledge elicitation," "intelligent tutoring," and "AI-supported learning." Boolean operators and truncations were used to capture variation in terminology across disciplines. To ensure rigor and relevance, studies were included if they (1) were peer-reviewed and published in English; (2) reported empirical findings (quantitative or mixed-methods); (3) involved adaptive automation in training or learning environments; (4) included participants across different expertise levels (novice, intermediate, or expert); and (5) described or implemented knowledge elicitation or extraction techniques (e.g., think-aloud protocols, debriefing, CTA). Studies were excluded if they (1) relied on static automation (non-adaptive systems), (2) were purely conceptual or opinion-based without empirical data, or (3) did not address training or skill acquisition outcomes. The initial literature search identified 1,245 records. After removal of 555 duplicates, 690 articles were screened at the title and abstract level. Of those, 60 full-text articles were assessed for eligibility, with 12 ultimately meeting full inclusion criteria. These studies spanned domains including aviation, healthcare, defense, emergency response, and industrial operations. Two independent reviewers screened studies with Cohen's kappa = 0.86 for inclusion decisions.

Analytical Procedures

Each included study was systematically coded using a structured coding sheet designed to capture both descriptive and analytical data. The coding variables included: (1) publication year and domain of application; (2) participant characteristics (sample size, level of expertise); (3) types and mechanisms of adaptive automation (e.g., dynamic feedback, level of autonomy adjustment, confidence calibration); (4) knowledge elicitation methods employed (e.g., think-aloud, eye-tracking, CTA); and (5) reported learning and performance outcomes.

Quantitative analysis was conducted to compute standardized mean differences (Cohen's d) and applied a random-effects model to account for heterogeneity in study design and outcome measures where sufficient statistical data was available. These effect sizes captured the magnitude of adaptive automation's impact on performance metrics such as task accuracy, response time, and error rates. Studies were grouped by domain and participant expertise level to assess differential effects. For qualitative synthesis, a thematic analysis was performed across the included studies to identify cross-study patterns, recurring design strategies, and emergent gaps grounded in Braun and Clarke's methodology (2006) with intercoder agreement on themes reaching 92%. Themes were iteratively refined to identify current gaps and instructional patterns in adaptive automation design for expertise. This included examination of how systems adjusted support based on learner performance, how knowledge elicitation data informed system adaptation, and whether support varied longitudinally to promote skill progression. Results from both quantitative and qualitative streams were triangulated to construct a comprehensive picture of current trends and gaps in adaptive automation design for training.

KEY FINDINGS & GAPS FROM THE RESEARCH

Across the 12 studies included in the meta-analysis, adaptive training interventions demonstrated a moderate overall impact on performance outcomes (pooled Cohen's $d = 0.56$, 95% CI [0.31, 0.82]). These effects were most pronounced in novice learners, where adaptive scaffolding and graduated task difficulty yielded significant improvements in accuracy, error reduction, and task completion time (average $d = 0.72$). For intermediate learners, effect sizes were smaller (average $d = 0.45$), suggesting that while adaptive interventions continued to provide value, their relative impact diminished as learners gained competence. In advanced learners, effects were weakest (average $d = 0.32$), reflecting the limited ability of current adaptive systems to support higher-order expertise development such as recognition-primed decision making, strategic flexibility, and transfer to novel contexts.

Domain-specific analyses revealed consistent benefits in aviation, defense, and medical training contexts, particularly in tasks with high cognitive workload and time pressure. Systems that dynamically adjusted autonomy levels and task

difficulty produced the strongest effects, whereas systems relying on post-hoc debriefing or static rules showed only marginal gains. Interventions that integrated multimodal knowledge elicitation, such as combining performance data with physiological indicators, demonstrated higher effectiveness compared to single-modality systems, but such approaches were rare across the dataset.

Taken together, these findings within the meta-analysis provide three broad conclusions and multiple overall gaps. First, adaptive training systems are demonstrably effective in accelerating early-stage learning, but they remain underdeveloped for sustained, longitudinal expertise growth. Second, current adaptive logic relies too heavily on surface-level performance data, which limits its ability to recognize and scaffold deeper cognitive changes essential for expert performance. Third, the absence of persistent learner models and reusable AI knowledge structures prevents systems from evolving pedagogically over time. These conclusions underscore the necessity of moving beyond short-term performance optimization toward frameworks that embed continuous knowledge capture, expertise-sensitive scaffolding, and mutual human–AI learning loops.

Gap 1: Overemphasis on Early-Stage Training: Limited Longitudinal Models of Expert Development

Across domains, adaptive training systems consistently modulate the level and type of support based on the trainee's expertise, aligning with learning science principles like the expertise reversal effect (where instruction beneficial to novices can hinder experts) (Ramirez et al., 2018). Our meta-analysis found that the most effective adaptive interventions were those that dynamically increased challenge or autonomy as learners gain skill and conversely provide scaffolding when learners are unskilled. Figure 1, below, illustrates the aggregated effect sizes for different adaptive strategies from this meta-analysis.

