Human-Agent Trust Mediation: Augmented Context Understanding as a Method to Dynamically Calibrate Machine-Reliance in Emergency Response.

Rodrigo Gutierrez Maquilon[1,2,*], Georg Regal[1,2] and Manfred Tscheligi[1,2]

[1]AIT-Austrian Institute of Technology, Vienna, Austria
[2]PLUS-Paris Lodron University of Salzburg, Salzburg, Austria

Abstract

The combination of extended Reality (XR) with conversational AI is emerging as a situational awareness (SA) augmentation layer for Emergency First Responders (EFRs), where stress, uncertainty, and time pressure degrade spatial reasoning and decision-making. Current approaches in conversational AI powered by spatial vision language models (S-VLMs) enable reliable SA support like a robot reporting metrically-grounded distances of detected real-world objects, e.g., "The car is 10 meters away.", to mitigate error-prone human distance estimations. This process constitutes a static human-agent trust relation that has the potential risk of causing over-reliance in EFRs. What happens if the robot's distance measurement is wrong because of light conditions? or if there is a hidden combustion inside the car? Therefore, we argue toward calibrated trust, including real-time trust measurement in machine-report, accounting for unexpected events, and interface strategies that respond to over and under trust as they emerge, e.g., "The car is estimated at 10 meters away with 65% confidence giving the light conditions." or "The car is now 5 meters away and a crackling sound is detected near the same location, careful for potential hidden combustion." Building on our prior prototype demonstrating that ground-truth depth injected into a VLM improves distance judgments and SA in an XR-robotic EFR scenario, we argue that the next iteration must steer spatial context understanding toward conversational agentic trusting: a robot teammate that actively explores, detects dynamic hazards (e.g., structural instability), and communicates changing threat trajectories grounded in multimodal evidence (e.g., depth, audio, thermal cues). We contribute (1) a measurement method for modeling trust in real time under pressure using behavioral, attentional, system, and psychophysiological signals; and (2) actionable XR interface principles for dynamically calibrating reliance, preventing automation misuse, and preserving human authority.

Keywords

agency, trust, XR, conversational AI, multimodal LLM, emergency responders, spatial reasoning

1. Introduction

Complex and time-critical scenarios such as mass-casualty incidents, structural collapses, or chemical, biological, radiological, nuclear and explosives (CBRNE) events require SA skills as a core competence for EFRs [1, 2, 3]. Beyond recognizing relevant objects and hazards, responders must continuously maintain a mental model of spatial relations: how far they are from a victim, whether an exit route is passable, or how close a robot can safely approach a suspected source of contamination [4, 5] (see Fig. 1). If a conversational AI simply echoes or loosely interprets an EFRs own distance estimates, it risks reinforcing incorrect judgments. Conversely, if it contradicts human intuition without clear justification, it may be perceived as untrustworthy and be ignored [4]. Stakes are high in the EFR domain: distance is not just a geometric parameter but a proxy for risk, reachability, and intervention timing [6].

CHI'26: XR4CE workshop, April 14, 2026, Barcelona, Spain
*Corresponding author.
rodrigo.gutierrez@ait.ac.at (R. Gutierrez Maquilon); georg.regal@ait.ac.at (G. Regal); manfred.tscheligi@ait.ac.at (M. Tscheligi)
0000-0002-6736-3418 (R. Gutierrez Maquilon); 0000-0003-4483-7710 (G. Regal); 0000-0001-6056-7285 (M. Tscheligi)

