An Investigation of Trust and Overreliance on Conversational Instructions in Augmented Reality

Anja Rejc
Know Center Research GmbH
Graz, Styria, Austria
arejc@know-center.at

Juliana Hildegard Madritsch
Know Center Research GmbH
Graz, Austria
jmadritsch@know-center.at

Neven ElSayed
Know Center Research GmbH
Graz, Austria
nelsayed@know-center.at

Dominik Kowald
Know-Center
Graz University of Technology
Graz, Austria
University of Graz
Graz, Austria
dkowald@know-center.at

Simone Kopeinik
Know Center Research GmbH
Graz, Austria
skopeinik@know-center.at

Abstract

Recent advances in Large Language Models (LLMs) have enabled conversational agents that provide dynamic guidance in augmented reality (AR) systems. By embedding guidance directly into users' physical environments, AR may further shape trust and reliance during task execution. However, limitations in LLM accuracy pose challenges for user trust and reliance, yet the joint dynamics of trust, reliance, and mental workload under erroneous guidance remain underexplored, particularly in AR-supported LLM interaction settings. We conducted a laboratory study using an LLM-based AR interface in a physical testbed, manipulating LLM accuracy through controlled errors. We measured trust, mental workload, overreliance, task performance, and eye gaze. Results show that higher trust increased behavioral reliance, which in turn reduced task performance, suggesting that trust can undermine decision quality when users fail to critically evaluate LLM outputs. Gaze data confirmed reduced verification behavior under higher trust, and comparison with a prior dataset showed that erroneous LLM-based guidance significantly reduced trust relative to high-accuracy conditions. These findings highlight the importance of trust adjustment in enabling users to appropriately rely on LLMs.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**.

Keywords

trustworty ai, overreliance, cognitive bias, artificial intelligence, augmented reality

ACM Reference Format:
Anja Rejc, Juliana Hildegard Madritsch, Neven ElSayed, Dominik Kowald, and Simone Kopeinik. 2026. An Investigation of Trust and Overreliance on Conversational Instructions in Augmented Reality. In *ACM Conversational User Interfaces 2026 (CUI '26), July 21–24, 2026, Bremen, Germany*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3816046.3816296

CUI '26, Bremen, Germany

ACM ISBN 979-8-4007-2741-2/26/07
https://doi.org/10.1145/3816046.3816296

1 Introduction

The integration of immersive technologies with AI-generated guidance, such as that provided by large language model (LLMs), facilitates real-time support and show promise for training and analytical support, thereby improving learning effectiveness and task performance [23]. However, the limitations of AI-generated outputs' accuracy, combined with the additional cognitive demands of immersive environments [3], raise additional challenges for trust and reliance on AI [15], increasing the likelihood of overreliance-related decision errors.

In this work, we understand AI as a superordinate term that encompasses large language models (LLMs); thus, AI-generated guidance is discussed broadly in the theoretical framework, while the present study specifically investigates an LLM-based guidance system in AR. Within this context, trust and reliance on AI are central concerns. Trust refers to the belief that an agent will support the user in achieving their goals under uncertainty [16], while reliance represents its behavioral manifestation [16], i.e., acting on system outputs, including following incorrect system recommendations [12]. While related, trust does not fully determine reliance [16], which is also shaped by system- [8] and model-specific characteristics [12] and cognitive processes (e.g., heuristics) that can also limit users' ability to verify system outputs [12, 21]. Trust and reliance are also shaped through interaction, with early errors in AI-generated outputs having lasting effects on user behavior [13]. If such errors remain undetected, trust may not be adjusted, leading to continued overreliance and reduced task performance. Accordingly, effective human-AI interaction depends on calibrated trust [13], where reliance aligns with system performance [16]. Prior work has investigated trust, reliance, and mental workload separately or in pairwise combinations, while integrated empirical studies that jointly examine these factors in task-based human–AI interaction remain limited, particularly in AR-supported LLM interaction settings. Moreover, reliance and trust are often assessed through