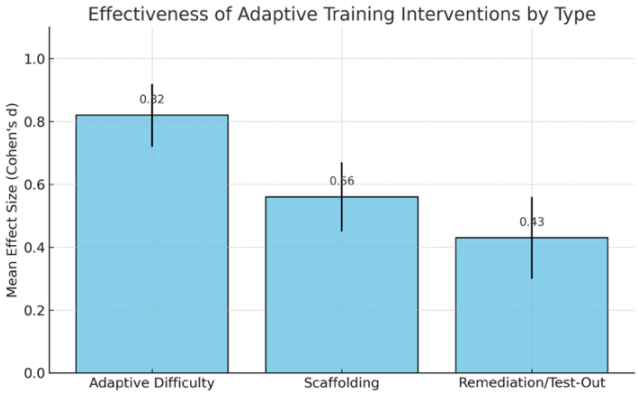

Figure 1. Effectiveness of adaptive training interventions by type.

A consistent pattern across the empirical studies analyzed is the disproportionate emphasis on novice performance gains. Adaptive automation has largely been validated in contexts where short-term success is prioritized, typically measured via immediate task performance metrics such as accuracy, error reduction, or task completion time. While these outcomes are valuable, they provide only a narrow window into the learning lifecycle and do not reflect the complexity of sustained expertise development, which requires deeper cognitive transformation, mental model refinement, and strategic adaptation over time. This fragmentation stems from the lack of longitudinal training data and system persistence. Few studies followed learners beyond initial training sessions, and even fewer tracked developmental markers (e.g., strategic decision-making, flexibility under stress, pattern recognition) that distinguish experts from skilled performers (Ericsson et al., 1993). As a result, adaptive systems are rarely designed to address later-stage learning needs of experts such as 1) metacognitive and macrocognitive skill refinement, 2) recognition-primed decision-making, and 3) transfer to novel contexts or edge cases.

Hilburn (2016) proposed a two-stage view of adaptive automation: early automation should conform to novice needs, but later, the expert should take the lead, with the system offering only minimal support. However, most current systems plateau before this shift occurs or only suffice for early-stage novice's learning as most expertise is then developed during "on the job training". This stalling represents a systemic limitation in design thinking and evaluation

frameworks. Without long-term tracking and evolving instructional logic, the promise of adaptive automation to produce experts, not just competent users, is unrealized.

Gap 2: Fragmented Integration of Knowledge Elicitation into Adaptive Control Logic

A defining feature of next-generation human–AI training systems is their ability not only to adapt to the learner but to learn from the learner. Our analysis examined how training systems elicited human knowledge, whether through verbal reasoning, behavioral performance, or psychophysiological signals, and how that information was converted into machine-interpretable models to drive adaptive instruction. We found a range of elicitation strategies in use, though they remain inconsistently implemented and often underdeveloped. Broadly, these methods include: 1) capturing verbal reasoning (e.g., think-aloud protocols, debriefs), 2) recording performance data (actions, errors, response times), and 3) sensing cognitive or affective states (e.g., via heart rate, skin conductance, EEG). Most current systems adapt based on what the trainee does, with some beginning to consider how the trainee feels (e.g., stress or confidence levels). Few systems, however, capture why a trainee makes a particular decision, missing opportunities to close the learning loop, where the human learns from the AI's feedback, and the AI continuously learns from the human's behavior and reasoning. While performance-based adaptation dominates, our findings point to the untapped potential of explicit and multimodal elicitation including real-time reasoning capture, expert-to-AI modeling, and physiological sensing. Figure 2, below, summarizes the prevalence of knowledge elicitation strategies identified in the meta-analysis and their relative frequency of use across systems:

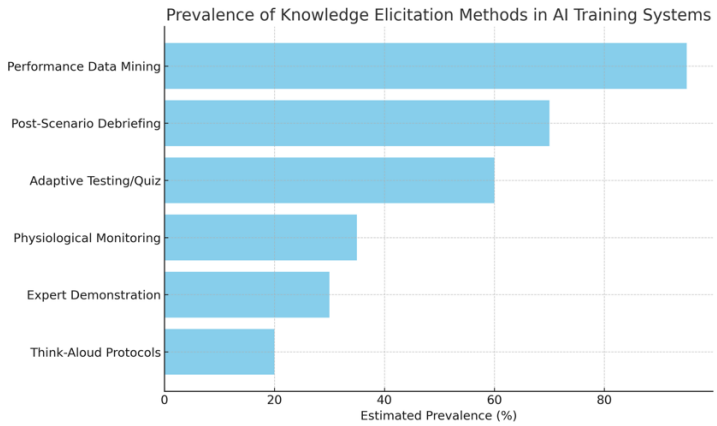

Figure 2. Prevalence of Knowledge Elicitation Methods in AI Training Systems

Despite growing interest in knowledge elicitation (KE) methods, ranging from think-aloud protocols and post-hoc interviews to physiological sensing, their integration into adaptive training systems remains fragmented, inconsistent, and primarily disconnected from real-time instructional logic. Rather than serving as dynamic inputs to guide instructional decisions moment-by-moment, KE techniques are often relegated to post-hoc analysis or external evaluation. In many systems, elicitation operates as a separate research tool rather than a core component of the adaptive engine itself. This disconnect limits systems' ability to interpret *why* learners succeed or fail, reducing adaptivity to a reaction to outcomes rather than a dialogue with underlying condition. For instance, Hoven et al. (2018) employed think-aloud protocols to better understand mental model development in medical trainees, revealing valuable insights into diagnostic reasoning and learner misconceptions. However, the adaptive system did not process this verbalized reasoning during training; instead, it remained a passive data source for researchers. In effect, the AI did not "learn" from the human in real-time.

