From Explainable AI to Human-Centered System Reliability: Quantifying and Visualizing Calibrated Trust in Mission-Critical XR

Matthew Wilchek[1,*], Kurt Luther[2] and Feras A. Batarseh[3]

[1]U.S. Army, Combat Capabilities Development Command, C5ISR Center, Fort Belvoir, VA, USA
[2]Department of Computer Science and Center for HCI, Virginia Tech, Alexandria, VA, USA
[3]Department of Biological Systems Engineering, Virginia Tech, Arlington, VA, USA

Abstract

Extended Reality (XR) systems increasingly integrate Artificial Intelligence (AI) to support professionals in high-risk settings such as search and rescue, law enforcement, and military operations. Yet common approaches to trust and explainability often rely on qualitative assessments or offline explanations that are poorly suited to embodied, time-critical work. This paper outlines an emerging perspective on how mission-critical XR systems may better support calibrated trust, resilient oversight, and situated explainability through human-centered system reliability (HCSR), a quantitative, user-parameterized estimate of reliability that evolves through accumulated evidence. Drawing on prior work in distributed XR and human–AI teaming, we describe three connected shifts: from generic trust to calibrated trust through updated reliability estimates; from seamlessness to resilience through human-centered oversight; and from transparency to situated explainability through lightweight spatial cues embedded in the XR interface. We conclude with implications and open challenges for reliability-aware XR systems.

Keywords

Embodied Explainable AI, Trust Calibration, Mission-Critical Systems, Mixed Reality, Human-in-the-Loop AI, Shared Perception

1. Introduction

Extended Reality (XR) systems increasingly integrate Artificial Intelligence (AI) to support professionals in high-risk settings such as search and rescue (SAR), law enforcement, and military operations, contexts that exemplify *extreme sensemaking* under uncertainty, time pressure, and incomplete information [1]. Decades of military XR research have shown that head-up, spatially registered overlays can improve situational awareness (SA) by embedding relevant information directly into the user's view of the physical environment [2]. However, this same body of work also shows that simply adding information can overwhelm users and degrade SA, motivating role-aware filtering and careful cue presentation [3].

Trust and explainability are widely viewed as prerequisites for effective AI-assisted decision-making [4]. Early work conceptualized trust through qualitative dimensions such as perceived reliability, technical competence, and understandability [5, 6], while more recent surveys [7] emphasize the importance of explainable AI (XAI) in fostering user confidence, particularly in computer vision systems [8]. Yet these approaches are largely designed for offline inspection and reflection and are not well suited to embodied, time-critical XR use [9, 10]. Trust is a key determinant of team efficiency in human–AI collaboration, though its role is interpreted inconsistently across "trustworthy AI" literature [11, 12, 13]. In distributed teams, calibrated trust helps members act on AI insights while reducing coordination delays and decision friction [14, 15]. Following Gunning [16], we treat trustworthy AI not as algorithms that report confidence, but as systems that communicate reliability, uncertainty, and provenance in ways that support human judgment.

CHI'26: XR4CE workshop, April 14, 2026, Barcelona, Spain
*Corresponding author
matthew.r.wilchek.civ@army.mil (M. Wilchek); kluther@vt.edu (K. Luther); batarseh@vt.edu (F. A. Batarseh)
0000-0001-8664-1586 (M. Wilchek); 0000-0003-1809-6269 (K. Luther); 0000-0002-6062-2747 (F. A. Batarseh)

The PerceptiSync framework builds on this foundation by providing a human-centered implementation of Dirichlet–Categorical (DC) trust modeling for distributed AI systems, using user-configurable features and real-time human feedback to adjust trust assessments [17]. Empirical evaluations show that integrating Human-in-the-Loop (HITL) and Crowd-in-the-Loop (CITL) mechanisms can improve trust assessment and system reliability in dynamic environments [18, 19]. Despite these advances, more research is needed to understand how human-centered system reliability (HCSR) should be designed and conveyed in XR systems. For this workshop, we organize that perspective around three connected shifts: from generic trust to calibrated trust through HCSR, from seamlessness to resilience through human-centered oversight, and from transparency to situated explainability through XR interface cues.