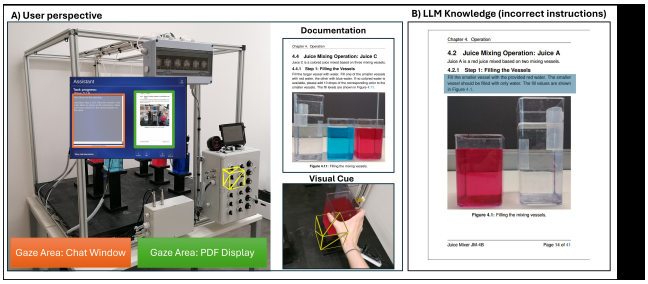

Figure 1: (A) User's point of view of an AR-supported task environment showing the physical testbed, visual cue, LLM-based assistant (Gaze Area: Chat Window), and documentation-grounded PDF instructions (Gaze Area: PDF Display). (B) Documentation used as the LLM-based assistant's knowledge source. In Task A, selected substeps were modified to introduce controlled errors while the user-facing PDF remained correct.

subjective self-reports or agreement with AI measures, with limited use of objective behavioral measures [5]. Eye tracking offers complementary, objective, real-time insights into user behavior, as it reflects attentional and information seeking processes [14], enabling the study of associations between reliance on AI [5], workload, and attention allocation [1].

To address these gaps, we formulate the following research questions:

- **RQ1:** Do early errors in LLM-based guidance influence changes in trust, reliance behavior, and task performance across tasks?
- **RQ2:** Are trust, reliance, and mental workload associated with users' gaze behavior and attention allocation when interacting with erroneous LLM-based guidance?
- **RQ3:** Does erroneous LLM-based guidance impact trust in AI?

We conduct a within-subject study in an AR-based task environment combining an LLM-based assistant and a PDF manual. We collect perceptional (trust in AI, trust in LLM, mental workload), and behavioral (task performance, overreliance) measures and incorporate gaze to capture how users allocate attention across competing information sources. This allows us to examine not only reliance on LLMs, but also how users engage with alternative sources during decision-making.

The study extends prior work [17] under a high-accuracy condition (100% accurate LLM output across all tasks) by introducing controlled errors in early task phases to examine how reduced accuracy of LLM-based guidance influences trust, reliance, and gaze behavior changes across tasks, further enabling comparison between conditions of high- and reduced-accuracy of LLM-based guidance. Guidance accuracy was experimentally manipulated through controlled errors introduced in the early engagement phases (Task A: ~81.25%, Tasks B/C: 100%), consistent with past research on trust sensitivity to system performance [2, 5, 10, 13, 22].

2 Methodology

2.1 Study Design

We conducted an ethically approved within-subject laboratory study to examine how LLM accuracy influences trust dynamics, mental workload, and reliance in an AR-supported environment[1]. Participants completed multi-step procedural tasks using an AR guidance system on a physical testbed, integrating three sources of information: a PDF-based instruction manual, a voice-driven LLM assistant, and AR visualizations on the physical setup. In this setting, participants completed three tasks of increasing complexity in a fixed order (see Sec . 2.4). The key independent variable was LLM accuracy (Task A: ~81.25%; see Sec 2.4). Dependent variables included behavioral (overreliance), subjective (trust in AI/LLM, mental workload), and objective (gaze, task performance) outcomes.

Procedure: Each 90-minute session began with a briefing in which participants provided informed consent, completed a guided system training session introducing the AR interface and the LLM-based assistant, followed by baseline questionnaires on trust in AI, trust in LLMs, demographics, and prior experience. In the study phase, participants completed three tasks using the system, while gaze data, task performance, and overreliance indicators were recorded; trust and mental workload questionnaires followed each task. The session concluded with a debriefing, informing participants of the LLM accuracy manipulation in Task A and collecting feedback. Participants received a €20 voucher as compensation.

2.2 Participants

A total of 18 participants (7 female, 11 male) were recruited online and screened for no prior experience with the experimental setup. Participants were aged 20–42 years (M = 29.72, SD = 5.77) and reported English proficiency (16 Advanced, 2 Native/Bilingual). Measured on a 4-point Likert-type scale (*None/Tried once or twice / Occasional use/ Regular use*), participants reported minimal prior experience with AR devices (M = 1.50, SD = 0.71) and with AR assistants (M = 1.06, SD = 0.24).