The lack of integration is further compounded by the heterogeneity in how KE inputs are captured and used across systems. Some rely solely on post-scenario debriefs or instructor-facilitated reflections, which delay the use of feedback until after the learning event has concluded, rendering the feedback less actionable. Other systems, like that of Ruberto et al. (2021), incorporate real-time physiological signals (e.g., heart rate variability, skin conductance) to infer cognitive load, which are used to modulate scenario difficulty. However, even in these cases, the data streams are often interpreted in isolation, lacking fusion with contextual task data or learner intent. As a result, while

physiological indicators may suggest stress or overload, the system may be unable to interpret why the learner is struggling or how best to intervene based on their specific cognitive strategy.

Critically, very few systems have achieved real-time fusion of cognitive, behavioral, and affective cues into a coherent learner state model that can drive instructional decision-making. This limits the ability of adaptive systems to recognize subtleties in learner progression, particularly during transitional phases, such as the shift from intermediate competence to strategic expertise. During these phases, surface-level metrics (like correctness or response time) are insufficient for detecting deeper learning needs, such as the refinement of mental models, strategy selection, or recognition-primed decision making. As a result, current systems risk misaligning their support: over-assisting advanced learners, under-challenging high performers, or failing to address conceptual gaps masked by competent task execution.

Hollnagel's (2002) work on cognitive control and joint systems further underscores this issue. He argues that effective human-automation collaboration depends on the system's ability to model and respond to the intent and contextual rationale behind human actions, not just outcomes (Hollnagel, 2002). A system that cannot perceive or interpret user strategy is limited in its ability to adapt intelligently, especially in complex or ambiguous scenarios. Without real-time access to the learner's internal decision processes, adaptive systems cannot engage in meaningful mutual learning and instead operate as reactive rather than reflective agents. Moreover, the fragmented use of KE reflects a deeper theoretical divide: current implementations often assume that performance equates to learning. Yet, Seda et al. (2021) emphasized that verbal reports and think-aloud data reveal critical insights into a learner's evolving mental representation of the task, which cannot be inferred solely from task completion metrics. Thus, without incorporating these deeper indicators into system logic, adaptive training platforms remain blind to how learners conceptualize problems, how they adjust strategies, and how misunderstandings evolve over time.

To move beyond this fragmentation, future systems must treat knowledge elicitation not as an ancillary component, but as a central mechanism for adaptation: treating KE as both diagnostic and instructional. This involves real-time collection, multimodal signal fusion, and integration of elicited knowledge into the learner model and feedback engine creating a bidirectional learning loop between human and AI. KE should not merely explain what happened after training, but shape how training unfolds. Further closing the loop between learner cognition and system response. The failure to integrate rich, multimodal knowledge elicitation into adaptive logic limits the personalization, accuracy, and instructional intelligence of training systems. Without mechanisms to interpret how and why learners act, not just what they do, adaptive automation will remain surface-level, incapable of supporting deep, sustained learning. Addressing this gap is essential for creating systems that scaffold not only behavior, but cognition itself, especially as learners progress toward expertise.

Gap 3: Weak AI Knowledge Reuse Mechanisms

Even when adaptive training systems succeed in capturing data, such as performance metrics, instructor feedback, biometric indicators, or behavioral logs, this information is rarely structured in a way that facilitates knowledge reuse. Most systems are still built upon static rule sets or deterministic decision trees that do not evolve in response to the patterns embedded in accumulated learner interaction data. As a result, the AI does not "learn" from past experiences; instead, it treats each new learner as a blank slate, missing the opportunity to continuously improve its pedagogical intelligence. This absence of a persistent, reusable memory architecture severely restricts the developmental potential of both the learner and the system.

This design limitation runs counter to principles in both machine learning and expertise development. As Wang et al. (2024) demonstrated with their Tutor CoPilot system, it is feasible to train AI agents on large corpora of expert-human interactions, such as tutoring transcripts, to scaffold novice learners with context-sensitive, real-time guidance. The strength of Tutor CoPilot lies in its ability to internalize expert instructional strategies and apply them flexibly. However, its scope was limited to classroom tutoring environments, where instructional goals, learner behaviors, and task contexts are often simpler and more constrained than those found in simulation-based training for aviation, defense, or medical domains. In high-stakes domains and complex collaborative training environments, the challenge is amplified by the multimodal nature of expertise: expert knowledge is expressed not only through verbal instruction or task completion, but also through stress regulation, gaze behavior, attention allocation, and contextual decision-making under uncertainty. Capturing such nuanced behavior across time and users requires systems that can ingest and structure data streams from diverse modalities (i.e., speech, eye tracking, EEG, task progression logs) and convert

them into machine-interpretable formats. However, the field lacks standardized schemas, data models, or ontologies that can support such integration in a scalable way (Brusilovsky & Millán, 2007).

