Learning to Lead Under Stress: Designing AI Guidance for In-Hospital Resuscitation Team Leader Training in XR

Myungjun Lee[1,*], Hyeongil Nam[1], Jazeb Zafar[1], Jennifer Davidson[1], Yiqun Lin[1], Adam Cheng[1] and Kangsoo Kim[1]

[1]*University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada*

Abstract

AI-assisted Extended Reality (XR) training is increasingly adopted in mission-critical domains, such as medical education; however, prevailing evaluations rely on task performance metrics—error rates, procedural accuracy, and completion times—that capture training *products* rather than the *processes* through which durable competence develops. We argue that genuine learning in high-stress XR training cannot be fully assessed through performance outcomes alone; it requires considering factors such as trust calibration, resilience, and explainability as interdependent *enabling conditions*. Meaningful learning in this context involves trainees appropriately calibrating their reliance on AI guidance, performing independently when that guidance is unavailable, and understanding *why* an action is appropriate (not merely *what* to do). Grounded in an ongoing XR-based resuscitation team leader training project, this position paper presents this reframing and outlines its implications for the evaluation of AI-assisted XR training systems.

Keywords

extended reality, AI guidance, resuscitation team leadership, medical education, trust, perception

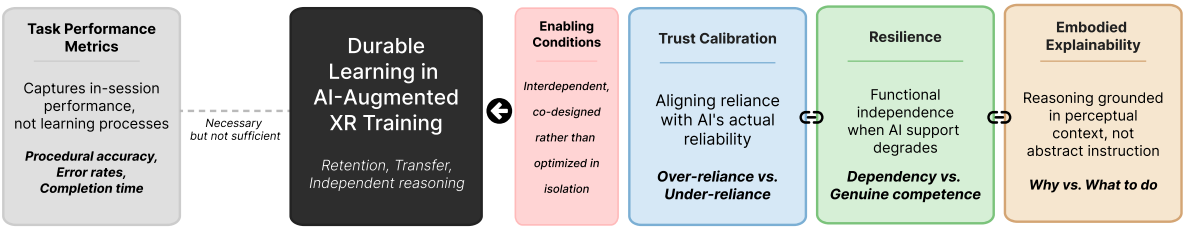

Figure 1: Conceptual framework of enabling conditions for durable learning in AI-assisted XR training.

1. Introduction

In-hospital cardiac arrest requires rapid and coordinated decision-making under extreme time pressure. During resuscitation, the team leader must prioritize interventions, interpret evolving clinical signals, and maintain situational awareness while managing a high workload [1]. Effective leadership extends beyond procedural accuracy to include decision ownership, coordination, and confidence under stress [2]. Training clinicians for this role remains a persistent challenge in medical education.

Extended Reality (XR) simulation is increasingly used to recreate the perceptual and temporal constraints of resuscitation scenarios [3, 4]. With the integration of artificial intelligence (AI) guidance, these systems can structure attention, reduce cognitive burden, and improve immediate task performance [5]. Compared to rule-based or static instructional systems, AI-driven guidance is adaptive and context-sensitive, producing variable outputs rather than fixed instructions. As a result, trainees must

CHI'26: XR4CE workshop, April 14, 2026, Barcelona, Spain
mj.lee1@ucalgary.ca (M. Lee); hyeongil.nam@ucalgary.ca (H. Nam); jazeb.zafar@ucalgary.ca (J. Zafar); jennifer.davidson@ahs.ca (J. Davidson); yiqlin@ucalgary.ca (Y. Lin); chenger@me.com (A. Cheng); kangsoo.kim@ucalgary.ca (K. Kim)
0009-0004-7463-9162 (M. Lee); 0000-0002-6017-8869 (H. Nam); 0009-0001-4076-4932 (J. Zafar); 0009-0009-9415-0598 (J. Davidson); 0000-0002-4726-5992 (Y. Lin); 0000-0002-1039-7502 (A. Cheng); 0000-0002-0925-378X (K. Kim)

interpret and evaluate guidance rather than simply following prescribed steps. By shaping attention and decision-making through selective cues, AI guidance may improve immediate performance but reduce opportunities for independent reasoning. These characteristics make trust calibration, resilience, and explainability necessary conditions for effective learning under AI mediation. Although the enabling conditions discussed in this paper may apply to automated guidance more broadly, they become especially critical in AI-assisted training, where guidance is adaptive and context-sensitive. However, this introduces a structural tension: guidance that optimizes in-session performance may simultaneously reduce opportunities for independent reasoning. In leadership training, where decision ownership is itself a learning objective, this tension becomes central.