2.3 System Description

The system used in this study is an LLM-powered AR assistance platform designed to support multi-step task execution in a simulated machine environment, integrating a document-based AR interface, a voice-driven LLM-based assistant, and a physical laboratory testbed [6, 17]. In line with prior work on platforms for immersive analytics development and evaluation [9], the testbed represents an industrial-like setup comprising six vessels mounted on stands, equipped with sensors (e.g., temperature, pH, fill level) and a control panel for pump operations. Users interact with the **AR application**, built with the Mixed Reality Toolkit and deployed on a Microsoft HoloLens 2. The interface provides access to a digital PDF instruction manual, enables voice-based interaction with the LLM-based assistant, and displays spatial AR cues that highlight relevant elements in the physical environment. Users initiate interaction with the **LLM-based assistant** via a handheld button, while

[1]The study received ethical approval from the Ethics Committee of Graz University of Technology (Approval ID: EK-92/2025)

a Bluetooth microphone captures speech. Speech is transcribed using OpenAI Whisper2, processed by the document-grounded LLM via the OpenAI Assistants API, and returned as audio (via OpenAI TTS3) and text in the chat interface. The underlying documentation is a structured industrial manual with text and images for each step. It is included in the model's prompt using an in-context learning (ICL) approach, so the assistant's responses stay aligned with the documented procedures. The assistant provides guidance, referencing relevant document pages and steps, which also trigger AR-guidance visualizations in the physical setup. The position and size of these cues are preconfigured using a reference marker to ensure correct alignment.

The LLM-based assistant incorporated a **system-level accuracy manipulation**: the PDF instruction manual remained fully accurate across all tasks, whereas the version provided to the LLM was modified in Task A, as described in Section 2.4.

2.4 Task Design

The study extends the previous study design [17] by introducing controlled errors in LLM output. The study comprised three tasks of increasing complexity, completed in a fixed order (A–C). In each task, participants performed a **standardized multi-step procedure** (e.g., preparing vessels, connecting the testbed's components, operating the system, retrieving the final mixture) using the AR system, extended by an LLM-based assistant and a PDF manual within the AR interface.

Task A. Participants followed a sequential procedure using two mixing vessels and three liquids. Manipulating accuracy, the LLM-based assistant rendered controlled errors in three out of 16 substeps: incorrect vessel size instruction, incorrect lid/sensor placement sequence, and incorrect control pump number assignment.

Task B. To increase the complexity, a troubleshooting task was designed that extended the standard procedure. The LLM-based assistant operated with 100% guidance accuracy.

Task C. Independent of Task A, the physical testbed's initial configuration differed from the setup shown in the instructions. Participants had to identify and correct the mismatch before continuing, increasing task complexity. The LLM-based assistant operated with 100% guidance accuracy.

Errors introduced in task A were designed to appear plausible within the task context rather than being obviously incorrect. They could either be identified through critical evaluation of the task context or verification against the PDF documentation. A fixed task order was used because the study focused on how early AI errors influenced subsequent trust and behavior over time; counterbalancing would have disrupted these temporal effects.

2.5 Measures

To examine how variations in LLM accuracy affect user behavior and system interactions, several subjective and objective measures were collected, as described in this section.

2.5.1 Perception Measures.

Mental Workload. Mental workload was assessed after each task using the Raw NASA-TLX [4]. The scale captures workload on six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration, using a 6-item 20-point scale (1–20). To ensure consistency across dimensions, the Performance subscale was reverse-coded. For each participant and task, Raw NASA-TLX (R-TLX) score was computed as the mean of all six subscales, with higher values indicating higher perceived workload. Internal consistency (Cronbach's alpha) indicated acceptable to good reliability of the R-TLX across tasks (Task A: $\alpha = .78$; Task B: $\alpha = .85$; Task C: $\alpha = .89$), with slightly higher reliability in later tasks.