This limitation is particularly evident in the failure of many adaptive systems to support cross-session memory. Without mechanisms for tracking and updating learner states across multiple training episodes, systems cannot support longitudinal learning, personalize instruction over time, or detect developmental patterns that signal deeper shifts in competence. Moreover, without cross-learner generalization, systems are unable to benefit from insights gathered from one user's interaction to improve support for others with similar profiles or behaviors. In effect, current architectures prevent the kind of collective instructional intelligence that should characterize next-generation AI-enabled learning environments. Addressing this issue requires that future systems be capable of several interrelated functions. First, they must incrementally update learner models using continuous data inputs by refining their understanding of user knowledge, preferences, and performance as training progresses. Second, they must be able to generalize these insights across learners, recognizing patterns that predict common challenges, misconceptions, or mastery trajectories. Finally, systems must support cohort-level model evolution, where insights gained from one generation of users inform the adaptive policies for the next. This is the foundation of what Wang et al. (2024) termed intelligent co-learning, an iterative, data-driven model of system improvement through human interaction.

This goal also reflects principles from human expertise modeling, where expert knowledge is considered both situated and adaptive (Ericsson et al., 1993; Klein, 1998). Experts are not merely repositories of procedures but are capable of flexibly adjusting their behavior in dynamic contexts. To support the development of such expertise in learners, adaptive systems must emulate this flexibility, evolving their instructional strategies over time based on the data they collect. This necessitates architectures that include retrainable models, knowledge fusion pipelines, and feedback loops that connect user input with system evolution. Without these capabilities, adaptive training systems will remain trapped in a cycle of reactive adaptation, responding to short-term performance indicators without ever internalizing the deeper patterns that define expert learning. As a result, they will fall short of their promise to develop skilled, adaptable professionals in domains where cognitive agility, strategic thinking, and decision-making under pressure are mission critical. The field must transition from static, performance-driven adaptation to dynamic, knowledge-driven co-evolution. This shift is central to realizing the vision of adaptive human–AI systems that are not only personalized, but pedagogically intelligent, reflective, and capable of growing alongside the learners they support.

MUTUAL LEARNING LOOP FRAMEWORK

The Adaptive Expertise Mutual Learning Loop conceptual framework (see Table 1, below) directly addresses key limitations surfaced in the meta-analysis of adaptive training systems: (1) an overemphasis on early-stage performance, (2) fragmented integration of knowledge elicitation, and (3) a lack of mechanisms for pedagogical adaptation in AI. By drawing upon the complementary theories of Ericsson's (1993) deliberate practice and Hollnagel's (2005) Joint Cognitive Systems (JCS), this framework reorients adaptive human–AI simulation-based training toward long-term developmental processes. The following five layers structure the Mutual Learning Loop as a learning engineering-operational model, each designed to fulfill critical instructional functions tied to known research gaps.

Table 1. Adaptive Expertise Mutual Learning Loop Conceptual Framework

Framework Layer	Purpose	Addressed Gap
Layer 1: Expertise-Aligned Adaptive Engine	Dynamically adjusts scaffolding and task difficulty; supports hybrid autonomy based on learner model output.	**G1**: Overemphasis on early-stage performance; need for long-term expertise development.
Layer 2: Multimodal Knowledge Elicitation System	Integrates performance, biometric, and verbal data to inform adaptive decisions in real-time.	**G2**: Fragmented and static use of knowledge elicitation methods.
Layer 3: Learner Modeling and Knowledge Fusion	Synthesizes cognitive, metacognitive, and affective signals to generate unified learner profiles.	**G2**: Lack of real-time, high-resolution learner modeling across cognitive domains.

Layer 4: Mutual Learning Loop	Facilitates bi-directional learning between human and AI; evolves instructional strategies with data.	**G3**: Weak AI knowledge reuse and lack of long-term instructional evolution.
Layer 5: Transparency & Trust Interface	Delivers explainable AI feedback and rationale to support learner reflection and trust calibration.	**G1-G3**: Supports metacognitive and macrocognitive growth; responsible automation trust calibration.

Design Imperatives: Layered Architecture of the Mutual Learning Loop Conceptual Framework

Layer 1: Expertise-Aligned Adaptive Engine
The first layer implements a dynamically responsive instructional engine that personalizes scaffolding across the novice-to-expert continuum. Rather than relying on static instructional sequences, this engine uses real-time learner data to modulate task difficulty, fade prompts, and escalate cognitive stressors. As the learner gains fluency, system support decreases, aligning with Ericsson's model of adaptive expertise development through challenging, feedback-rich practice. Notably, the engine supports hybrid adaptation, offering both AI-led (adaptive) and learner-controlled (adaptable) adjustments to the training interface. This flexibility is critical for supporting autonomy and fostering metacognitive regulation. Addressing Gap 1, this layer ensures the system moves beyond early-stage metrics and provides developmental scaffolds that persist and adapt over time, nurturing expert-level cognitive performance. Intention of this layer is to be systematically linked to complex environments that utilize on the job training.