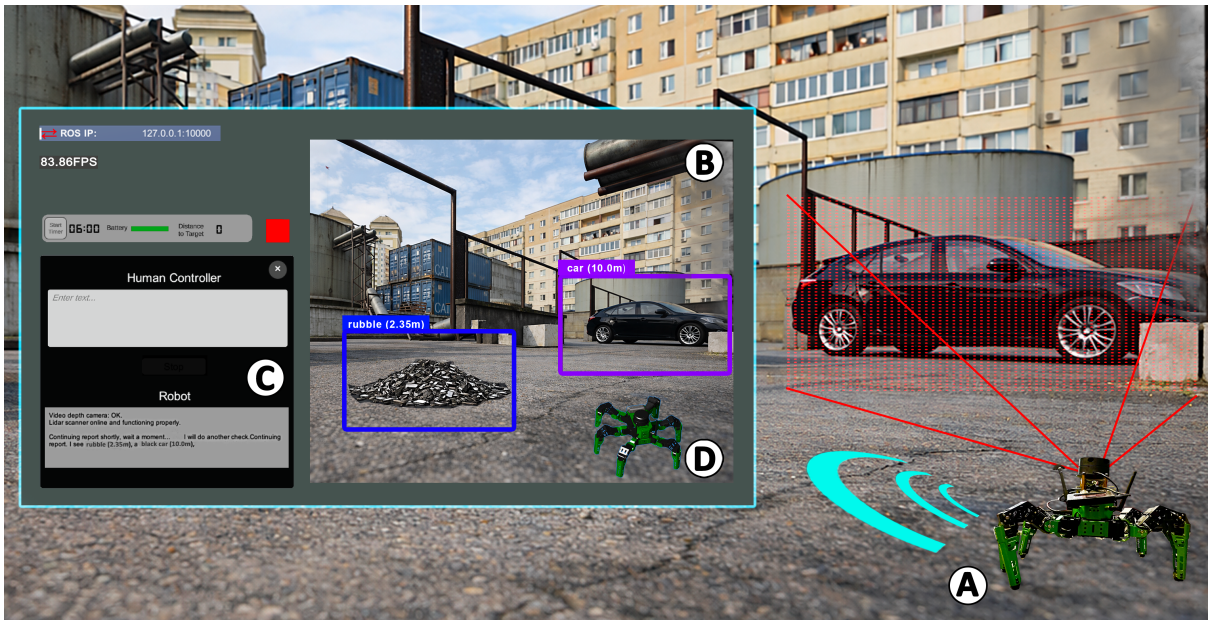

Figure 1: Mixed reality (MR) emergency response simulation of toxic smoke in an oil refinery. Depth measurements from the camera mounted on the robot (A) are shown in the labels of detected objects in the MR view (B). Corresponding verbal feedback, e.g., "The car is 10.0m away.", is shown on the chat window (C) and spoken through headphones. The robot also transmits its pose to enable 3D model tracking and visualization (D), e.g., behind obstacles in a mixed reality simulation. These audiovisual descriptions of the real-world environment are informed by metrically-grounded distances to support human error-prone estimations.

In this context, [7] assessed confidence and workload as indirect indicators of trust and cognitive efficiency. While these measures provided valuable insight into participants subjective experiences, they did not capture how trust and reliance on multimodal LLMs might evolve over time or under repeated exposure. The study then suggested that future research should examine how users trust calibration and situational awareness develop as they gain familiarity with the system and encounter varying levels of automation reliability. Moreover, integrating adaptive dialogue strategies, in which the LLM modulates the amount, timing, and modality of feedback based on estimated workload or situational demand, may further optimize the balance between SA and cognitive load. Such adaptive mechanisms, combined with depth-augmented spatial reasoning, have the potential to create robust, context-aware assistants that can flexibly support human operators in high-risk and rapidly evolving situations. Ultimately, these developments could pave the way for more trustworthy and resilient humanAI collaboration frameworks, where the system not only conveys spatial understanding but also aligns its communication dynamically with the humans cognitive state and task context.

2. Position

This paper argues that spatial language has to expand as a calibration mechanism: robots agency to (a) acquire disambiguating evidence, (b) predict evolving threats, and (c) communicate both risk and uncertainty in ways that prevent overtrust and undertrust. This aligns with the core trust framing in [8], where the goal is not maximum trust but appropriate reliance under uncertainty. Therefore, we propose the following:

2.1. Measuring and modeling trust in real time under pressure

Trust in automation should converge toward appropriate reliance under uncertainty, not toward maximal acceptance [8, 9]. In the mixed reality pipeline of Fig. 1, where robot depth, pose, and

conversational reports shape the responders situational picture, trust is enacted as a sequence of control choices: comply with guidance, verify it, or override it. Our earlier depth-grounded conversational support improved spatial judgments and situational awareness, yet relied mainly on post hoc confidence and workload as indirect proxies for trust [7]. Moving from proxies to trust calibration requires measurements that can be captured without interrupting work.

We propose a multi-channel measurement method that triangulates trust using behavioral, attentional, system, and psychophysiological signals (see Fig. 2). Behavioral traces operationalize reliance as action, capturing compliance with warnings, override frequency, latency to acknowledge and act, and verification strategies such as requesting a re-scan, asking for confidence, or seeking a second viewpoint. These traces allow calibration to be expressed as conditional reliance, accepting the agent when it is correct and withholding it when it is wrong, directly targeting misuse and disuse phenomena [10]. Attentional traces estimate whether calibration is feasible in the moment, using head pose, gaze, and interaction focus to capture whether world-locked evidence is inspected or ignored, especially under stress-driven narrowing views [4]. System traces capture what the agent knows about its current trustworthiness, including sensing quality, latency, staleness, and cross-modal agreement, consistent with work showing that situated uncertainty presentation and explanation timing shape trust and action [11, 12, 13]. Psychophysiological measures, such as HRV, EDA, and pupil dilation, are treated as moderators rather than trust meters, aligning with evidence that these signals are informative but non-specific [14].