Trust in AI. Trust was assessed using an adapted version of the Trust in XAI Index [11, 20]. In line with prior validation [20], Item 6 was excluded. Item 7 was removed as it was not applicable to the AR setting of this study. Minor wording adjustments were introduced to ensure suitability across measurement points, primarily by adapting verb tenses. The final scale comprised six items rated on a 5-point Likert scale (1 = "Strongly disagree" to 5 = "Strongly agree"). The questionnaire was administered at baseline (prior to the task exposure) and after each task. Trust in AI scores were computed as the mean across all items for each task, with higher values indicating greater trust. Internal consistency (Cronbach's alpha) for each measurement point (baseline and after tasks A–C) indicated poor reliability at baseline ($\alpha = .57$), acceptable reliability in task A ($\alpha = .80$) and good reliability in Tasks B ($\alpha = .89$) and C ($\alpha = .88$).

Trust in the LLM. Trust in the LLM was assessed using the Trust-in-LLMs Index (TILLMI) [7], which captures cognitive (perceived competence, reliability) and affective (emotional response to interaction) trust, following McCallister's framework [18]. TILLMI was administered at baseline and after each task. TILLMI scores were calculated as the mean across all items for each task, with higher values indicating greater trust in LLM. Internal consistency (Cronbach's α) indicated questionable to acceptable reliability of the TILLMI across measurement points (Baseline: $\alpha = .72$; Task A: $\alpha = .58$; Task B: $\alpha = .65$; Task C: $\alpha = .69$), with the lowest reliability observed in Task A and a gradual increase in subsequent tasks. Despite this, the measure was retained due to its theoretical relevance, and results were interpreted with caution alongside additional measures.

2.5.2 Observational Measures.

Task Performance. Task performance was operationalized for each participant as the proportion of correctly completed substeps within each task (Task A: 16; Task B: 18; Task C: 16), including correctly and incorrectly instructed substeps. Each substep was coded by the experimenter as correct (1) or incorrect (0), and performance scores were calculated as the ratio of correct to total substeps.

Overreliance. Behavioral overreliance was operationalized following prior work on automation misuse [19], quantified as the "percentage of agreement with the AI when the AI made incorrect predictions" [2]. In Task A, where the LLM-based assistant's responses were intentionally manipulated to include errors, overreliance was calculated per participant using equation 1:

$$OverrelianceRate = \frac{NumberOfCommissionErrors}{TotalNumberOfAIErrorsPresented} \quad (1)$$

In Task A, this applied to three predefined substeps (3/16), enabling controlled assessment of participants' responses to incorrect guidance. A trial was coded as overreliance when participants followed incorrect LLM-based guidance without verification or correction, or despite having verified it against the PDF manual.

Eye Gaze Behavior. Eye-tracking (collected via Microsoft HoloLens 2) was used to analyze visual attention to physical components, where we defined two areas of interest (AOIs): (i) the chat window (LLM-based guidance) and (ii) the PDF display (original instructions). Fixation duration and total dwell time per AOI were computed per participant and task. Gaze allocation was interpreted as an exploratory indicator of attention distribution and information processing. Greater attention to the chat window was associated with reliance on LLM-based guidance, whereas increased PDF attention indicated verification behavior. Reduced attention to PDF was considered indicative of reliance on LLM-based guidance without verification.

3 Results

Data were analyzed using a combination of descriptive and inferential statistical methods to examine the effects of LLM accuracy on trust, mental workload, task performance, gaze behavior, and overreliance. To address RQ3, trust scores from a prior dataset (High-Accuracy Condition) [17] were compared with those from the present experimental condition. Baseline trust was assessed for comparability, followed by a comparison of trust after Task A using independent-samples t-tests and Mann–Whitney U tests.

3.1 RQ1: How do early errors in LLM-based guidance influence trust, reliance behavior, and task performance over time?

Participants exhibited a moderate level of overreliance on the LLM ($M = 1.17$, $SD = 1.20$), with 61.11% of participants committing at least one commission error, and 38.89% none. The most common outcome was a single error (27.78%), followed by three (22.22%) and two errors (11.11%). Overreliance decreased slightly across the three error opportunities in Task A (44%, 39%, 33%). Based on these responses, participants were categorized into an Error group ($n = 11$) and a NoError group ($n = 7$).