Layer 2: Multimodal Knowledge Elicitation System
The second layer serves as the system's sensory interface, capturing a breadth of learner data to drive adaptive responses. This includes behavioral performance metrics (e.g., task accuracy, response latency), physiological indicators (e.g., heart rate variability, eye tracking), and verbal or reflective inputs (e.g., think-alouds, post-task rationales). Unlike traditional adaptive systems that rely solely on performance data, this layer enables real-time fusion of cognitive-affective signals, grounding the system's decision-making in a holistic understanding of the learner. By continuously collecting and routing these inputs to the learner model, the system transitions from a reactive tutor to a context-sensitive learning partner. This directly addresses Gap 2, overcoming the fragmented and retrospective application of knowledge elicitation in prior systems by enabling proactive and interpretive adaptation.

Layer 3: Learner Modeling and Knowledge Fusion
Layer three synthesizes the incoming multimodal data into a unified, high-resolution learner model. It maps the learner's state across four interrelated domains: cognitive (task mastery and decision-making logic), metacognitive (self-awareness, confidence, error monitoring), macrocognitive (adapting to complex, dynamic, ambiguous real-world situations), and affective (stress, motivation). This comprehensive learner profile informs instructional logic, enabling the AI to tailor scaffolding, feedback type, and challenge level in real time based on levels of expertise. The model not only identifies what the learner knows, but also how they think and feel during task performance such as insights that are essential for fostering adaptive expertise. In this way, the layer operationalizes Hollnagel's emphasis on designing systems that are sensitive to human variability, promoting a Joint Cognitive System where machine behaviors evolve in alignment with human cognitive processes (2002). This layer further solidifies the response to Gap 2 by establishing a robust and interpretable learner modeling architecture centralized on expertise.

Layer 4: Mutual Learning Loop
At the heart of the framework lies the Mutual Learning Loop itself, an engine of bi-directional co-adaptation between human and machine. As the learner progresses, the AI continuously adjusts its pedagogical strategy by analyzing patterns in learner performance, strategy use, and response to feedback. Conversely, the human receives customized instructional support that evolves in granularity and intensity based on their development. This reciprocity aligns with Ericsson's view that expertise emerges from iterative, feedback-based practice and with Hollnagel's argument that automation must be designed for human–machine collaboration as opposed to substitution. Importantly, this layer enables persistence: the AI can learn from one session to the next, generalize across learners, and reuse instructional insights across scenarios. By incorporating long-term data structures and retrainable instructional logic, Layer 4 addresses Gap 3, transforming the AI from a static rule-executor into an evolving pedagogical agent.

Layer 5: Transparency & Trust Interface

The final layer ensures the system remains interpretable, ethical, and learner-centered through transparency and explainability. This interface delivers real-time rationale for AI interventions in language that is understandable and actionable. For example, learners may be told, "I increased difficulty because you demonstrated consistent accuracy under pressure," or "You can now disable AI hints, your decision-making is independently sound." These insights support metacognitive and macrocognitive development, help calibrate trust in automation, and prevent over-reliance or premature rejection of system support. Grounded in Hollnagel's (2002) emphasis on human–system resilience and trust calibration, this layer strengthens user engagement and accountability. While not tied to a singular research gap, it is essential for operationalizing all others, ensuring that the system's intelligent behaviors are transparent and aligned with learner needs and goals.

SCALABILITY and USE CASE

Our research team is currently advancing the prototyping of Layer 3: Learner Modeling and Knowledge Fusion of the AEMLL framework within a high-stakes maritime skilled trades training context. This layer, which constructs unified learner profiles across cognitive, metacognitive, macrocognitive, and affective domains, is the foundation upon which intelligent human–AI training systems must be built. Drawing directly from our recent meta-analysis on adaptive automation and simulation-based learning environments and grounded in theoretical insights from Ericsson's deliberate practice framework and Hollnagel's Joint Cognitive Systems model (2002), we determined that beginning with Layer 3 was both strategic and necessary.

The rationale for prioritizing Layer 3 is clear: without a robust and interpretable learner model, downstream adaptive features, such as scaffolded feedback, autonomy calibration, and instructional evolution, cannot function effectively. Our meta-analysis revealed that most existing systems are over-reliant on surface-level performance metrics and lack high-resolution insights into how a learner reasons, adapts, and responds under stress. Addressing this, our maritime use case targets learners enrolled in a multi-year welding apprenticeship pipeline. By collecting multimodal data (eye-tracking to capture visual attention patterns, behavioral logs to analyze strategy use and errors, and think-aloud protocols to surface decision-making rationales) we are creating detailed learner profiles that reflect not just what learners do, but how they think and feel during problem-solving scenarios as expertise is developed. This comprehensive profiling supports real-time and post-training personalization, enabling instructors and future AI agents to better tailor guidance to where learners are on their trajectory toward adaptive expertise. Importantly, our design operationalizes Gap 2 of the framework, fragmented and static use of knowledge elicitation, by fusing multiple streams of data to infer learner state dynamically and contextually.