2. Shift 1: Calibrated Trust via Human-Centered Reliability

A common approach to studying trust in AI-assisted systems relies on post-hoc, subjective methods such as Likert-style questionnaires or self-reported scales [20]. While useful for capturing perception, these measures offer limited support for real-time decision-making in dynamic, high-risk environments. In challenging environments, professionals must continually decide whether to rely on specific AI outputs, teammates, or sensing modalities as situations evolve. This motivates a move from static trust scores to reliability as a measurable, continuously updated signal.

Prior work has begun to formalize trust as a probabilistic, evidence-driven process. Guo et al. introduced a DC trust model in which trust between a source i and receiver j is represented as an opinion parameterized by accumulated positive, negative, and uncertain evidence [21]. The resulting reliability estimate is $\omega_{ij} = \frac{\alpha_{ij}}{\alpha_{ij}+\beta_{ij}+\gamma_{ij}}$, where α_{ij}, β_{ij}, and γ_{ij} denote positive, negative, and uncertain evidence, respectively. As observations are exchanged, these parameters are updated, yielding a time-evolving estimate of source reliability. Cirne et al. further show that such trajectories can exhibit recognizable patterns of growth, decay, or oscillation as evidence accumulates or conflicts over time [22]. Reliability differs from both model confidence and explanation: confidence reflects uncertainty in a single prediction, explanations clarify why a prediction was produced, and reliability captures how dependable a source has been across interactions and time. This distinction matters in distributed XR systems, where users must reason across heterogeneous AI agents, sensing modalities, and teammates whose performance may vary across contexts.

We extend this probabilistic model into *human-centered system reliability* (HCSR), a user-parameterized estimate of an information source's reliability that evolves through accumulated evidence exchange and individual trust preferences. The PerceptiSync framework implements this idea by combining user-configurable features with real-time human feedback to shape trust updates [17]. Treating HCSR as a foundational mechanism allows XR systems to convey both what an AI detected and how much a user should rely on that source, supporting calibrated trust, resilient oversight, and situated explainability.

3. Shift 2: Resilience through Human-Centered Oversight

Prior research on AI-enabled wearable and XR systems identifies trust, adaptability, cognitive burden, and error tolerance as central challenges for human-in-the-loop operation, noting that small errors or failures can present significant safety and performance risks in real-world deployments [19]. These findings suggest that mission-critical XR systems need to support human oversight when uncertainty arises [23]. In extreme sensemaking, resilience depends on interfaces and HITL mechanisms that help users perceive uncertainty, add contextual knowledge, and intervene when automated assessments diverge from reality. Within this framing, changing reliability estimates should inform when oversight is needed.

Our prior work on PerceptiSync builds on HCSR by providing explicit mechanisms for human-centered oversight grounded in CITL and HITL principles. Users define a trust persona that shapes when humans are consulted, when intervention is required, and how user judgment affects reliability

estimates [17]. These include *trust level*, *trust history*, and *trustworthy assessments*, which together determine how conservatively new information is incorporated, how much prior evidence is required, and how many reliable observations are needed before a source is reinforced or degraded. These features are not fixed settings. They determine how incoming evidence changes HCSR over time and when the system shifts from autonomous reliance to human-centered oversight. After a new observation, detection, or message is received, PerceptiSync records the evidence as positive, negative, or uncertain depending on whether the information is corroborated, contradicted, or remains ambiguous. The user's trust persona then shapes how strongly that evidence affects the current reliability estimate. In practice, this adjustment occurs in two ways: users may configure their trust persona for a mission context, such as low-visibility SAR operations or newly deployed sensing modalities, and the system may adapt implicitly as evidence accumulates. Repeated agreement between distributed agents can strengthen reliability and reduce the need for intervention, while conflicting observations, uncertain detections, or missing corroboration can slow trust growth, lower reliability, or trigger oversight prompts. In this way, calibrated trust is updated through a continuous loop of evidence, user preferences, and possible human review rather than through a fixed score.

This dynamic behavior also appeared in a prior user study of distributed AI in SAR teams, where PerceptiSync was deployed within a scenario-driven, multi-role simulation comparing AI-assisted and non-assisted conditions [18]. Participants selected trust configurations before the study, and these settings directly influenced how reliability scores were assessed between users. More moderate or trusting configurations were associated with higher Average Trust Scores (ATS) and shorter decision times, while more cautious configurations were associated with lower trust scores and longer deliberation. Qualitative feedback further suggested that trust assessments sometimes served as shared reference points for team dialogue and coordination.

