RelEYEance: Gaze-based Assessment of Users' AI-reliance at Run-time

ZEKUN WU, Saarland University, Saarland Informatics Campus, Germany
YAO WANG, University of Stuttgart, Germany
MARKUS LANGER, University of Freiburg, Department of Psychology, Germany
ANNA MARIA FEIT, Saarland University, Saarland Informatics Campus, Germany

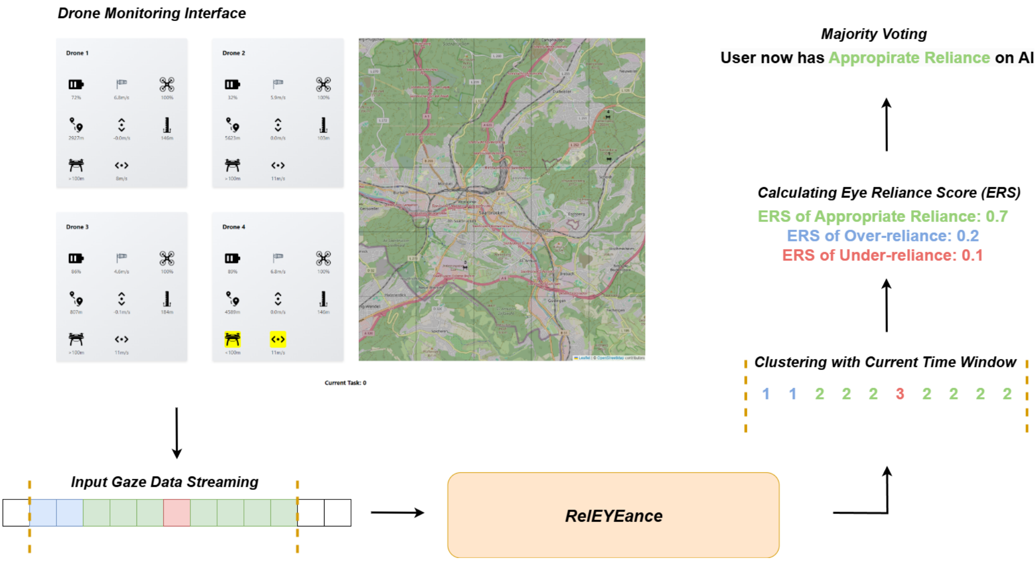
Drone Monitoring Interface
Drone 1
Drone 2
Drone 3
Drone 4
Current Task: 0
Majority Voting
User now has Appropirate Reliance on AI
Calculating Eye Reliance Score (ERS)
ERS of Appropriate Reliance: 0.7
ERS of Over-reliance: 0.2
ERS of Under-reliance: 0.1
Clustering with Current Time Window
1 1 2 2 2 3 2 2 2 2
Input Gaze Data Streaming
RelEYEance

Fig. 1. We propose *RelEYEance*, an online clustering pipeline to detect user reliance (over-, under-, or appropriate) in real-time from a stream of gaze data during drone monitoring tasks. The method takes a stream of gaze data as input, clusters with the current time window, and yields Eye Reliance Score (ERS) for real-time feedback to users using an AI system. Shown separately in the center is the drone monitoring interface used in this research, which displays four drone status panels on the left – each with icons indicating the real-time state of each drone – and a map on the right that visualizes drone positions. This interface enables operators to identify critical situations, such as Drone 4 approaching a no-fly zone at overspeed.

In time-critical detection tasks, such as drone monitoring, a key condition for users to effectively leverage AI assistance is to find an appropriate trade-off between making fast decisions and verifying AI suggestions, which we refer to as appropriate user reliance. However, assessing such reliance is often oversimplified by focusing solely on task outcomes, potentially overlooking whether users properly verify AI messages. We collected eye-tracking data from an AI-assisted monitoring task and developed a gaze-based reliance model: *RelEYEance*, to assess the extent of user reliance on AI-suggested alarms. We found that gaze patterns

Authors' Contact Information: Zekun Wu, wuzekun@cs.uni-saarlaned.de, Saarland University, Saarland Informatics Campus, Saarbrücken, Saarland, Germany; Yao Wang, yao.wang@vis.uni-stuttgart.de, University of Stuttgart, Stuttgart, Germany; Markus Langer, markus.langer@psychologie.uni-freiburg.de, University of Freiburg, Department of Psychology, Freiburg im Breisgau, Germany; Anna Maria Feit, feit@cs.uni-saarland.de, Saarland University, Saarland Informatics Campus, Saarbrücken, Germany.

© 2025 Copyright held by the owner/author(s).
This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *Proceedings of the ACM on Human-Computer Interaction*, https://doi.org/10.1145/3725841.

related to verification behaviors distinguish between appropriate reliance, over-reliance, and under-reliance, influencing task performance. We validated our model in a second user study, showing it can reliably detect users' over- and under-reliance at run-time, which could be used e.g. for issuing intervention messages. The results demonstrate the potential for real-time human-AI reliance assessment, facilitating adaptive reliance calibration.

CCS Concepts: • **Human-centered computing → User models**; **User studies**; *Graphical user interfaces*; *Empirical studies in interaction design.*

Additional Key Words and Phrases: visual highlighting; saliency models; human oversight; monitoring interfaces; gaze behavior; eye tracking; visual attention

ACM Reference Format:
Zekun Wu, Yao Wang, Markus Langer, and Anna Maria Feit. 2025. RelEYEance: Gaze-based Assessment of Users' AI-reliance at Run-time. *Proc. ACM Hum.-Comput. Interact.* 9, 3, Article ETRA16 (May 2025), 18 pages. https://doi.org/10.1145/3725841

1 Introduction

Maintaining appropriate reliance on AI support is critical for human operators working in time-sensitive scenarios, such as air traffic control or network operations centers with alarm systems. In these environments, inappropriate reliance on AI can result in severe consequences. For example, a pilot ignoring a cockpit warning in favor of personal judgment could have disastrous outcomes. Therefore, there is a pressing need for an instant and reliable measure of operator reliance on AI, ideally in real-time without knowing the actual correctness of decisions (ground truth). Additionally, AI support in these scenarios can take various forms—visual, auditory, or a mix of both—raising the challenge of understanding how reliance varies across modalities. Drone monitoring, in particular, exemplifies these challenges, as operators are tasked with managing multiple autonomous drones in dynamic environments, requiring quick yet careful responses to AI-suggested alarms. Identifying a robust and consistent reliance indicator that applies across different forms of AI support in such task scenarios is vital to fostering effective human-AI collaboration.

The current assessment of appropriate reliance is mainly focused on two approaches: outcome agreement and subjective perception [32, 33]. Outcome agreement considers reliance appropriate only when the user makes the correct decision, whether by relying on themselves or following the AI's suggestion [1, 33, 36]. However, this measure risks overlooking the user's actual utilization of AI information by not considering their decision-making process. For example, in the case of a highly reliable AI system, a user who blindly accepts the AI's suggestion and a user who meticulously verifies all information before making a decision may have similar outcome agreements, despite differing significantly in their reliance on the AI. In comparison, using subjective perceptions of reliance —often measured through self-reports— come with other limitations. First, they are susceptible to biases, as users may overestimate or underestimate their reliance due to individual differences [21, 25]. More importantly, self-reports are impractical for continuous, real-time assessment during tasks, as they interrupt the workflow and cannot capture moment-to-moment fluctuations in reliance.