Operationalizing Layer 3 also aligns with AEMLL's aims to be a plug-and-play architecture. By building the core learner modeling infrastructure first, we create a data and logic backbone that subsequent layers can easily integrate with. For example, Layer 1 (Expertise-Aligned Adaptive Engine) can later plug in to use real-time outputs from Layer 3 to dynamically adjust feedback and challenge levels. Similarly, Layer 2 (Multimodal Knowledge Elicitation System) is already functionally linked to our prototyping by serving as the input mechanism feeding the learner model. In time, Layer 4 (Mutual Learning Loop) and Layer 5 (Transparency & Trust Interface) will be implemented to evolve AI instructional strategies and support human–AI interpretability and trust calibration, respectively. Ultimately, this work exemplifies how the AEMLL can be operationalized in decentralized, modular research environments. Importantly, the prototyping of Layer 3 also directly supports DoD and training modernization priorities, including adaptive learning ecosystems, distributed simulation-based readiness, and human–machine teaming readiness initiatives outlined in the National Defense Strategy and the DoD AI Strategy. By building a scalable infrastructure for context-aware learner profiling, where training gaps remain persistent, this work addresses strategic needs for developing resilient, adaptable personnel capable of operating alongside AI systems in real-world missions. Our team's focused development of Layer 3 establishes the conditions for intelligent, context-aware expertise adaptation to flourish infrastructures that enable the remaining layers to be activated iteratively and cohesively over time. This phased, integrative approach ensures that expertise development in high-consequence domains can be both scientifically rigorous and practically scalable.

LIMITATIONS

Meta-Analysis Limitations

While this meta-analysis offers meaningful insights into the effectiveness of adaptive training systems and their role in fostering expertise development, several limitations should be acknowledged that may affect the generalizability and precision of the findings. First, the heterogeneity of study designs and evaluation methodologies introduces variability into the aggregated effect sizes. Studies differed significantly in their sample sizes, duration of training interventions, domain-specific tasks (e.g., aviation, medicine, military simulation), and operational definitions of "expertise." Some studies evaluated immediate performance outcomes (e.g., accuracy or time-on-task), while others considered behavioral measures or subjective self-assessments. This variation complicates direct comparisons and introduces potential noise into meta-analytic estimates, particularly when interpreting intervention subtypes such as scaffolding, adaptive difficulty, or remediation. Second, the sample sizes within individual studies were often modest, with many experiments relying on fewer than 30 participants per condition. Such small sample sizes can inflate effect sizes and reduce statistical power, increasing the likelihood of both Type I and Type II errors (Cohen, 1992). As a result, while the aggregated data trends provide valuable directional insight, caution must be taken in interpreting specific numerical values of mean effect sizes as definitive. Third, the diversity in data collection techniques and metrics limits cross-study consistency. Some systems used biometric inputs such as GSR or eye tracking (e.g., Ruberto et al., 2021), while others relied solely on keystroke logs, quizzes, or post-session debriefs. This lack of standardization in data types and analytic methods creates difficulties in drawing precise conclusions about the relative effectiveness of different adaptive mechanisms across domains. Fourth, many of the studies reviewed focused on short-term training outcomes, with few extending their data collection over time to examine long-term retention, transfer, or sustained expertise development. As a result, the meta-analysis captures snapshots of learner performance rather than comprehensive learning trajectories. This shortcoming is particularly critical when considering systems designed to support progressive development from novice to expert over months or years.

Finally, there remains a publication bias risk in the literature, where studies demonstrating significant or positive effects are more likely to be published. As with many meta-analyses, this may result in an overestimation of the true effectiveness of adaptive training interventions if null or negative results are underreported. These limitations highlight the need for future work that adopts more longitudinal, multi-modal, and standardized approaches to evaluating adaptive systems in training contexts. Expanding datasets, increasing methodological transparency, and aligning studies around common outcome metrics will be essential to deepen our understanding of how human–AI systems can reliably support expertise development at scale.

Adaptive Expertise Mutual Learning Loop Framework Limitations

While the Mutual Learning Loop presents a theoretical and learning engineering-operational model for advancing adaptive expertise development within AI training systems, it remains conceptual in nature and has several limitations that must be acknowledged. Chief among these is its reliance on a longitudinal developmental view of learning that is not yet well-supported by empirical data across domains. Much of the existing research on adaptive systems, both in academic literature and operational settings, focuses on short-term gains in performance, typically within the early stages of novice training. Although the framework is grounded in Ericsson et al.'s theory of deliberate practice (1993) and Hollnagel's cognitive systems engineering principles (2005), the empirical infrastructure required to fully validate long-term, co-evolutionary training mechanisms is still underdeveloped.