Trust in AI decreased after Task A and then stabilized across subsequent tasks in both groups (Error: $M = 3.14 \rightarrow 2.92 \rightarrow 2.86 \rightarrow 2.86$; NoError: $M = 2.98 \rightarrow 2.17 \rightarrow 2.19 \rightarrow 2.17$), with a more pronounced decrease in the NoError group. Trust in LLM remained relatively stable across tasks in both groups (Error: $M = 2.89 \rightarrow 2.82 \rightarrow 2.95 \rightarrow 2.89$; NoError: $M = 3.12 \rightarrow 2.83 \rightarrow 2.79 \rightarrow 2.86$), showing only minor changes across tasks.

A mixed-design ANOVA was conducted with Group (Error vs. NoError) as the between-subjects factor and measurement point (Baseline, after tasks A–C) as the within-subjects factor. Assumption checks indicated no substantial deviations from normality (all $ps > .05$). Violations of sphericity were observed ($W = .41$, $p = .022$), and Greenhouse–Geisser corrections were applied ($\epsilon = .67$). The analysis revealed a significant main effect of measurement point on trust in AI, $F(3, 48) = 7.15$, $p < .001$, $\eta^2_G = .10$ (corrected $p = .003$), indicating changes in trust over time. There was no significant main effect of Group, $F(1, 16) = 3.64$, $p = .074$, $\eta^2_G = .15$, or of interaction, $F(3, 48) = 1.97$, $p = .131$, $\eta^2_G = .03$ (corrected $p = .156$), suggesting similar patterns across groups. For trust in LLM, no significant effects of measurement point, group, or their interaction were observed (all $ps > .05$), indicating stable trust in LLM across tasks.

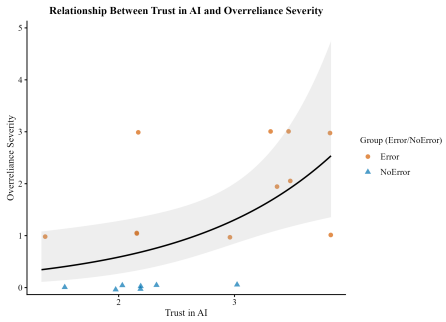

Figure 2: Relationship Between Trust in AI and Overreliance Severity in Task A.

In terms of task performance, participants in the NoError group consistently outperformed those in the Error group across tasks (NoError: $94.20 \rightarrow 96.03 \rightarrow 87.50$; Error: $69.32 \rightarrow 88.89 \rightarrow 76.99$).

This is further supported by a strong negative correlation between overreliance and task performance in Task A ($\rho = -.80$, $p < .001$), indicating that higher overreliance was associated with substantially lower performance. Poisson regression analyses further indicated that trust in AI significantly predicted overreliance in the single-predictor model including only trust in AI variable ($b = 0.80$, $p = .010$, IRR = 2.22; see Fig. 2) and remained significant when psychological variables were included ($b = 1.19$, $p = .042$, IRR = 3.30). However, this effect was no longer significant in the full model ($b = 1.41$, $p = .093$, IRR = 4.11).

3.2 RQ2: How are trust, reliance, and mental workload associated with users' gaze behavior and attention allocation when interacting with erroneous LLM-based guidance?

As shown in Fig. 3, trust in AI was associated with users' gaze behavior and attention allocation across tasks. Overall, higher trust in AI was related to longer gaze duration on the chat window (Pearson: $r = .33$, $p = .015$; Spearman: $\rho = .29$, $p = .033$) and shorter gaze duration on the PDF display (Spearman: $\rho = -.33$, $p = .013$). This pattern reached significance in Task C for chat gaze ($r = .57$, $p = .014$).