In response to the need for a prompt and accurate method to assess user reliance on real-time AI notifications in time-critical tasks, we developed *RelEYEance*, a clustering model leveraging real-time eye-tracking data as a reliable indicator of human reliance on AI in decision-making contexts [4]. RelEYEance uses gaze features such as fixation count on the targeted AI alarm to infer reliance levels during monitoring tasks.

Our research involved two user studies within a consistent drone monitoring scenario assisted by AI alarms. The first study aimed to identify which gaze features most accurately capture different levels of user reliance on AI. We analyzed gaze behavior across various AI assistance modalities and determined the features that effectively distinguish between over-reliance, under-reliance, and appropriate reliance. Using the collected gaze data, we applied the RelEYEance model to cluster user verification gaze behavior in each task trial into distinct reliance categories, leading to the development of an Eye Reliance Score (ERS) to quantify the user's inclination toward specific reliance types.

Building on the insights from the first study, the second study evaluated the effectiveness of RelEYEance in real-time scenarios. We implemented the clustering model in an online pipeline to identify user reliance instantly during task performance. When inappropriate reliance was detected through majority voting based on ERS, intervention messages were issued to recalibrate user reliance on AI.

Our contributions are in three aspects:

- First, we found significant differences in user verification gaze behavior across varying levels of reliance on AI during a time-sensitive monitoring task. The identified gaze features consistently indicated user reliance across different AI assistance modalities, helping to establish a strong foundation for our model.
- Second, we utilized RelEYEance, an unsupervised clustering model using eye-tracking data, to classify user verification gaze behavior into distinct reliance categories: over-reliance, under-reliance, and appropriate reliance. These categories revealed significant differences in verification effort and task performance, with appropriate reliance leading to the highest F1 scores in this binary detection task.
- Third, we integrated the RelEYEance into an online reliance clustering pipeline. The model accurately identified inappropriate reliance induced by varying AI reliability levels and provided detailed insights into ERS dynamics over time. This demonstrates its potential for real-time assessment of user reliance and for delivering timely interventions to address inappropriate reliance on AI alarms.

2 Background and Related work

2.1 Reliance in AI-assisted decision making

The rapid advancement of automation has driven extensive research on user reliance on automated systems [19, 24]. We acknowledge that prior studies differentiated between trust and reliance and build upon this distinction by focusing on reliance as a behavioral manifestation of trust [35]. In other words, we did not aim to assess trusting intentions (e.g., via self-report measures) but on reliance behavior during the interaction of a person and an AI system. [15, 26]. Broadly, user reliance can be classified as appropriate or inappropriate, with inappropriate reliance further divided into over-reliance and under-reliance [21, 31]. Over-reliance occurs when users accept incorrect AI recommendations without adequate judgment [31], while under-reliance reflects a reluctance to trust or use AI, even when it offers correct or useful advice [33]. Inappropriate reliance can stem from various factors, including biases toward AI such as expecting near-perfect performance [21, 25, 27], a lack of experience with an AI system thus having little evidence about the actual performance of those systems [16, 29], or unfit explanations for system behavior and outputs [2, 20]. Research shows that inappropriate reliance often leads to poorer task performance compared to independently, which can have catastrophic consequences in safety-critical contexts, such as UAV monitoring [24, 30]. Therefore, accurately assessing and promptly identifying inappropriate reliance is essential for effective human-AI collaboration.

Various methods have been developed to assess human reliance on AI and identify inappropriate reliance [3, 16]. One common approach is to use decision-based metrics, such as the agreement rate between user decisions and AI suggestions [25] or the switch rate of users changing answers to align with AI recommendations [16, 25]. Other outcome-based measures include tracking human errors [3], full delegation to AI [6], or weighting of AI advice [23]. Additionally, self-reports are widely used to estimate user reliance on AI [4, 39]. Such measures are limited in their usefulness to detect inappropriate reliance in time-critical detection tasks, given that they are based on asking users directly or require knowledge about the accuracy of the outcome. Moreover, they fail to consider the decision-making behavior. Recently, gaze tracking has been explored as an alternative to assess users' trust during Human-Robot Interaction [18, 39] and gaze metrics, such as fixation duration, were shown to correlate with perceived reliance and agreement rate in AI-supported decision-making tasks [4, 5]. Building on these preliminary findings, we thus develop an approach to detect inappropriate reliance in real-time through measures of gaze behavior.

2.2 Real-time Feedback in Human-AI Collaboration

Real-time feedback systems have demonstrated the potential to enhance decision-making by alerting users to risky AI recommendations [25], fostering reflective engagement [20], and supporting triangulated decision-making [31]. For instance, Lai and Tan [20] found that explicitly conveying machine performance improved human accuracy in deception detection tasks beyond what explanations alone achieved. However, such approaches risk inducing over-reliance, as users may interpret accuracy scores at face value without considering underlying uncertainties [25, 31]. Real-time feedback mechanisms are increasingly explored in areas with a high demand for quick responses, including sports [17] and healthcare [38]. Adaptive systems research has further investigated context-aware models that tailor information displays based on cognitive load and task context [22], as well as collaborative VR frameworks enabling real-time creative collaboration, such as animated sketching and scene editing [14]. Despite these advancements, few studies directly address recalibrating user reliance on AI through feedback. This work aims to bridge this gap by delivering feedback designed to recalibrate AI reliance.

3 Investigating User Reliance Through Gaze Behavior: Study Design and Findings

To develop a method for assessing the user's reliance on AI assistance in a time-critical task, we first set out to understand whether gaze data could indicate users' reliance levels on AI systems — specifically distinguishing between appropriate reliance, over-reliance, and under-reliance. In relation to that we also wanted to understand how different user reliance measures (detection agreement, self-reported reliance), and gaze behavior differ across various AI modalities (such as visual versus audio recommendations). Thus, we designed a study that collected eye-tracking data from users while engaging in a time-critical monitoring task. Concretely, users' task was to monitor multiple autonomous drones and to report quickly and accurately when any of the drones encountered a critical situation. This scenario was chosen to study user reliance on AI because it balances complexity and ease of use. While the task involves monitoring different types of information, it is still manageable without specialized knowledge. Participants experienced different conditions related to the AI support they received.

3.1 Monitoring Interface and Task Overview

Informed by previous research [7, 12, 13, 34, 37] and resembling existing multi-drone monitoring interfaces [10, 11, 37], we developed a multi-drone monitoring user interface that was suitable for controlled user studies. As illustrated in Figure 1, the interface combines icon-based elements, facilitating quick assimilation of drone parameters, with a map display to grant enhanced spatial awareness and immersion. Each drone block features eight elements, showcasing core drone metrics and a representative image (for simplicity we refer to the combination of both the image and textual or numeric data as *icon* in the following). These were chosen to cover different categories of data relevant to drone monitoring [37]. The icons visually symbolize these core metrics which makes it easy to interpret the corresponding sensor values displayed below. Throughout the study tasks described below, the icons retained a static visual representation. Only the underlying values are updated according to the drone's simulated state.