Current evaluations of AI-assisted XR training primarily emphasize task performance metrics, such as procedural accuracy, error rates, and completion time [6, 3]. Although necessary, these metrics capture training products rather than the processes through which durable competence develops. In learning science, improvements in short-term performance do not necessarily indicate long-term retention, transfer, or independent decision-making capacity under stress [7, 8]. For leadership training, the question is therefore not only whether trainees perform correctly with AI support, but whether they develop the ability to act appropriately when that support is absent or unreliable.

In mission-critical contexts, such as in-hospital resuscitation, three dimensions become particularly salient: (1) trust calibration, which shapes epistemic alignment with AI guidance [9]; (2) resilience, which preserves functional independence under system degradation [10]; and (3) embodied explainability, which grounds reasoning in a perceptual context rather than abstract instructions [11, 12]. Together, these dimensions describe the conditions under which trainees may internalize reasoning, retain agency, and transfer skills to real-world practice.

In this position paper, grounded in an ongoing XR-based resuscitation team leader training project, we ask the following question: *"In XR-based resuscitation team leader training, how should AI guidance be designed to support calibrated trust while preserving independent clinical judgment and leadership agency under stress?"* We argue that evaluating AI-assisted XR training requires reframing commonly measured constructs—not as standalone endpoints, but as enabling conditions for durable learning. While these enabling conditions are grounded in training contexts, they have direct implications for how XR systems are used in real-world operational settings.

In mission-critical domains, such as resuscitation, practitioners rely on patterns of judgment, attention, and reliance that are shaped during training. In this sense, training and operational XR augmentation are sequentially coupled. If AI guidance during training induces overreliance or suppresses independent reasoning, these patterns may carry over to real-world scenarios where XR systems provide real-time support under uncertainty. Training design, therefore, functions as a prerequisite for operational augmentation and conditions how practitioners interpret, trust, and act upon XR-mediated guidance in practice.

2. Background: Evaluation Patterns in AI-Assisted XR Training

As AI-assisted XR training expands in high-stakes domains, especially in medical education, including resuscitation and emergency care contexts [13, 14], systematic and umbrella reviews have evaluated its effectiveness primarily through measurable outcomes, such as procedural accuracy, completion time, knowledge gain, confidence, and satisfaction [6, 3, 4]. In resuscitation training, reported metrics commonly include cardiopulmonary resuscitation (CPR) depth and rate compliance, protocol adherence, and immediate performance scores, with comparatively limited attention to longitudinal retention or transfer to practice [3]. These measures provide evidence of feasibility and skill acquisition, but primarily capture in-session performance.

A complementary body of work examines user perceptions and trust in AI-guided systems. Studies on embodiment and virtual agents demonstrate that visual form and social behavior influence perceived trustworthiness, confidence, and social presence [11, 12]. In parallel, research on trust in automation emphasizes calibrated reliance, warning that both over-reliance and under-reliance can degrade decision

quality [9, 10]. In many cases, however, trust is treated as an evaluative outcome variable rather than a component of a broader learning process.

Across these strands of work, performance metrics, user perception, and trust are often examined as distinct evaluative constructs. However, learning science distinguishes between short-term performance gains and durable learning, including retention, transfer, and independent reasoning under variable conditions [8, 7]. Improvements observed under continuous guidance do not necessarily indicate that learners can perform independently when support is removed. For mission-critical leadership training, where decision ownership and adaptive judgment are core competencies, this distinction becomes especially consequential.

These patterns motivate reframing: rather than treating performance, trust, and perception as isolated endpoints, AI-assisted XR training may need to be evaluated in terms of the conditions that enable durable learning under stress.

3. Reframing AI-Assisted XR Training: Enabling Conditions for Durable Learning

To resolve the tension between AI-supported performance and independent competence, we propose reframing evaluation around *enabling conditions* that shape how learning unfolds under guidance. In mission-critical XR training, three dimensions have become structurally central. First, trainees must maintain appropriate epistemic alignment with AI outputs (*trust calibration*). Second, they must retain functional independence when AI support is degraded or unavailable (*resilience*). Third, they must ground their reasoning in perceptual context rather than rely solely on abstract instructions (*embodied explainability*). These dimensions correspond to the cognitive, structural, and perceptual aspects of learning under AI mediation. Together, they characterize the conditions that support durable learning.