PerceptiSync also includes an optional *trust monitoring* feature that enables users to inspect received information and override system-computed reliability values when necessary. This becomes especially important during breakdowns, such as when distributed sources disagree, a detection is weakly supported, or operational context suggests that the AI is missing something important. In these moments, HCSR should function not only as a descriptive score but also as a trigger for oversight. When enabled, users are shown both their perceived scene and information from the sending source, along with the system's assessed reliability, and may adjust or replace that value based on their real-time interpretation. While this can introduce bias, it enables subject matter experts (SMEs) such as SAR team leaders, drone operators, or canine handlers to correct cases where automated reliability assessments are wrong or incomplete due to occlusions, sensor blind spots, or operational context not represented in the model [24]. More broadly, SME oversight can also help regulate what information is shared and under what constraints, balancing openness with security and privacy in high-stakes collaboration [25]. Framing resilience around human-centered oversight helps reposition breakdowns as manageable moments for intervention rather than as failures of the system alone.

4. Shift 3: Situated Explainability in XR

Explainable AI (XAI) has traditionally been delivered through post-hoc explanations such as textual descriptions, saliency maps, or dashboard-style interfaces [26]. While valuable for debugging and offline inspection, these forms of explanation are often too demanding for embodied, time-critical work [27]. In high-risk settings, professionals need information that can be perceived at a glance and interpreted in context [14]. This points to a need for XAI as an integrated perceptual cue embedded directly within the XR interface, particularly when users must recognize changing reliability and decide whether intervention is needed during a breakdown.

Mixed reality provides a medium for such situated explanations by spatially anchoring digital information to physical objects, locations, and teammates. Prior work shows that such overlays can support rapid sensemaking. KHAIT uses first-person canine video streams as XR overlays to help handlers interpret hazards and detect survivors in real time [28], while Ajna fuses detections from

multiple sensing modalities into a common spatial frame for collaborative interpretation [1]. Building on these examples, we argue that reliability cues can function as explanations in their own right. Rather than only clarifying why an output was produced, situated explanations can communicate how that output should be interpreted in context and whether it should be acted on, treated cautiously, or reviewed further.

Our prior SAR user study using PerceptiSync trust cues illustrates both the promise and the current limitations of this idea [18]. Trust assessments supported collaboration and selective reliance across distributed AI agents, but they were conveyed primarily through verbal or textual forms rather than embodied XR cues. This leaves an important design question: how should reliability be visualized to support resilience during breakdowns? We argue that situated explainability is the interface mechanism that makes this possible. When sources disagree or evidence remains too uncertain to justify autonomous acceptance, the XR interface can surface that state directly at the point of use through peripheral highlights, anchored widgets, alert-style overlays, or other lightweight spatial signals that indicate when verification, override, or delayed action may be warranted. In this sense, situated explainability in XR not only reveals why AI produced an output, but also how that output should be interpreted and acted upon in the moment.

5. Implications and Open Challenges

This work points to a need for systematic methods to prototype and evaluate how HCSR operates within distributed XR systems before testing in challenging environments. User studies that evaluate AI systems and interfaces together can blur the distinction between system behavior and interface design, making it difficult to determine how reliability cues influence trust calibration, intervention timing, and team coordination [29, 30]. In particular, evaluation should isolate the progression outlined in this paper: how evidence updates HCSR, how HCSR informs oversight during uncertainty or disagreement, and how situated cues shape user response. One promising way to address this gap is the Council of Wizards (CoW), a Wizard-of-Oz method we previously proposed for studying distributed AI [18]. By pairing multiple human operators with Wizards to simulate distributed perception systems and information exchange, CoW enables controlled study of how different reliability cues and oversight strategies shape sensemaking, coordination, and decision-making.

How to visualize HCSR in XR so that it supports situated explainability remains an open challenge. Because these cues are intended to operationalize resilience during breakdowns, they should help users notice degraded reliability, localize the affected source or detection, and understand when confirmation or override is warranted. Reliability cues could be conveyed through changes in color, audio, iconography, or subtle spatial objects embedded within the user's field of view, such as a peripheral border or halo that shifts in color or intensity to signal cautious, moderate, or trusting reliability when viewing information from another user or AI system. Alternatively, reliability may be conveyed through small, glanceable XR widgets or alert-style overlays anchored to incoming detections, offering controls for validation or override. Determining which cue combinations are most interpretable, least disruptive, and most supportive of HITL override remains an open research question.