In the monitoring task, participants had two goals: (1) to detect and acknowledge critical situations by pressing the space bar, and (2) to monitor the drone locations on the map displayed on the right side of the interface.

3.1.1 Detecting critical situations. Participants were asked to identify all critical situations by observing any changes in the indicator values for each drone block. During the study, participants encountered four distinct critical situations, each representing a different contextual scenario, and each lasting for 7 seconds. When the critical situation happened, two out of the eight icons transitioned into a critical range, as shown in the Figure 1 where drone four is having the zone breach critical situation (see the supplemental materials for details). Upon detecting a critical situation, the participants should press the space bar to acknowledge detection. We separated each monitoring task into 30 detection trials, each lasting between 15 and 20 seconds. A critical event had a 60% chance of occurring in each trial, with the onset of these events randomized to start between 5 and 10 seconds after the trial began. To prevent order effects, the sequence of trials within each task was randomized individually for each participant, ensuring that no two participants experienced the same order of trials.

3.1.2 Drone locating. In addition to detecting critical situations, participants were also required to track the position of all drones on the map, displayed on the right side of the interface. At the end of each trial, participants were asked to select the appropriate grid corresponding to the location of each specific drone. This subtask was there to simulate the limited attentional resources that human overseers have when monitoring autonomous systems while also having to perform other (work-related) tasks. It also encourages participants to rely on the AI system to be able to perform well in both tasks.

3.2 Study Design

In this study, we employed a one-factorial design with AI assistance as a within-subjects factor, encompassing four levels, each representing a different AI assistance modality. Each participant began with no AI assistance and then completed tasks with three different kinds of AI assistance in different modalities presented in a random order.

3.2.1 AI Assistance Conditions. To simulate AI assistance in a controlled manner, we employed a Wizard-of-Oz setup, where AI responses were manually pre-programmed to simulate specific reliability levels and assistance modalities

- Visual AI Assistance: The system highlighted the corresponding icon in yellow, blinking for 1 second at each alarm instance. For example, in the case of a zone breach, the icon for Drone 4 was highlighted, as shown in Figure 1. This representation ensured visibility while minimizing distraction.
- Audio AI Assistance: The warnings were also delivered via concise vocalized messages in English, using a male voice. These messages, such as "Drone four, zone breach!" were synchronized with the visual alarm.
- Mixed AI Assistance: In this condition, the visual highlight and the audio message were synchronized, with both modalities activating simultaneously. This ensured that the visual icon highlight and the spoken message were delivered unison.

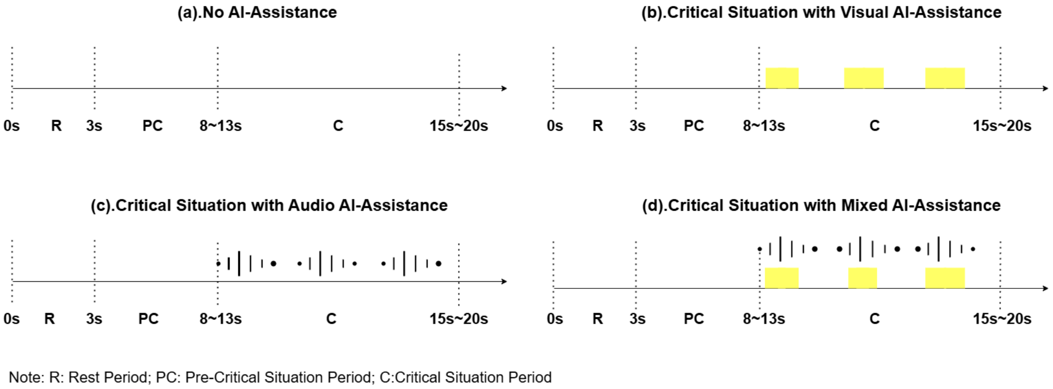
(a).No AI-Assistance
0s
R
3s
PC
8~13s
C
15s~20s
(b).Critical Situation with Visual AI-Assistance
(c).Critical Situation with Audio AI-Assistance
(d).Critical Situation with Mixed AI-Assistance
Note: R: Rest Period; PC: Pre-Critical Situation Period; C:Critical Situation Period

Fig. 2. Timeline of the critical situation task with four AI assistance conditions (no AI, visual AI, audio AI, and mixed AI). Participants began with the no-AI condition as the baseline, followed by randomized AI-assisted conditions. Each trial begins with a 3-second rest period, followed by a pre-critical period (5-10 seconds) where no critical situations occur. During the response period (15-20 seconds), participants must detect a critical situation with or without AI assistance. The visual, audio, and mixed AI assistance conditions are represented by yellow highlights, audio waveforms, or a combination of both, respectively.

As shown in Figure 2, these alarms were triggered at the onset of a critical situation and repeated three times. The AI assistance had an overall accuracy of 80%, with 15 hits, 9 correct rejections, 3 false alarms, and 3 misses across the 30 AI-assisted task trials. In addition to the three AI-assisted alarm modalities, we included a no-AI condition where participants completed the task based solely on their judgment. This condition was presented first to capture participants' natural approach to the task without AI influence. Starting with no assistance allowed us to observe how participants engaged with the task independently, providing a reference point to better understand reliance behaviors as they transitioned into AI-assisted conditions. Importantly, participants completed a practice task prior to the main trials, which included exposure to all AI modalities and the task scenarios. This design helped mitigate potential learning effects by ensuring participants were already familiar with the task structure and AI assistance before starting the no-AI condition.

3.2.2 Apparatus. The experiment utilized a 24-inch desktop screen displaying the monitoring user interface with a resolution of 1920 × 1200 px. Eye movements of the participants were tracked with a Tobii Pro Fusion eye tracker, operating at a sampling rate of 250 Hz. The tracker, positioned beneath the screen and oriented upwards, was adjusted to accommodate each participant. With the optimal distance maintained between 50-65 cm, as verified by the calibration software (Tobii Pro Eye Tracker Manager).

3.2.3 Participants. We recruited 24 participants via local advertisements and word-of-mouth communication at the university. After excluding the data of two participants (one due to recording failure and two for not completing the entire task) the final dataset consisted of 21 participants (10 female, 11 male). Their age ranged from 21 to 34, with a median age of 25. Two participants indicated intermediate to advanced experience with flying drones, while the remaining 19 participants reported no prior experience specifically related to drone operations.

3.2.4 Procedure. Each participant began the study with a detailed introductory video explaining the user interface and tasks (see supplementary materials). Participants provided informed consent before the experiment, acknowledging their participation and understanding of the study's procedures. Participants were explicitly informed that the AI system was not perfect and could make mistakes before the study. However, they were not informed about the nature of the AI assistance (i.e., that it was simulated via a Wizard-of-Oz approach) to provide a more realistic experience of interacting with an AI system. The calibration procedure utilized the Tobii Pro Eye Tracker Manager's five-point calibration method. Once calibrated, participants performed a practice task consisting of ten trials, which incorporated all four critical situations and included tasks both with and without AI assistance across all three AI modalities. Additionally, participants were fully familiarized with the drone locating subtask in the practice. They then undertook four tasks with different AI assistance as detailed in Section 3.2.1. Breaks were put between tasks, during which the eye tracker was recalibrated to maintain data quality. Participants were paid 15€ as time compensation. The study procedure and task were approved by the university's ethics committee.