In contrast, trust in LLMs was not significantly associated with gaze behavior on either display (all $ps > .05$), nor was mental workload ($ps > .05$). Although workload decreased between tasks in both groups (Error: $M = 8.48 \rightarrow 6.39 \rightarrow 6.35$; NoError: $M = 8.12 \rightarrow 6.00 \rightarrow 5.55$), with a significant main effect of Task, $F(2, 32) = 6.67$, $p = .004$, $\eta^2_G = .084$ (corrected $p = .007$), no Group effects or interactions were observed (all $ps > .05$).

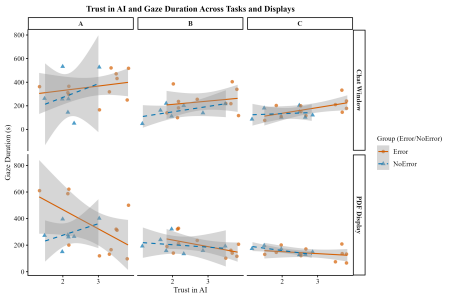

Figure 3: Relationship Between Trust in AI and Gaze Duration Across Tasks and Display Types.

Overreliance was not significantly associated with gaze behavior; in Task A, it was not correlated with gaze duration on the PDF display ($\rho = -.10$, $p = .71$) or the chat window ($\rho = -.13$, $p = .60$).

Descriptive analyses indicated a decrease in gaze duration across tasks for both the PDF display and the chat window in both groups (PDF display: Error group (333,322.7 ms → 195,660.5 ms → 138,324.1 ms); NoError group (289,163.7 ms → 199,804.3 ms → 153,852.3 ms); Chat Window: Error group (364,600.4 ms → 236,248.4 ms → 178,619.6 ms); NoError: 290,628.0 ms → 157,338.6 ms → 135,355.4 ms). Mixed-design ANOVAs confirmed significant main effects of Task for both PDF gaze, $F(2, 32) = 14.45$, $p < .001$, $\eta^2_G = .29$, and chat window gaze, $F(2, 32) = 17.68$, $p < .001$, $\eta^2_G = .35$, with no significant Group effects or interactions (all $ps > .05$).

Although no additional statistically significant associations were observed, several patterns emerged: Trust in LLM showed a trend-level negative association with PDF gaze duration in Task B ($r = -.42$, $p = .080$; $\rho = -.43$, $p = .077$). Weak negative associations between Trust in AI and mental workload were also observed across tasks (Task A: $\rho = -.32$, $p = .201$; Task B: $\rho = -.06$, $p = .825$; Task C: $\rho = -.19$, $p = .453$), although these were non-significant.

3.3 RQ3: Does erroneous LLM-based guidance impact trust in AI?

The assumptions of normality and homogeneity of variance were met; therefore, independent-samples t-tests were conducted. Baseline trust in AI did not differ significantly between conditions, $t = -1.83$, $p = .079$, 95% CI $[-0.86, 0.05]$, $d = -0.61$. However, trust in AI after Task A was significantly lower in the Experimental condition ($M = 2.63$) than in the High-Accuracy Condition ($M = 3.51$), $t = -2.84$, $p = .008$, 95% $CI[-1.51, -0.25]$, with a large effect size ($d = -0.95$), indicating a substantial difference between the conditions.

4 Conclusion, Discussion, and Outlook

This study investigated how erroneous LLM-based guidance influences user trust, reliance behavior, and attention allocation in an AR-supported environment. Consistent with prior work [2], results showed that higher trust in AI was associated with increased behavioral overreliance, which was negatively associated with task performance. This indicates that trust can promote reliance behaviors, undermining decision quality when users fail to critically evaluate LLM outputs.

Gaze data provided insight into underlying behavioral mechanisms: higher trust in AI was linked to less attention to the PDF and more to the Chat window, suggesting reduced verification, while lower trust showed the opposite pattern, indicating continued monitoring and independent information seeking. These findings indicate that trust shapes both reliance decisions and attention allocation across information sources. Mental workload did not differ between groups. Participants who verified and corrected LLM errors did not report higher effort than those who overrelied, suggesting that overreliance is not primarily driven by cognitive demand, but by differences in trust and critical engagement.