3.1. Trust Calibration: Epistemic Alignment

Trust calibration refers to the alignment between a trainee's reliance on AI guidance and the system's actual reliability [9, 10]. Over-reliance may lead to the uncritical acceptance of incorrect recommendations, whereas under-reliance can increase monitoring demands and cognitive load. Through the lens of Signal Detection Theory (SDT), trainees must distinguish genuine AI errors from valid guidance while maintaining attention to concurrent task demands. In this framing, trust calibration is not an outcome to maximize, but rather a condition that preserves active reasoning and decision ownership during training.

3.2. Resilience: Functional Independence

Resilience concerns the trainee's ability to act effectively when AI support degrades. In mission-critical settings, systems inevitably encounter noise, uncertainty, or failure. If training occurs only under continuous AI scaffolding, improvements in performance may reflect dependency rather than competence. Designing for resilience involves intentionally creating moments in which guidance is reduced or withdrawn, requiring trainees to exercise independent judgment. In this sense, resilience is both a system property and a learning condition that supports the transfer to real-world practice.

3.3. Embodied Explainability: Perceptual Grounding

Effective training requires understanding why an action is appropriate, not only what to do. Traditional transparency mechanisms, such as text overlays or alerts, may increase cognitive load under stress. Embodied explainability instead leverages spatial cues, gaze direction, gestural guidance [15], and environmental feedback to scaffold reasoning through perceptual channels [11, 12]. By situating guidance within the trainee's embodied interaction with the environment, explainability supports reasoning that more closely resembles real clinical practice.

3.4. Interdependence of Enabling Conditions

These three conditions are interdependent. Calibrated trust without resilience may allow trainees to detect AI errors but leave them unable to proceed independently. Resilience without perceptual grounding may produce procedural compliance without reasoning. Explainability without appropriate trust may lead to confusion regarding when to rely on guidance. Evaluating AI-assisted XR training therefore involves examining how these dimensions jointly support retention, transfer, and independent decision-making under stress.

Figure 1 illustrates this framework, depicting trust calibration, resilience, and embodied explainability as interdependent dimensions that jointly enable durable learning under AI mediation.

4. Illustrative Case: Empirical Entry Point

As an empirical entry point, we are conducting an ongoing XR-based pediatric resuscitation team leader study in collaboration with the Alberta Children's Hospital. The study compares three visual agent modalities—text-based prompts, an abstract visual agent, and a human-like embodied agent—while holding vocal guidance constant. By examining how agent representation shapes perceived authority, trust, workload, and learning experience, this work provides an empirical entry point for investigating how AI guidance influences reliance and explainability in high-stress contexts. Although the study does not directly operationalize all the enabling conditions proposed in this paper, it allows us to examine how design choices in representation may shape trust calibration and embodied explainability, informing future investigations of resilience and independent performance when guidance is degraded or withdrawn.

5. Discussion and Conclusion

This study reframes the evaluation of AI-assisted XR training from isolated performance and perception metrics toward enabling conditions for durable learning under stress. In mission-critical leadership contexts, trust calibration, resilience, and embodied explainability are not independent objectives but interdependent dimensions that shape whether trainees retain agency and transfer competence beyond guided interactions.

This reframing also suggests concrete design implications. First, the visual representation of the guiding agent may serve as a lever for trust calibration. In our ongoing study, we compare multiple visual agent conditions to examine how the agent's visual form shapes trust, acceptance, and preference.

Second, resilience may be supported through the guided withdrawal of AI scaffolding. A system could initially provide full guidance and then progressively reduce support as the trainee demonstrates competence, and reintroduce it when performance declines. This process, informed by real-time indicators such as action timeliness or protocol adherence, allows the evaluation of whether trainees maintain functional independence under reduced support.

Third, embodied explainability may be supported through the agent's spatial and social behavior (e.g., directing the gaze toward a relevant monitor or gesturing toward medication) rather than relying on text overlays that can increase cognitive load under stress. Such spatial and social behaviors have been shown to increase confidence and social presence in AR agent interactions [11].

This perspective carries several implications for the XR for Challenging Environments workshop community. First, evaluation should extend beyond in-session accuracy, satisfaction, or trust scores to examine retention, transfer, and independent decision-making under degraded or absent support. Second, these enabling conditions must be co-designed rather than optimized in isolation: calibrated trust without resilience risks dependency, resilience without perceptual grounding risks procedural compliance without reasoning, and explainability without appropriate trust may destabilize reliance. Finally, because high-stress environments vary substantially across domains, these conditions require domain-specific validation and adaptation.

As AI-augmented XR training continues to expand across challenging environments, designing and evaluating systems through the lens of enabling conditions may better support the development of professionals who can act appropriately when guidance is incomplete, uncertain, or unavailable.

Acknowledgments

This work was supported by a Transdisciplinary Scholarship Connector Grant from the University of Calgary.