3.2.5 Measures. To assess user reliance on AI alarms, we utilized the following measures [4]: *user AI agreement rate*, self-reported *perceived reliance*, and user's *gaze behavior*. The user AI agreement rate was calculated as the percentage of trials where the user's response matched the AI alarm. The perceived reliance was gathered from participants' responses on a 5-point Likert scale to the statement, "I relied on the AI-assisted alarm in the previous task trial," collected every five trials.

The captured gaze coordinates were processed into fixations according to two criteria: low dispersion (35 px) and adequate duration (50 ms), using PyGaze [8]. We defined 16 Areas of Interest (AOI), each corresponding to two neighboring icons associated with an alarm or a critical situation. Detailed information about the AOIs are presented in the supplemental material. Gaze behavior was assessed for each AOI related to a critical situation using three commonly used gaze metrics that reflect *verification behavior*, capturing the user's attention and engagement with the displayed information [4]: *fixation count, fixation duration* and *Revisits*. Fixation count represents the total number of fixations on an AOI. A higher count directly signifies greater attentional engagement. Fixation duration, also referred to as overall dwell time, calculates the total time spent looking at an AOI by combining all fixations. Extended duration within the activated AOI could hint at deeper cognitive processing, as participants work to interpret or verify the critical information. The number of revisits counts how many times participants shifted their gaze away and then back to the AOI. This gauges the recurrent attention or potential uncertainty users might have regarding a specific area. The metrics were computed from the start of a critical situation until the end of a trial or until the participant pressed the space bar to indicate they detected the critical situation.

We also analyzed gaze metrics related to scan and search behavior, such as mean saccade amplitude and AOI transition rate, but did not find a clear relationship between these metrics and reliance. We exclude them here for brevity and provide the corresponding results in the Supplementary Material.

We assessed participants' performance in detecting critical situations using several metrics: *recall*, *precision* and *F1 Score*, as commonly used in binary classification [4]. *Response time* captures the speed with which users acknowledged critical situations, with faster times suggesting better performance. For the drone locating subtask, *accuracy* is calculated as the proportion of correctly identified drone positions among all required locations.

3.3 Analysis and Results

For the hypothesis testing analysis below, data normality was first checked using the Shapiro-Wilk Test. For normally distributed data, a one-way ANOVA test was conducted, and statistical significance was determined at a p-value below 0.05 (reported as F-statistic with degrees of freedom). When significant differences were found, Tukey's HSD post-hoc test was applied for pairwise comparisons. For non-normally distributed data, the Kruskal-Wallis test was used (reported as

H-statistic with degrees of freedom). If significant differences were found using the Kruskal-Wallis test, Dunn's test was applied as a post-hoc analysis to identify specific pairwise differences. To ensure statistical validity, all analyses were performed at the individual subject level, where data was first averaged across trials for each participant before statistical testing. This approach guarantees independence of observations, addressing concerns related to pseudoreplication.

3.3.1 User Reliance and Detection Performance Across AI Modalities. We visualized the agreement rate for each user across different AI assistance conditions (see Figure 3 (left)). In all three conditions, the agreement rate exceeded 80%, indicating substantial reliance on AI suggestions. However, no significant differences were observed between the conditions (F(2,63)=0.95, p = .392). For the perceived reliance, the threshold for users who demonstrated their reliance on self-reported score is 3.0. As shown in Figure 3 (right), the average perceived reliance scores in all three AI assistance conditions were above this threshold, suggesting that participants did perceive great reliance on the AI. Nevertheless, no significant differences were found between the conditions (F(2,63)=0.365, p = .696).

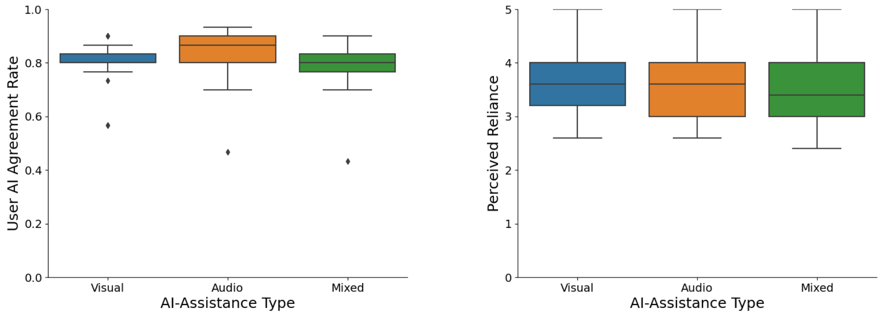

Fig. 3. Summary of agreement reliance and perceived reliance in different AI-assistance conditions

We then investigated gaze behavior and task performance change across different AI assistance, as summarized in Table 1. The test results indicated a significant difference across these conditions for fixation count (H(3) = 11.30, p = .010), revisits (H(3) = 19.665, p < .001), and a near-significant difference for fixation duration (H(3) = 7.67, p = .053). Pairwise comparisons confirmed that the average fixation count rose significantly from 3.15 (no AI) to 5.15 (Visual, p = .035), and 5.02 (Mixed, p = .023). Revisits followed the same trend, increasing from 2.17 (no AI) to 3.74 (Visual, p < .001), 3.78 (Audio, p = .006), and 3.72 (Mixed, p = .002). These results indicate that AI assistance increased user engagement with the alarm areas. Additionally, compared to the User AI agreement and perceived reliance shown in (Figure 3), the variance for all three gaze metrics (fixation count, fixation duration, and revisits) was much higher within each AI assistance condition, indicating that gaze behavior captures a broader range of reliance behaviors.

In addition to gaze behavior, we examined user performance across the different AI assistance conditions, as indicated in the bottom three rows from Table 1. For the F1 score, there was a significant difference across the AI-assisted conditions (H(3) = 11.27, p = .010). Post-hoc comparisons showed that the F1 score for critical situation detection significantly increased from 0.78 in the no-AI condition to 0.90 in the Visual condition (p = .012) and 0.89 in the Mixed condition (p = .049). However, the increase to 0.88 in the Audio condition was not statistically significant (p = .073). Similarly, AI assistance led to significant reductions in response time (H(3) = 192.32, p < .001). The response time decreased from 3.12 seconds in the no-AI condition to 1.63 seconds in the Visual