Extending prior findings [13], results suggested that changes in trust depend on error detection: participants who did not follow incorrect LLM guidance showed patterns consistent with trust adjustment and continued verification, whereas those who did follow incorrect LLM guidance relied on instructions without adjusting trust. This reveals a key mechanism of human-AI interaction failure: trust, un-adjusted to erroneous system performance, reduces verification, leading to overreliance and failure to detect errors. Consistent with this, the comparison between high-accuracy and experimental conditions showed that exposure to erroneous LLM behavior significantly reduced trust after initial interaction.

Several theoretically meaningful patterns were observed, including a negative association between trust in AI and mental workload, as well as a trend-level negative association between trust in LLM and PDF gaze duration, warranting further investigation.

4.1 Limitations

The relatively small sample size limits the generalizability of the findings; the laboratory setting may not fully capture the complexity of real-world environments; behavior was measured in a controlled scenario with predefined errors, which may differ from naturally occurring LLM errors; gaze behavior provides valuable insights into attention allocation, but does not directly capture users' cognitive processes, and should be interpreted with caution; lastly, the findings are based on a specific AR-based LLM system and may not generalize to other interfaces, domains, or AI modalities.

4.2 Outlook

Findings of this research contribute to research on trust in human–AI interaction by highlighting the distinction between trust as a subjective attitude and reliance as its behavioral manifestation. They further highlight the importance of integrating subjective, behavioral, and physiological measures to better understand real-time trust dynamics and how trust shapes human-AI interaction. Building on this, conversational user interfaces, especially when used in decision-making or training scenarios, should become more interactive and context-aware, using indicators to detect overreliance in real time and enable adaptive interventions that support appropriate reliance and critical evaluation. This is particularly important in immersive environments with integrated LLMs, where increased cognitive demands and the specific nature of conversational LLMs may further reduce users' tendency to verify outputs.

Acknowledgments

This work has been partially funded by the Horizon Europe project LUMEN (GA: 101187940) and the FFG COMET module DDIA. The Know Center is funded within the Austrian COMET Program – Competence Centers for Excellent Technologies – under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth, and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

[1] Paul Ayres, Joy Yeonjoo Lee, Fred Paas, and Jeroen J. G. van Merriënboer. 2021. The Validity of Physiological Measures to Identify Differences in Intrinsic Cognitive Load. *Frontiers in Psychology* 12 (Sept. 2021). doi:10.3389/fpsyg.2021.702538

[2] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (April 2021), 21 pages. doi:10.1145/3449287

[3] Josef Buchner, Katja Buntins, and Michael Kerres. 2022. The impact of augmented reality on cognitive load and performance: A systematic review. *Journal of Computer Assisted Learning* 38, 1 (2022), 285–303. doi:10.1111/jcal.12617 Published online: 4 November 2021.

[4] James C. Byers, Alvah C. Bittner, and Susan G. Hill. 1989. Traditional and Raw Task Load Index (TLX) Correlations: Are Paired Comparisons Necessary? In *Advances in Industrial Ergonomics and Safety*, Anil Mital (Ed.). Taylor & Francis, 481–485.

[5] Shiye Cao and Chien-Ming Huang. 2022. Understanding User Reliance on AI in Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–23. doi:10.1145/3555572 Published online: 11 November 2022.

[6] Tomislav Duricic, Peter Müllner, Nicole Weidinger, Neven A. M. ElSayed, Dominik Kowald, and Eduardo E. Veas. 2024. AI-Powered Immersive Assistance for Interactive Task Execution in Industrial Environments. In *ECAI 2024 - 27th European Conference on Artificial Intelligence (Frontiers in Artificial Intelligence and Applications, Vol. 392)*. IOS Press, 4491–4494. doi:10.3233/FAIA241037

[7] Edoardo Sebastiano De Duro, Giuseppe Alessandro Veltri, Hudson Golino, and Massimo Stella. 2025. Measuring and identifying factors of individuals' trust in Large Language Models. arXiv:2502.21028 [cs.HC] https://arxiv.org/abs/2502.21028

[8] Sven Eckhardt, Niklas Kühl, Mateusz Dolata, and Gerhard Schwabe. 2025. A Survey of AI Reliance. *Comput. Surveys* 58, 6 (Dec. 2025), 1–37. doi:10.1145/3776528

[9] Neven ElSayed, Oliver Prentner, Nicole Weidinger, and Eduardo Veas. 2025. The Juice Mixer: A Platform for Developing and Evaluating Immersive Analytics. In *2025 IEEE Conference on Human Factors in Immersive Analytics (HFIA)*. IEEE, Vienna, Austria, 1–5. doi:10.1109/HFIA68651.2025.00005 2–3 November 2025.