One major challenge is the need for longitudinal research that can track learner progression from novice to expert across extended timelines. Capturing meaningful indicators of expertise development, such as the evolution of mental models, strategic flexibility, and transfer to novel tasks, requires studies that span years. However, such work is methodologically and logistically demanding. Defining "expertise" itself is often domain-dependent and context-specific, complicating comparisons across fields like aviation, medicine, cybersecurity, and defense. What counts as expertise in one domain may not translate neatly into another, making it difficult to design a unified evaluation framework for adaptive training systems. Additionally, cognitive modeling of dynamic decision making, a key feature of expertise, remains a work in progress. While the Adaptive Expertise Mutual Learning Loop proposes real-time learner modeling that fuses behavioral, cognitive, and affective data, implementing such a model requires large-scale, high-resolution datasets that are often difficult to collect, especially in high-stakes environments. Capturing multimodal signals such as physiological stress indicators, verbal reasoning traces, and attentional shifts demands not only technical integration but also theoretical coherence in how these signals are interpreted and used by the system. Moreover, the data requirements for such a system are immense. Training effective AI-driven adaptation engines, particularly those capable of longitudinal personalization and instructional reuse, necessitates massive quantities of labeled training data. These data sets must be representative, domain-specific, and diverse enough to support

generalization across individual learners and tasks. Without large-scale deployments and data-sharing mechanisms, it is difficult to validate the framework's assumptions about mutual adaptation and co-evolution in practice.

Finally, the framework presumes a level of system persistence, infrastructure maturity, and institutional support that may not be present in many training contexts currently. In practice, training systems are often fragmented across learning management platforms, operational units, and assessment regimes, making the seamless integration of all five layers a considerable engineering challenge. Thus, while the AEMLL provides a compelling roadmap for future training systems, it must be tested and refined through iterative prototyping, pilot studies, cross-domain collaborations, and a sustained investment in longitudinal learning science research. The framework offers a promising conceptual advance, but one that requires significant empirical, technical, and institutional support to be realized at scale. Further research is needed to operationalize its mechanisms, evaluate its claims across contexts, and generate the longitudinal evidence base necessary to validate its developmental impact on expertise formation.

DISCUSSION

As simulation-based training systems evolve alongside AI affordances, the imperative to embed adaptive formation of expertise into design is rapidly intensifying. The increasing prevalence of human–AI teaming in operational environments, ranging from autonomous navigation and remote medicine to tactical defense operations, demands training systems that mirror these hybrid contexts. Traditional one-size-fits-all or static training platforms no longer suffice. Instead, learners must be prepared to operate as part of dynamic human–machine teams where adaptability, strategic reasoning, and mutual coordination are essential. AEMLL anticipates this future by placing co-adaptation and bidirectional learning at the core of its architecture. Moreover, the rapid acceleration of software capabilities, particularly in areas such as natural language processing, emotion detection, and real-time cognitive state modeling, opens new opportunities for embedding intelligent adaptivity into training systems at scale. Machine learning infrastructures are becoming more accessible, sensor integration more seamless, and cloud-based simulation platforms more powerful. These technological advances create the conditions for operationalizing each layer of the AEMLL in real-world training contexts, if design remains grounded in cognitive science and supported by empirical validation.

Looking ahead, the next steps in research for adaptive expertise in simulation-based training systems should include the following priorities:

1. Layer-Specific Experimental Validation: Conduct controlled studies testing the effectiveness of individual layers, such as real-time expert learner modeling or transparency-enhancing interfaces, across various training scenarios and domains.
2. Longitudinal Implementation Studies: Establish long-term research partnerships to track how learners develop expertise over time within AEMLL-enabled environments, particularly focusing on transfer, retention, and mental model refinement.
3. Data Infrastructure Architectures: Create shared, interoperable data architectures and standards that allow for persistent learner models of expertise, multi-modal data fusion, and secure cross-session AI retraining.

Ultimately, the Adaptive Expertise Mutual Learning Loop is not a fixed solution but a conceptual blueprint for an evolving ecosystem of intelligent training systems specifically targeting advancing expertise. Its layered design allows for targeted research and prototyping today while enabling integration and scale tomorrow. As human–AI collaboration becomes the norm across complex domains, simulation-based training systems must rise to meet the challenge not simply by teaching tasks but by cultivating the adaptive expertise necessary to thrive in uncertain, high-stakes environments. The AEMLL framework provides a foundational model for that future.

CONCLUSION

This meta-analysis demonstrates that adaptive training systems hold significant promise for advancing expertise development, but current implementations remain fragmented and overly focused on short-term novice performance. By synthesizing evidence across two decades of studies, this work clarifies how adaptive human–AI training can be structured to support longitudinal expertise growth, embed multimodal knowledge elicitation, and enable mutual learning between human and machine. For the I/ITSEC community, these findings highlight the urgent need to move beyond "one-size-fits-all" adaptive systems toward scalable, expertise-driven training architectures aligned with defense modernization priorities. The Adaptive Expertise Mutual Learning Loop framework provides a roadmap for simulation designers, researchers, and practitioners to embed co-adaptation, transparency, and long-

term learner modeling into next-generation training systems. Its layered design anticipates the demands of distributed training, AI-enabled readiness, and resilient human–machine teaming. As the community advances toward Industry 5.0 principles, this research underscores the relevance of aligning adaptive training design with cognitive science to prepare warfighters, engineers, and operators not only for competence, but for true adaptive expertise in complex, uncertain environments.