Table 1. Gaze metrics and task performance by AI assistance condition

	No AI	Visual	Audio	Mixed
Fixation Count	3.15 ± 1.82	5.15 ± 2.49	5.73 ± 4.06	5.02 ± 2.04
Fixation Duration (s)	0.55 ± 0.31	0.89 ± 0.58	0.94 ± 0.65	0.81 ± 0.35
Revisits	2.17 ± 0.99	3.74 ± 1.14	3.78 ± 1.78	3.72 ± 1.33
F1 Score ↑	0.78 ± 0.17	0.90 ± 0.08	0.88 ± 0.07	0.89 ± 0.07
Response Time (s) ↓	3.12 ± 1.91	1.63 ± 1.27	2.63 ± 1.58	1.68 ± 1.24
Drone Map Locating Accuracy ↑	0.78 ± 0.15	0.86 ± 0.11	0.88 ± 0.12	0.88 ± 0.12

condition (p < .001) and 1.68 seconds in the Mixed condition (p < .001), whereas the reduction to 2.63 seconds in the Audio condition was not statistically significant (p = .427). For drone-locating accuracy, there was a significant improvement from 0.78 in the no-AI condition to 0.88 in the Mixed condition (p = .047). However, the increases to 0.86 in the Visual condition and 0.88 in the Audio condition were not statistically significant (p = .394, p = .068, respectively). One thing to note is that while certain AI assistance conditions, such as the Mixed condition, significantly improved user performance in the drone monitoring task compared to the no-AI condition, there were no statistically significant differences among the three AI modalities in terms of F1 score, response time, and drone-locating accuracy (p > .05). These findings suggest that AI assistance generally enhances user performance by improving detection accuracy and reducing response time, while also allowing them to allocate more attention to tracking drone locations on the map. However, the extent of such improvements varies depending on the specific modality of the AI assistance provided.

3.3.2 Gaze Behavior Reveals User Reliance on AI Suggestions. We first analyzed how participants' gaze behavior differed based on their decision to agree or disagree with AI suggestions. As shown in Figure 4, significant differences were found in fixation duration (H(1) = 3.97, p = .046) and revisits (F(1,19) = 6.86, p = .013), while the difference in fixation count approached significance (H(1) = 3.78, p = .054). Specifically, the average fixation count increased from 1.52 when agreeing with the AI to 4.24 when disagreeing (Figure 4 (left)). The average fixation duration increased from 348.31 ms for agreement to 741.79 ms for disagreement (Figure 4 (middle)), and the average number of revisits increased from 1.43 to 2.60 (Figure 4 (right)). These results indicate that participants generally exerted cognitive effort in verifying the AI's recommendations when rejecting them, compared to when they agreed.

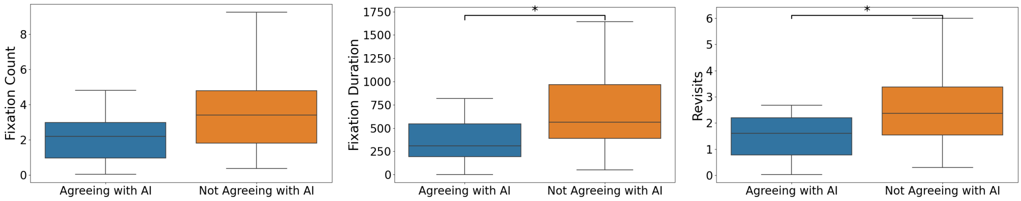

Fig. 4. Summary of Verification Gaze Metrics in Different Human-AI Agreement Conditions. Asterisks indicate statistical significance: * $p < 0.05$, ** $p < 0.01$.

Following these findings, we further analyzed how gaze metrics related to specific detection outcomes, reflecting how reliance levels vary based on verification efforts. As shown in Figure 5, significant differences were found in all gaze metrics: fixation count (H(3) = 12.59, p = .006), fixation duration (H(3) = 13.90, p = .003) and revisits (F(3,84) = 11.92, p < .001). Post-hoc comparisons revealed that FN cases, where participants rejected correct AI suggestions, exhibited significantly higher fixation counts (p = .003) and longer fixation durations (p = .002) than FP cases, where participants followed incorrect AI suggestions. FN cases also had significantly more revisits than all other conditions (p < .001 for TP, TN, and FP). FP cases consistently had the lowest mean gaze metrics, though the differences were not statistically significant. This trend reinforces the contrast between low-verification FP cases and high-verification FN cases. These results suggest that gaze verification levels align with different human reliance on AI: FN cases are associated with extensive verification, indicating potential under-reliance on AI, while FP cases involve minimal verification, reflecting possible over-reliance. In contrast, maintaining a moderate level of verification was more often linked to correct detections, suggesting an appropriate reliance on AI alarms.

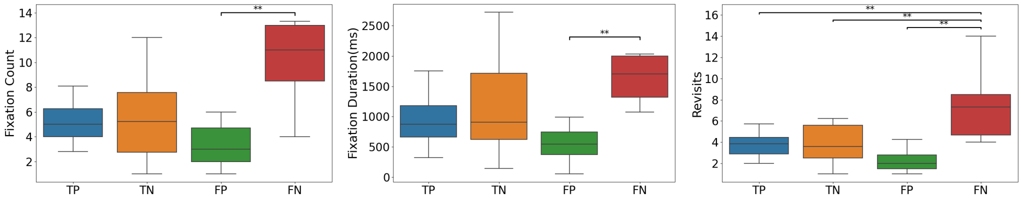

Fig. 5. Summary of Verification Gaze Metrics in Different Detection Results. Asterisks indicate statistical significance: * $p < .05$, ** $p < .01$.

4 RelEYEance: gaze-based clustering approach on user reliance of AI

In this section, we introduce RelEYEance, a gaze-based clustering approach to assess user reliance on AI in this drone monitoring task. We begin by applying unsupervised clustering on existing trial data, revealing three distinct reliance groups based on users' gaze behavior. These groups allow us to interpret patterns of user verification effort and reliance tendencies. Following this, we demonstrate how the clustering model can be adapted for real-time application, enabling continuous assessment of reliance levels during monitoring.

4.1 Feature Processing and Clustering Model

Building on the findings in Section 3.3.2, we used fixation count, fixation duration, and revisit as the input to the model. For each detection trial with an AI alarm, these features were normalized using Z-score normalization, based on the mean and standard deviation calculated from the gaze data across all task trials completed by the same user under the same AI assistance modality. This user-specific baseline helps account for individual differences in verification behavior across trials while preserving variability related to the AI assistance modality. In addition to the multi-feature clustering approach taken here, we conducted an additional analysis using only fixation duration as input. The results, detailed in the supplemental material, highlight that while fixation duration alone captures reliance differences to some extent, it overlooks nuanced behaviors revealed by fixation count and revisits

We chose k-medoids-based partitioning, specifically PAM (Partitioning around Medoids), as the clustering method [9]. To determine the optimal number of clusters, we used a combination of the elbow method, which identifies the point where adding more clusters no longer significantly improves model fit, and silhouette analysis, which measures how well each point fits within its assigned cluster compared to other clusters. This analysis suggested three as the optimal number of clusters. Each task trial is represented by a data point defined by the normalized verification gaze metrics in a high-dimensional space. In this space, the Euclidean distance between trials quantifies how much they differ in their verification efforts of the AI message.

4.2 Interpretation of User Reliance Clusters

For interpreting the clustered results, we compared these clustered groups based on their verification behavior (fixation count, fixation duration, revisits), detection performance (precision, recall, F1 score), reliance measures (user AI agreement, self-reported reliance), and subtask performance (drone map locating accuracy). The group interpretations are as follows:

Group OR (Over-reliance): This group exhibited the lowest verification effort, with the fewest fixation count (3), fixation duration (535 ms), and revisits (3). Precision in detection was the lowest at 0.81, indicating a tendency to follow false alarms blindly. Despite the worst detection performance, the very high user AI agreement (0.87) suggests a false sense of reliability in AI suggestions. The users overly trust AI without sufficient verification.