6. Declaration on Generative AI

The authors used ChatGPT 5.2 for grammar and spell checking, then reviewed and edited the manuscript and take full responsibility for its content.

References

[1] M. Wilchek, K. Luther, F. A. Batarseh, Ajna: A wearable shared perception system for extreme sensemaking, ACM Trans. Interact. Intell. Syst. 15 (2025). URL: https://doi.org/10.1145/3690829. doi:10.1145/3690829.

[2] M. A. Livingston, L. J. Rosenblum, D. G. Brown, G. S. Schmidt, S. J. Julier, Y. Baillot, J. E. Swan, Z. Ai, P. Maassel, Military Applications of Augmented Reality, Springer New York, New York, NY, 2011, pp. 671–706. URL: https://doi.org/10.1007/978-1-4614-0064-6_31. doi:10.1007/978-1-4614-0064-6_31.

[3] M. A. Livingston, Z. Ai, K. Karsch, G. O. Gibson, User interface design for military ar applications, Virtual Reality 15 (2011) 175–184. URL: https://doi.org/10.1007/s10055-010-0179-1. doi:10.1007/s10055-010-0179-1.

[4] F. A. Batarseh, L. Freeman, C.-H. Huang, A survey on artificial intelligence assurance, Journal of Big Data 8 (2021) 60. URL: https://doi.org/10.1186/s40537-021-00445-7. doi:10.1186/s40537-021-00445-7.

[5] G. C. Moore, I. Benbasat, Development of an instrument to measure the perceptions of adopting an information technology innovation, Info. Sys. Research 2 (1991) 192–222. URL: https://doi.org/10.1287/isre.2.3.192. doi:10.1287/isre.2.3.192.

[6] M. Madsen, S. D. Gregor, Measuring human-computer trust, in: Proceedings of the 11th Australasian Conference on Information Systems (ACIS 2000), Information Systems Management Research Centre, Queensland University of Technology, Brisbane, QLD, Australia, 2000, pp. 6–8. URL: https://api.semanticscholar.org/CorpusID:18821611.

[7] O. Vereschak, G. Bailly, B. Caramiaux, How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies 5 (2021). URL: https://doi.org/10.1145/3476068. doi:10.1145/3476068.

[8] F. Ekman, M. Johansson, J. Sochor, Creating appropriate trust in automated vehicle systems: A framework for hmi design, IEEE Transactions on Human-Machine Systems 48 (2018) 95–101. doi:10.1109/THMS.2017.2776209.

[9] M. Billinghurst, A. Clark, G. Lee, A survey of augmented reality, Foundations and Trends in Human-Computer Interaction 8 (2015) 73–272. URL: https://doi.org/10.1561/1100000049. doi:10.1561/1100000049. arXiv:https://www.emerald.com/fthci/article-pdf/8/2-3/73/11031378/1100000049en.pdf.

[10] M. Billinghurst, H. Kato, Collaborative augmented reality, Commun. ACM 45 (2002) 64–70. URL: https://doi.org/10.1145/514236.514265. doi:10.1145/514236.514265.

[11] K. A. Hoff, M. Bashir, Trust in automation: Integrating empirical evidence on factors that influence trust, Human Factors 57 (2015) 407–434. URL: https://doi.org/10.1177/0018720814547570. doi:10.1177/0018720814547570, pMID: 25875432.

[12] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human Factors 46 (2004) 50–80. URL: https://journals.sagepub.com/doi/abs/10.1518/hfes.46.1.50_30392. doi:10.1518/hfes.46.1.50_30392, pMID: 15151155.

[13] S. M. Merritt, H. Heimbaugh, J. LaChapell, D. Lee, I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system, Human Factors 55 (2013) 520–534. URL: https://doi.org/10.1177/0018720812465081. doi:10.1177/0018720812465081, pMID: 23829027.

[14] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, H. P. Beck, The role of trust in automation reliance, International Journal of Human-Computer Studies 58 (2003) 697–718. URL: https://www.sciencedirect.com/science/article/pii/S1071581903000387. doi:https://doi.org/10.1016/S1071-5819(03)00038-7, trust and Technology.

[15] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, R. Parasuraman, A meta-analysis of factors affecting trust in human-robot interaction, Human Factors 53 (2011) 517–527. URL: https://doi.org/10.1177/0018720811417254. doi:10.1177/0018720811417254.