[10] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–29. doi:10.1145/3610067

[11] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5 (2023), 1096257. doi:10.3389/fcomp.2023.1096257 Published online: 6 February 2023.

[12] Lujain Ibrahim, Katherine M. Collins, Sunnie S. Y. Kim, Anka Reuel, Max Lamparth, Kevin Feng, Lama Ahmad, Prajna Soni, Alia El Kattan, Merlin Stein, Siddharth Swaroop, Ilia Sucholutsky, Andrew Strait, Q. Vera Liao, and Umang Bhatt. 2025. Measuring and mitigating overreliance is necessary for building human-compatible AI. arXiv:2509.08010 [cs.CY] https://arxiv.org/abs/2509.08010

[13] Patricia K. Kahr, Gerrit Rooks, Martijn C. Willemsen, and Chris C. P. Snijders. 2024. Understanding Trust and Reliance Development in AI Advice: Assessing Model Accuracy, Model Explanations, and Experiences from Previous Interactions. *ACM Trans. Interact. Intell. Syst.* 14, 4, Article 29 (Dec. 2024), 30 pages. doi:10.1145/3686164

[14] Adem Korkmaz. 2026. Mapping Eye-Tracking Research in Human–Computer Interaction: A Science-Mapping and Content-Analysis Study. *Journal of Eye Movement Research* 19, 1 (Feb. 2026), 23. doi:10.3390/jemr19010023

[15] Dominik Kowald, Sebastian Scher, Viktoria Pammer-Schindler, Peter Müllner, Kerstin Waxnegger, Lea Demelius, Angela Fessl, Maximilian Toller, Inti Gabriel Mendoza Estrada, Ilija Šimić, et al. 2024. Establishing and evaluating trustworthy AI: overview and research challenges. *Frontiers in Big Data* 7 (2024), 1467222.

[16] J. D. Lee and K. A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (Jan. 2004), 50–80. doi:10.1518/hfes.46.1.50_30392

[17] Juliana H Madritsch, Tomislav Duricic, Neven ElSayed, Simone Kopeinik, and Eduardo Veas. 2026. AI-Powered Conversational Assistance in Augmented Reality for Multi-Step Tasks. In *2026 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 519–529.

[18] Daniel J. McAllister. 1995. Affect- and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations. *Academy of Management Journal* 38, 1 (1995), 24–59. doi:10.5465/256727 Published: 1 February 1995.

[19] Raja Parasuraman and Victor Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39, 2 (June 1997), 230–253. doi:10.1518/001872097778543886

[20] Sebastian A. C. Perrig, Nicolas Scharowski, and Florian Brühlmann. 2023. Trust Issues with Trust Scales: Examining the Psychometric Quality of Trust Measures in the Context of AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. ACM, New York, NY, USA, 1–7. doi:10.1145/3544549.3585808 19 April 2023.

[21] Linda J. Skitka, Kathleen L. Mosier, and Mark Burdick. 1999. Does Automation Bias Decision-Making? *International Journal of Human-Computer Studies* 51, 5 (1999), 991–1006. doi:10.1006/ijhc.1999.0252

[22] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 129 (April 2023), 38 pages. doi:10.1145/3579605

[23] Steven Yoo, Sakib Reza, Hamid Tarashiyoun, Akhil Ajikumar, and Mohsen Moghaddam. 2024. AI-Integrated AR as an Intelligent Companion for Industrial Workers: A Systematic Review. *IEEE Access* 12 (2024), 191808–191827. doi:10.1109/access.2024.3516536