Group AR (Appropriate reliance): Users demonstrated balanced verification behavior, with moderate fixation counts (7), durations (1207 ms), and revisits (5). Detection performance was the highest, with a precision of 0.91, recall of 0.97, and an F1 score of 0.94, indicating appropriate reliance on AI alarms. User agreement (0.74) reaches the average of one among all three clustered groups, suggesting that users engaged in more cautious verification, leading to better detection results. This group represents the ideal balance between trusting AI and verifying its suggestions.

Group UR (Under-reliance): This group showed excessive verification efforts, with the highest fixation count (11), fixation duration (1,945 ms), and revisits (7). While precision remained relatively high at 0.89, recall dropped to 0.89, leading to the lowest F1 score (0.89) and a high rate of false negatives. User agreement (0.68) was the lowest, reflecting the caution and hesitancy in trusting AI alarms. This group indicates under-reliance, where users spent too much time verifying AI suggestions, leading to missed opportunities and delayed decisions.

Table 2. Summary of gaze metrics and detection performance across groups. All metrics represent mean values except task performance ones and user AI agreement.

Group	Gaze Features			Task Performance							Reliance Measures		
	Fix Cnt	Fix Dur (ms)	Rev	TP	TN	FP	FN	Prec	Rec	F1	User Agmt	Rel Score	Map Acc
OR	3	535	3	141	20	33	5	0.81	0.97	0.88	0.87	3.63	0.90
AR	7	1,207	5	127	45	13	4	0.91	0.97	0.94	0.74	3.64	0.89
UR	11	1,945	7	63	26	8	8	0.89	0.89	0.89	0.68	3.50	0.88

Abbreviations: OR: Over-reliance, AR: Appropriate Reliance, UR: Under-reliance, Fix Cnt: Fixation Count, Fix Dur: Fixation Duration, TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative, Prec: Precision, Rec: Recall, F1: F1 Score, User Agmt: User AI Agreement, Rel Score: Self-Reported Reliance, Map Acc: Drone Map Locating Accuracy.

4.3 Using RelEYEance in Real-Time

We developed an online clustering pipeline to apply the RelEYEance in real-time, detecting user reliance (over-, under-, or appropriate reliance) during monitoring tasks from a stream of gaze data (see Figure 1). The process begins by establishing a baseline for each user's verification gaze behavior, calculated from their prior interactions with AI alarms under the same AI assistance in the first user study. Specifically, we compute the mean and standard deviation of three gaze features—fixation count, fixation duration, and revisits—during these interactions. Once this baseline is set, these values are used for Z-score normalization of the same features in subsequent task trials.

As users perform AI-assisted monitoring tasks, the RelEYEance pipeline tracks their gaze data whenever an AI prompt appears. After each appearance, the three verification gaze features are calculated and normalized before input into the clustering model. To assess the user's inclination toward a specific reliance type, we introduce the **Eye Reliance Score (ERS)**, calculated from a sliding window that stores clustering results from recent monitoring events. For the following study, we used a window size of **10** events; however, this size is adjustable based on the AI system and user behavior. The ERS represents the proportion of each reliance type (over-reliance, under-reliance, and appropriate reliance) observed within this window, with majority voting applied to determine the user's current reliance type. This pipeline enables continuous tracking of user reliance, providing real-time assessments of reliance behavior throughout the task.

5 Evaluation: Real-Time Reliance assessment with RelEYEance

In this section, we implemented RelEYEance to instantly analyse user reliance on AI during the critical situation detection task. Building on the real-time clustering pipeline introduced in the previous section (Section 4.3), this study evaluates whether the model can accurately identify inappropriate reliance (over- or under-reliance) from real-time gaze data. We conducted a second experiment, where participants performed the same drone monitoring task as in Section 3. As we did not observe any significant reliance difference across different alarm modalities in the previous analysis (Section 3.3.1), we focused solely on the mixed AI-assistance condition for this study. User gaze data was continuously tracked, and the clustering model inferred reliance types based on gaze behavior. When over-reliance or under-reliance was detected across multiple trials, interventions, such as visual prompts and audio notices, were delivered to guide users toward more appropriate reliance. We assess the effectiveness of the RelEYEance by evaluating whether reliance types were accurately identified within each group and show that the ERS scores provide a more detailed picture of user reliance over time compared to outcome-based metrics.

5.1 User Reliance Manipulation

AI performance is widely recognized as one of the most critical factors influencing user reliance in automated systems [4, 24]. In this experiment, we kept manipulating the perceived AI reliability in the critical situation detection task using a Wizard-of-Oz approach. Instead of an actual machine learning model, predefined AI performance levels (e.g., 80% accuracy) were assigned to systematically vary the reliability of AI-generated alarms. This controlled setup ensured consistency across participants and allowed us to examine how different perceived AI performance levels influence user reliance. We designed three distinct AI performance levels to influence user reliance:

- **High performance**: AI accuracy was set to 100%, representing a flawless system.
- **Medium performance**: AI accuracy was set to an average of 80%, consistent with the AI used in the first experiment. Every 10 trials included approximately 5 hits, 1 correct rejection, 3 false alarms, and 1 miss.
- **Low performance**: AI acted as a random binary classifier with an average accuracy of 50%. Every 10 trials included approximately 4 hits, 1 correct rejection, 4 false alarms, and 1 miss.

To manipulate user reliance, we employed the priming effect by varying the sequence in which participants experienced different AI performance levels [31]. Participants were exposed to either the high- or low-performance AI system, influencing their reliance patterns in subsequent trials. When participants experienced the high-performance AI, we expected them to develop over-reliance on the AI's alarms due to its flawless early performance. Conversely, when participants interacted with the low-performance AI system, we anticipated that they would underrely on the AI later due to early exposure to errors [28, 29]. In addition to the AI reliability manipulation, we also provided explicit information about the AI's performance to prime participants' expectations. For example, before deploying the high-performance AI system, we displayed a message: "For the next task, we will use a well-developed AI system, which is expected to deliver highly accurate alarms...". Detailed results can be found in the supplemental material.

5.2 Study Design

The experiment employed an over-reliance vs. under-reliance between-subjects factorial design, resulting in two experimental groups. Participants were exposed to varying levels of AI performance across multiple tasks, as summarized in Figure 6.

Twelve participants (six in each experimental group) were initially recruited for the study. One participant from the under-reliance group was excluded due to a bug in the demo that incorrectly issued the intervention at the wrong time. The final dataset included 11 participants, with ages ranging from 22 to 34 years (Median = 24). Among them, four participants were female and seven were male. None of the participants had prior exposure to drone monitoring tasks. The experiment utilized the same eye-tracking setup as in the first study.