[16] D. Gunning, Darpa's explainable artificial intelligence (xai) program, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19, Association for Computing Machinery, New York, NY, USA, 2019, p. ii. URL: https://doi.org/10.1145/3301275.3308446. doi:10.1145/3301275.3308446.

[17] M. Wilchek, M. Nguyen, Y. Wang, K. Luther, F. A. Batarseh, Perceptisync: Trustworthy object detection using crowds-in-the-loop for cyber-physical systems, ACM Transactions on Cyber-Physical

Systems (2025). URL: https://doi.org/10.1145/3746644. doi:10.1145/3746644, just Accepted (advance online publication); volume/issue not yet assigned.

[18] M. Wilchek, S. Dickinson, K. Luther, F. A. Batarseh, The influence of distributed ai in trust and collaboration for search-and-rescue teams, in: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), Association for Computing Machinery, New York, NY, USA, 2026, pp. 1–20. URL: https://doi.org/10.1145/3772318.3791523. doi:10.1145/3772318.3791523, just Accepted.

[19] M. Wilchek, W. Hanley, J. Lim, K. Luther, F. A. Batarseh, Human-in-the-loop for computer vision assurance: A survey, Engineering Applications of Artificial Intelligence 123 (2023) 106376. URL: https://www.sciencedirect.com/science/article/pii/S0952197623005602. doi:https://doi.org/10.1016/j.engappai.2023.106376.

[20] S. C. Kohn, E. J. de Visser, E. Wiese, Y.-C. Lee, T. H. Shaw, Measurement of trust in automation: A narrative review and reference guide, Frontiers in Psychology 12 (2021) 604977. URL: https://doi.org/10.3389/fpsyg.2021.604977. doi:10.3389/fpsyg.2021.604977.

[21] J. Guo, Q. Yang, S. Fu, R. Boyles, S. Turner, K. Clarke, Towards trustworthy perception information sharing on connected and autonomous vehicles, in: 2020 International Conference on Connected and Autonomous Driving (MetroCAD), 2020, pp. 85–90. doi:10.1109/MetroCAD48866.2020.00021.

[22] D. Cirne, V. Calambur, Fostering trust and quantifying value of ai and ml, in: K. Arai (Ed.), Proceedings of the Future Technologies Conference (FTC) 2024, Volume 2, Springer Nature Switzerland, Cham, 2024, pp. 586–603.

[23] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, E. Horvitz, Guidelines for human-ai interaction, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 1–13. URL: https://doi.org/10.1145/3290605.3300233. doi:10.1145/3290605.3300233.

[24] M. De-Arteaga, R. Fogliato, A. Chouldechova, A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1–12. URL: https://doi.org/10.1145/3313831.3376638. doi:10.1145/3313831.3376638.

[25] S. Venkatagiri, A. Gautam, K. Luther, Crowdsolve: Managing tensions in an expert-led crowdsourced investigation, Proc. ACM Hum.-Comput. Interact. 5 (2021). URL: https://doi.org/10.1145/3449192. doi:10.1145/3449192.

[26] T. Nguyen, A. Canossa, J. Zhu, How human-centered explainable ai interface are designed and evaluated: A systematic survey, 2024. URL: https://arxiv.org/abs/2403.14496. arXiv:2403.14496.

[27] J. Kim, H. Maathuis, D. Sent, Human-centered evaluation of explainable ai applications: a systematic review, Frontiers in Artificial Intelligence Volume 7 - 2024 (2024). URL: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1456486. doi:10.3389/frai.2024.1456486.

[28] M. Wilchek, L. Wang, S. Dickinson, E. Feuerbacher, K. Luther, F. A. Batarseh, Khait: K-9 handler artificial intelligence teaming for collaborative sensemaking, in: Proceedings of the 30th International Conference on Intelligent User Interfaces, IUI '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 925–937. URL: https://doi.org/10.1145/3708359.3712107. doi:10.1145/3708359.3712107.

[29] S. Naveed, G. Stevens, D. Robin-Kern, An overview of the empirical evaluation of explainable ai (xai): A comprehensive guideline for user-centered evaluation in xai, Applied Sciences 14 (2024). URL: https://www.mdpi.com/2076-3417/14/23/11288. doi:10.3390/app142311288.

[30] A. Rechkemmer, M. Yin, When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22, Association for Computing Machinery, New York, NY, USA, 2022. URL: https://doi.org/10.1145/3491102.3501967. doi:10.1145/3491102.3501967.