Figure 2: Trust multi-channel measurement method. Behavioral, attentional, system, and psychophysiological signals are fused into a real-time probabilistic trust estimate that supports calibrated reliance and detection of misuse and disuse.

Methodologically, trust is modeled as a latent state in an online estimator, for example a Bayesian filter, that fuses these signals into a probabilistic trust estimate with explicit uncertainty bounds. Validation focuses on calibration outcomes rather than self-report, by systematically varying automation reliability and sensing conditions and testing whether trust estimates predict observable misuse or disuse before errors occur [15, 10, 9]. A common criticism is that real-time inference will be noisy, invasive, and overly personalized. However, the method privileges behavioral evidence, uses physiology only as context, propagates uncertainty rather than forcing crisp labels, and can rely on on-device feature extraction to minimize privacy risk.

2.2. XR interface principles for dynamic calibration of trust

A trust-aware interface must enable rapid recalibration while preserving human authority. Spatial conversational output becomes a calibration instrument when it clearly separates observation, inference, and recommendation, and when it makes uncertainty and staleness glanceable rather than hidden in fluent language. Evidence should be situated in the environment through anchored cues and optional show me pathways, so users can verify claims without disengaging from the task. Progressive disclosure is essential, delivering short, actionable utterances first, then revealing rationale and sensor detail only when requested or when risk is high. Interface assertiveness should adapt to overtrust and undertrust risk, softening guidance, foregrounding uncertainty, and introducing confirmation gates for high-consequence actions under potential overtrust, while strengthening grounding through cross-modal evidence and concise rationales under undertrust [12, 13]. When the system fails or revises a claim, trust repair should be

explicit and local, stating what changed, why it changed, and how behavior will adjust next time. See Fig. 3 for an overview of the proposed trust-aware XR interface design principles.

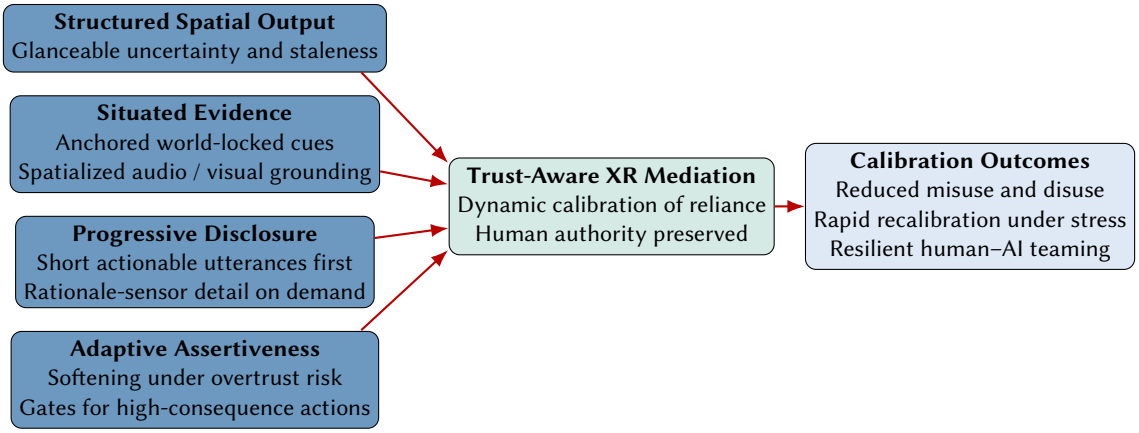

Figure 3: XR interface principles for dynamic calibration of reliance. Structured spatial output, situated evidence, progressive disclosure, and adaptive assertiveness converge in trust-aware XR mediation that preserves human authority while reducing misuse and disuse under stress.

It can be argued that adaptive XR calibration will distract responders or create alert fatigue. On the other hand, calibration is achieved through better timing and structure, not more information. By tying interventions to hazard severity, sensor uncertainty, and observed attention, and by privileging brief, situated cues over verbose explanations, adaptation can reduce cognitive translation costs and help responders retain a coherent mental model while remaining the ultimate decision maker.