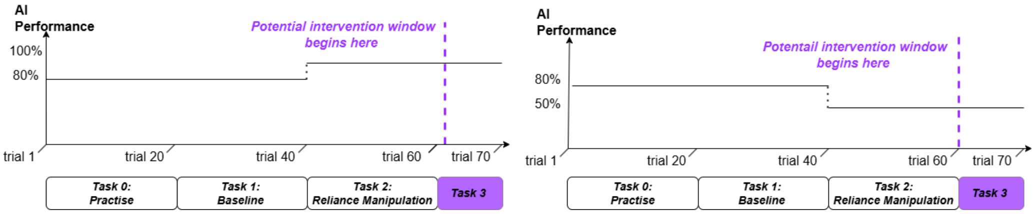

Fig. 6. The experimental procedure includes four tasks. Task 0 is a practice round to help participants familiarize themselves with the task, while Task 1 establishes a baseline by capturing the mean and standard deviation of gaze features. In Task 2, participants' reliance on AI is manipulated, with AI performance increased for the over-reliance group (left) and decreased for the under-reliance group (right), continuing until the end of the experiment. In the final task (Task 3), reliance levels are assessed by RelEYEance for each trial, and intervention would be issued if inappropriate reliance is detected.

Tasks in the experiment followed the same format as in the first study with the AI modality fixed to the mixed condition. As indicated in Figure 6, all participants began with a practice task where the AI operated at medium performance (80% accuracy) to familiarize them with the detection task. Following this, participants completed another task with the same AI performance to establish a baseline for gaze behavior, used for normalization in RelEYEance. After these two initial tasks, participants were randomly assigned to one of two groups:

- **Over-reliance group**: Participants experienced high-performance AI (100% accuracy) during Task 2, intended to induce over-reliance. In Task 3, if over-reliance was detected, an intervention was triggered. This intervention redirected participants to a page displaying the visual message, "Take a moment to verify AI's suggestion," accompanied by an audio prompt: "Please verify the alarm before making your decision..."
- **Under-reliance group**: Participants were assigned to a low-performance AI condition (50% accuracy) in Task 2 to induce under-reliance. In Task 3, interventions were triggered when under-reliance was detected. Participants were redirected to a page with the visual message, "The AI alarm might help you make decisions faster," and an audio prompt: "...use the AI system and make your decision promptly to ensure efficiency."

5.3 Results

We first examined the differences in user gaze behavior in Task 2 between the two groups to confirm that the manipulation was effective. The results showed significant differences in all verification gaze metrics (see supplemental material for statistical details). In the over-reliance group, five out of six participants received interventions triggered by real-time detection of inappropriate reliance with RelEYEance. Similarly, in the under-reliance group, four out of five participants were issued interventions. This shows that the clustering model successfully identified reliance deviations for most participants.

Additionally, we aggregated all participants' ERS for each reliance group across all trials in Task 2 and Task 3. For the over-reliance group (Figure 7 (left)), the high-performance AI quickly gained participants' trust, leading to a high ERS in the initial trials. This indicates that participants showed minimal verification, with ERS values mostly ranging between 0.6 and 1, even after the intervention, suggesting a strong tendency towards over-reliance. In contrast, the under-reliance group's ERS showed greater fluctuation during Task 2 when exposed to low-performance AI. This variability reflects participants' efforts to dynamically adjust their verification behaviors, with average ERS values ranging between 0.2 and 0.8. Notably, a sharp drop in ERS after the intervention indicates a prompt recalibration of visual attention, suggesting the intervention's effectiveness in promoting a more balanced reliance on AI.

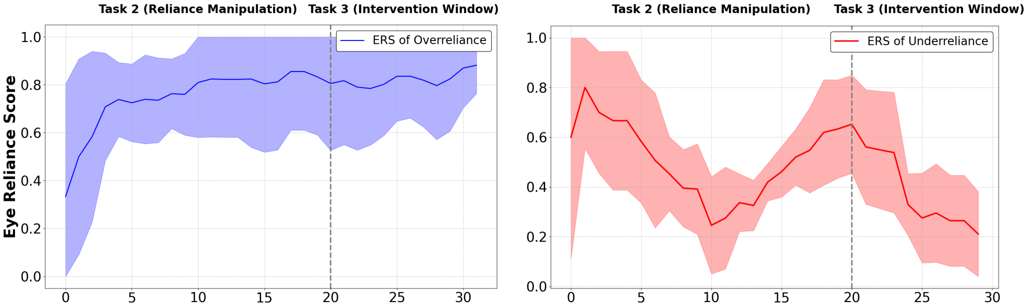

Fig. 7. Aggregated Eye Reliance Score of over-/under-reliance in Task 2 and 3 across all participants in the over-reliance (left) and under-reliance (right) groups. The gray line indicates from which point on an intervention happened when inappropriate reliance was detected.

Building on the aggregated analysis, we examined individual participants' responses to interventions to understand how they adjusted their reliance on AI following targeted feedback. The behavior-based ERS provided more detailed insights into changes in user reliance in particular once the intervention was issued. Figure 8 shows the over- / under-reliance ERS scores along with the User-AI agreement and the users' perceived reliance for two example users from the over- and under-reliance groups. Notably, for the participant in the over-reliance group (Figure 8 (left)), while both user AI agreement and perceived reliance showed no visible change, the over-reliance ERS exhibited a subtle yet immediate decrease, suggesting a slight increase in attention to the alarms. For the participant in the under-reliance group (Figure 8 (right)), the under-reliance ERS displayed the highest variability compared to the other two reliance measures, ranging from a peak of 0.7 before the intervention to a low of 0.2 afterward, at which point no inappropriate reliance is observed anymore. In contrast, User AI Agreement remained steady between 0.4 and 0.6, and perceived reliance consistently stayed low (2.0) during the same period. These visualizations of individual ERS further validate the efficiency of RelEYEance in capturing nuanced shifts in reliance based on users' gaze behavior. We also examined the differences in ERS before and after the intervention for both groups, and the statistical testing results confirmed that the intervention significantly reduced under-reliance ERS in under-reliance group, while its impact on ERS of over-reliance in over-reliance group was more limited (see supplemental material for details).

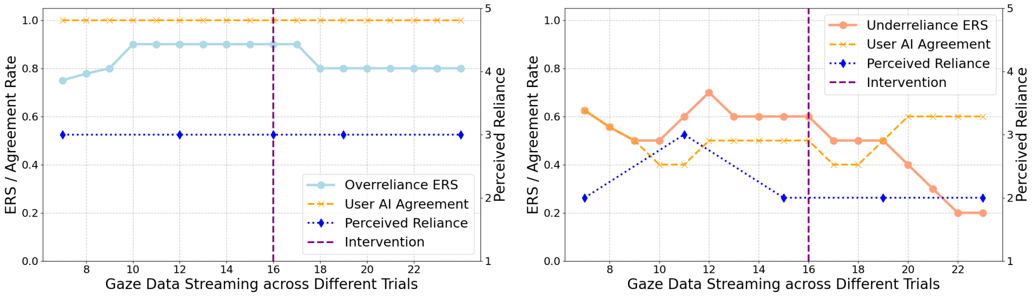

Fig. 8. Individual Eye Reliance Score before and after Intervention

6 Discussion & Conclusion

In this paper we developed the first gaze-based method for detecting user's reliance on AI recommendations in real-time. RelEYEance overcomes limitations of previously used reliance measures, which are mainly outcome-based or rely on self-reports. A decision outcome might not always be available (e.g., when no immediate decision is made) and self-reports are time-consuming and difficult to collect at run-time. Moreover, the correctness of AI recommendations is typically not known in practice and thus it is difficult to judge what would be an appropriate user agreement. Our work shows, that we can use gaze data to detect inappropriate reliance by taking into account the decision-making behavior of users while verifying AI suggestions. In the following, we reflect on the outcome of our studies and discuss the limitations of our proposed approach and how it could be used to help users in calibrating their reliance.