Declaration on Generative AI

During the preparation of this work, the author(s) used Anthropic Claude in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] L. L. Brown, Y. Lin, N. M. Tofil, F. Overly, J. P. Duff, F. Bhanji, V. M. Nadkarni, E. A. Hunt, A. Bragg, D. Kessler, I. Bank, A. Cheng, Impact of a cpr feedback device on healthcare provider workload during simulated pediatric resuscitation, Resuscitation 130 (2018) 111–117. doi:10.1016/j.resuscitation.2018.06.035.

[2] S. Hunziker, A. C. Johansson, F. Tschan, N. K. Semmer, L. Rock, M. D. Howell, S. Marsch, Teamwork and leadership in cardiopulmonary resuscitation, Journal of the American College of Cardiology 57 (2011) 2381–2388. doi:10.1016/j.jacc.2011.03.017.

[3] A. Cheng, et al., Use of augmented and virtual reality in resuscitation training: A systematic review, Resuscitation Plus 18 (2024) 100643. doi:10.1016/j.resplu.2024.100643.

[4] T. Tene, D. F. Vique López, P. E. Valverde Aguirre, L. M. Orna Puente, C. Vacacela Gomez, Virtual reality and augmented reality in medical education: an umbrella review, Frontiers in Digital Health 6 (2024). URL: https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2024.1365345. doi:10.3389/fdgth.2024.1365345.

[5] D. Dasa, M. Board, U. Rolfe, T. Dolby, W. Tang, Evaluating AI-driven characters in extended reality (XR) healthcare simulations: A systematic review, Artificial Intelligence in Medicine 170 (2025) 103270. URL: https://www.sciencedirect.com/science/article/pii/S0933365725002052. doi:10.1016/j.artmed.2025.103270.

[6] Y. Baashar, et al., Effectiveness of using augmented reality for training in the medical professions: Meta-analysis, JMIR Serious Games 10 (2022) e32715. doi:10.2196/32715.

[7] N. C. Soderstrom, R. A. Bjork, Learning versus performance: an integrative review., Perspectives on Psychological Science 10 (2015) 176–99. doi:10.1177/1745691615569000.

[8] E. L. Bjork, R. A. Bjork, Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning, in: M. A. Gernsbacher, R. W. Pew, L. M. Hough, J. R. Pomerantz (Eds.), Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society, Worth Publishers, 2011, pp. 56–64.

[9] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human Factors 46 (2004) 50–80. doi:10.1518/hfes.46.1.50_30392.

[10] R. Parasuraman, D. H. Manzey, Complacency and bias in human use of automation: An attentional integration, Human Factors 52 (2010) 381–410. doi:10.1177/0018720810376055.

[11] K. Kim, L. Boelling, S. Haesler, J. Bailenson, G. Bruder, G. F. Welch, Does a digital assistant need a body? the influence of visual embodiment and social behavior on the perception of intelligent virtual agents in ar, in: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR), IEEE, 2018. doi:10.1109/ISMAR.2018.00039.

[12] I. Wang, J. Smith, J. Ruiz, Exploring virtual agents for augmented reality, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, ACM, New York, NY, USA, 2019. doi:10.1145/3290605.3300511.

[13] A. I. Iqbal, A. Aamir, A. Hammad, H. Hafsa, A. Basit, M. O. Oduoye, M. W. Anis, S. Ahmed, M. I. Younus, S. Jabeen, Immersive Technologies in Healthcare: An In-Depth Exploration of Virtual Reality, Journal of Primary Care & Community Health 15 (2024) 21501319241293311. doi:10.1177/21501319241293311.

[14] R. Harari, S. Schulwolf, P. Borges, H. Salmani, F. Hosseini, S. Bailey, B. Quach, E. Nohelty, S. Park, Y. Verma, E. Goralnick, S. Goldberg, H. Shokoohi, R. Dias, A. Eyre, Applications of Augmented Reality for Prehospital Emergency Care: Systematic Review of Randomized Controlled Trials, JMIR XR and Spatial Computing 2 (2025) e66222. URL: https://xr.jmir.org/2025/1/e66222. doi:10.2196/66222.

[15] J. N. Siebert, F. Ehrler, A. Gervaix, K. Haddad, L. Lacroix, P. Schrurs, A. Sahin, C. Lovis, S. Manzano, Adherence to AHA Guidelines When Adapted for Augmented Reality Glasses for Pediatric Cardiopulmonary Resuscitation: A Randomized Controlled Trial, Journal of Medical Internet Research 19 (2017) e183. doi:10.2196/jmir.7379.