REFERENCES

Aricò, P., Borghini, G., Di Flumeri, G., Colosimo, A., Bonelli, S., Golfetti, A., Pozzi, S., Imbert, J.P., Granger, G., Benhacene, R. & Babiloni, F. (2016). Adaptive Automation Triggered by EEG-based Mental Workload Index: A Passive Brain-Computer Interface Application in Realistic Air Traffic Control Environment. *Frontiers in Human Neuroscience, 10*, 539.

Bernabei, M., & Costantino, F. (2024). Adaptive Automation: Status of Research and Future Challenges. *Robotics and Computer-Integrated Manufacturing, 88*(3), 102724.

Billings, D. R. (2010). Adaptive Feedback in Simulation-Based Training (Doctoral dissertation). University of Central Florida repository publicly accessible.

Braun, V., & Clarke, V. (2006). Using Thematic Analysis in Psychology. *Qualitative Research in Psychology, 3*(2), 77-101.

Bruder, C., & Hasse, C. (2019). Differences Between Experts and Novices in the Monitoring of Automated Systems. *International Journal of Industrial Ergonomics, 72*, 1–11.

Brusilovsky, P., & Millán, E. (2007). User Models for Adaptive Hypermedia and Adaptive Educational Systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The Adaptive Web* (pp. 3–53). Springer.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The Role of Deliberate Practice in the Acquisition of Expert Performance. *Psychological Review, 100*(3), 363–406.

Fraulini, N. W., Marraffino, M. D., Garibaldi, A. E., Johnson, C. I., & Whitmer, D. E. (2024). Adaptive Training Instructional Interventions: A Meta-Analysis. Military Psychology.

Hilburn, B. (2016). A Hybrid Approach to Training Expert Skills in Highly Automated Systems: Lessons from Air Traffic Management. *IFAC-PapersOnLine, 49*(19), 207–211.

Hoc, J.M., Cacciabue, P.C., & Hollnagel, E. (1995). *Expertise and Technology: Cognition & Human-Computer Cooperation*. Psychology Press.

Hollnagel, E. & Woods, D. D. (2005). *Joint cognitive systems: Foundations of Cognitive Systems Engineering*. CRC Press/Taylor & Francis.

Hoven, S., Seniuk, A., & Martel, M. (2018). Adaptive Visual-Diagnostic Training: User Mental Model Development. In *Proceedings of Graphics Interface 2018* (pp. 44–45). ACM Press.

Kaber, D. B., et al. (2006). Situation Awareness Implications of Adaptive Automation for Information Processing in an Air Traffic Control-Related Task. *Intl. Journal of Industrial Ergonomics, 36*(5), 447–462.

Klein, G. (1998). *Sources of Power: How People Make Decisions*. MIT Press.

Ramirez, A. G., Hu, Y., Kim, H., & Rasmussen, S. K. (2018). Long-Term Skills Retention Following a Randomized Prospective Trial on Adaptive Procedural Training. *Journal of Surgical Education, 75*(6), 1589–1597.

Ruberto, A. J., Rodenburg, D., Ross, K., Sarkar, P., Hungler, P. C., Etemad, A., Howes, D., Clarke, D., McLellan, J., Wilson, D., Szulewski, A. (2021). Adaptive Simulation Utilizing a Deep Multitask Neural Network. *AEM Education and Training, 5*(3)

Sadler, G., et al. (2016). Effects of Transparency on Pilot Trust and Agreement in the Autonomous Constrained Flight Planner. In Proc. *IEEE DASC 2016*.

Seda, P., Vykopal, J., Švábenský, V., & Čeleda, P. (2021). Reinforcing Cybersecurity Hands-On Training with Adaptive Learning. In *Proceedings of the 51st IEEE Frontiers in Education Conference (FIE 2021*). IEEE.

Stanney, K. M., Archer, J., Skinner, A., Horner, C., Hughes, C., Brawand, N. P., Martin, E., Sanchez, S., Moralez, L., Fidopiastis, C. M., Perez, R. S. (2021). Performance Gains from Adaptive eXtended Reality Training Fueled by Artificial Intelligence. *Journal of Defense Modeling and Simulation Applications Methodology Technology, 19*(2).

Sarkar, P., Ross, K., Ruberto, A. J., Rodenburg, D., Hungler, P., & Etemad, A. (2019). Classification of Cognitive Load and Expertise for Adaptive Simulation Using Deep Multitask Learning. In *Proceedings of 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*.

Verniani, A., Galvin, E., Tredinnick, S., Putman, E., Vance, E. A., & Anderson, A. P. (2024). Features of Adaptive Training Algorithms for Improved Complex Skill Acquisition. *Frontiers in Virtual Reality, 5*, Article 1322656.

Vogl, J. T., D'Alessandro, M., Wilkins, J., Ranes, B., Persson, I., McCurry, C. D., & Bommer, S. (2024). Optimizing Adaptive Automation in Aviation: A Literature Review on Dynamic Automation System Interaction. USAARL Technical Report 2025-09.

Wang, R.E., Ribeiro, A.T., Robinson, C.D., Loeb, S., & Demszky, D. (2024). Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise. *arXiv preprint arXiv:2410.03017*. https://arxiv.org/abs/2410.03017.