3. Conclusion

Calibrated trust in emergency response demands both measurement and design. We advanced, first, a real-time trust modeling method that triangulates behavioral, attentional, system, and psychophysiological signals into a probabilistic estimate of trust calibration, second, XR interface principles that make uncertainty actionable, support verification, adapt assertiveness, and repair reliance while preserving human authority. The next step is a shared evaluation practice that tests these mechanisms under controlled reliability shifts and realistic stressors, so trust mediation becomes a measurable safety property rather than an assumed benefit.

Declaration on Generative AI

We used a ChatGPT exclusively to support language editing and phrasing in a small portion of the manuscript. All core design choices, analysis, figures, tables, and conclusions were produced by the authors.

References

[1] A. S. Baetzner, R. Wespi, Y. Hill, L. Gyllencreutz, T. C. Sauter, B.-I. Saveman, S. Mohr, G. Regal, C. Wrzus, M. O. Frenkel, Preparing medical first responders for crises: a systematic literature review of disaster training programs and their effectiveness, Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine 30 (2022). URL: http://dx.doi.org/10.1186/s13049-022-01056-8. doi:10.1186/s13049-022-01056-8.

[2] O. Zechner, D. García Guirao, H. Schrom-Feiertag, G. Regal, J. C. Uhl, L. Gyllencreutz, D. Sjöberg, M. Tscheligi, Nextgen training for medical first responders: Advancing mass-casualty incident preparedness through mixed reality technology, Multimodal Technologies and Interaction 7 (2023) 113. URL: http://dx.doi.org/10.3390/mti7120113. doi:10.3390/mti7120113.

[3] A. O'Brien, G. J. Read, P. M. Salmon, Situation awareness in multi-agency emergency response: Models, methods and applications, International Journal of Disaster Risk Reduction 48 (2020) 101634. URL: https://www.sciencedirect.com/science/article/pii/S2212420920300443. doi:https://doi.org/10.1016/j.ijdrr.2020.101634.

[4] M. R. Endsley, Toward a theory of situation awareness in dynamic systems, Human Factors 37 (1995) 32–64. URL: https://doi.org/10.1518/001872095779049543. doi:10.1518/001872095779049543. arXiv:https://doi.org/10.1518/001872095779049543.

[5] E. S. Spelke, K. D. Kinzler, Core knowledge, Developmental Science 10 (2007) 89–96. doi:10.1111/j.1467-7687.2007.00569.x.

[6] G. DA, B. A, C. I, D. G. L, D. D, D. A, G. L, L. A, M. M, P. S, P. C, S. V, S. M, S. L, Artificial intelligence applied to disasters and crises management (2024). doi:10.2760/0323818(online).

[7] R. G. Maquilon, M. Hueber, G. Regal, M. Tscheligi, Ground-truth depth in vision language models: Spatial context understanding in conversational ai for xr-robotic support in emergency first response, 2026. URL: https://arxiv.org/abs/2602.15237. arXiv:2602.15237.

[8] R. R. Hoffman, M. Johnson, J. M. Bradshaw, A. Underbrink, Trust in automation, IEEE Intelligent Systems 28 (2013) 84–88. doi:10.1109/MIS.2013.24.

[9] E. J. de Visser, M. M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, M. A. Neerincx, Towards a theory of longitudinal trust calibration in human–robot teams, International Journal of Social Robotics 12 (2020) 459–478. doi:10.1007/s12369-019-00596-x.

[10] R. Parasuraman, V. Riley, Humans and automation: Use, misuse, disuse, abuse, Human Factors 39 (1997) 230–253. doi:10.1518/001872097778543886.

[11] E. Brewer, G. Wu, B. Steers, S. Castelo, R. Lopez, I. Roman, S. Chen, J. Rulff, C. Zhao, A. D. Wilson, et al., Designing for situated interpretation of uncertainty in augmented reality, in: 2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), IEEE, 2023, pp. 616–625.

[12] C. Chen, M. Liao, S. S. Sundar, When to explain? exploring the effects of explanation timing on user perceptions and trust in AI systems, in: Proceedings of the Second International Symposium on Trustworthy Autonomous Systems (TAS '24), Association for Computing Machinery, New York, NY, USA, 2024. doi:10.1145/3686038.3686066.

[13] G. Papagni, J. de Pagter, S. Zafari, M. Filzmoser, S. T. Koeszegi, Artificial agents' explainability to support trust: considerations on timing and context, AI & Society 38 (2023) 947–960. doi:10.1007/s00146-022-01462-7.

[14] I. B. Ajenaghughrure, S. D. C. Sousa, D. Lamas, Measuring trust with psychophysiological signals: A systematic mapping study of approaches used, Multimodal Technologies and Interaction 4 (2020) 63. doi:10.3390/mti4030063.

[15] 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.