6.1 Eye Gaze as an Indicator of Human Reliance on AI

In our first user study, we established that user verification gaze behavior is a meaningful indicator of reliance on AI, comparable to user-AI agreement and perceived reliance scores across different modalities of AI assistance. The results showed that verification efforts—reflected in fixation count, duration, and revisits—significantly varied when participants agreed or disagreed with the AI's suggested alarms. These gaze differences were very pronounced when examining different detection outcomes. Specifically, reduced verification (lower fixation count, duration, and revisits on AI-alarmed AOIs) signaled over-reliance, leading to more false positives and decreased precision. Conversely, excessive verification (higher values for these metrics) indicated under-reliance, contributing to false negatives and lower recall. Proper verification efforts—indicative of appropriate reliance—resulted in optimal detection performance, positioning them between over-reliance and under-reliance regarding metrics. Based on these findings, we successfully clustered gaze-based reliance into three primary categories: over-reliance, under-reliance, and appropriate reliance, each corresponding with distinct user verification behaviors and detection outcomes.

6.2 Real-Time Reliance Assessment

Moving beyond previous reliance measures like user-AI agreement and perception, which are unsuitable for real-time assessment, our empirical findings motivated the design of an online user reliance clustering pipeline. The second user study demonstrated the practicality of integrating the clustering model: RelEYEance into this pipeline to continuously assess user reliance. By processing gaze data in real-time during AI-assisted monitoring tasks, the RelEYEance model accurately predicted user reliance levels. When inappropriate reliance was detected, tailored interventions were delivered, prompting users to adjust their behavior. In addition, the ERS derived from the clustering results in the RelEYEance pipeline provided a clear temporal profile of reliance dynamics. Analyzing changes in ERS offered a nuanced understanding of how user reliance evolves over time, highlighting shifts in behavior that are not captured by traditional reliance measures.

6.3 Limitations and Future Works

First, the gaze metrics used here are specific to the drone monitoring task. While we expect the general clustering approach to be adaptable to other tasks, the specific gaze metrics indicating over-reliance or under-reliance may vary depending on the task requirements and the user interface layout, influencing visual attention patterns [4]. Consequently, when applying RelEYEance to other setups, it may be necessary to tailor the gaze features to fit the specific context. In future work, we plan to conduct a more comprehensive analysis of gaze features to identify those that can be generalized across different tasks.

Also, through the ERS provided by real-time RelEYEance assessment, we observed substantial differences in the effects of interventions between overreliance and underreliance conditions. This finding suggests the need to adapt intervention strategies based on the type of reliance a user is exhibiting. Future studies could explore personalized intervention strategies that adjust dynamically based on real-time reliance assessments to improve user performance and decision-making in AI-supported tasks.

In addition, while the Wizard-of-Oz setup enabled controlled testing of AI reliance, it may not fully capture user interactions with real AI systems, which could introduce additional uncertainties or variabilities in reliance behavior.

Furthermore, the error rates of the AI system were fixed at specific levels (50%, 80%, and 100%) in this study. Since error rates can have non-linear effects on decision-making and reliance, the generalizability of the findings may be limited. Future work should investigate how varying base error rates influence user reliance and gaze behavior to ensure that the RelEYEance model is robust across diverse reliability levels.

Finally, the small sample size (11 participants) in the second user study limits the generalizability of the findings. While the results demonstrate the feasibility of using RelEYEance for real-time reliance assessment, future work should include larger and more diverse samples to ensure broader applicability and validate the robustness of the approach.

6.4 Conclusion

In summary, this work introduced RelEYEance, a gaze-based approach for real-time reliance detection in AI-assisted monitoring tasks. Through two user studies, we demonstrated that gaze metrics provide meaningful insights into user reliance behaviors, distinguishing between over-reliance, appropriate reliance, and under-reliance. Our results confirmed that verification gaze behavior (fixation count, fixation duration, and revisits) aligns with different levels of reliance and influences detection performance. Moreover, our second study validated the feasibility of real-time reliance monitoring and gaze-based intervention, showing that interventions led to measurable changes in ERS and gaze behavior. These findings highlight the potential of gaze-based adaptive interventions in AI-assisted decision-making and open pathways for future research on real-time reliance calibration strategies.

Privacy and Ethics Statement

This study adheres to ethical guidelines for eye-tracking research, ensuring participant privacy and informed consent. Our findings contribute to AI-assisted decision-making while recognizing the potential risks of user over-reliance on AI. Future applications should consider safeguards against AI misuse and user autonomy erosion.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 389792660 – TRR 248 – CPEC, see https://perspicuous-computing.science and by the Daimler and Benz Foundation (project TITAN - Technologische Intelligenz zur Transformation, Automatisierung und Nutzerorientierung des Justizsystems, grant no. 45-06/24). Y. Wang was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161.

References

[1] Zahra Ashktorab, Michael Desmond, Josh Andres, Michael Muller, Narendra Nath Joshi, Michelle Brachman, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Christine T Wolf, et al. 2021. Ai-assisted human labeling: Batching for efficiency without overreliance. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.

[2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16.

[3] 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. *Proceedings of the ACM on Human-computer Interaction* 5, CSCW1 (2021), 1–21.

[4] 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.

[5] SHIYE CAO, SHICHANG KE, YANYAN MO, LIU ANQI, and CHIEN-MING HUANG. 2023. Eyes are the Windows to AI Reliance: Toward Real-Time Human-AI Reliance Assessment. (2023).

[6] Chun-Wei Chiang and Ming Yin. 2021. You'd better stop! Understanding human reliance on machine learning models under covariate shift. In *Proceedings of the 13th ACM Web Science Conference 2021*. 120–129.

[7] ML Cummings and PJ Mitchell. 2007. Operator scheduling strategies in supervisory control of multiple UAVs. *Aerospace science and technology* 11, 4 (2007), 339–348.

[8] Edwin S Dalmaijer, Sebastiaan Mathôt, and Stefan Van der Stigchel. 2014. PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavior research methods* 46 (2014), 913–921.

[9] Vivek Dhakal, Anna Maria Feit, Per Ola Kristensson, and Antti Oulasvirta. 2018. Observations on typing from 136 million keystrokes. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[10] DJI. 2023. Mavic 3 - Downloads. https://www.dji.com/de/mavic-3/downloads Accessed: August 26, 2023.

[11] FlytBase. 2023. Drone Delivery System: Everything You Need to Know. https://www.flytbase.com/blog/drone-delivery-system Accessed: August 26, 2023.

[12] Christian Fuchs, Clark Borst, Guido CHE de Croon, MM Van Paassen, and Max Mulder. 2014. An ecological approach to the supervisory control of UAV swarms. *International Journal of Micro Air Vehicles* 6, 4 (2014), 211–229.

[13] Marianna Di Gregorio, Marco Romano, Monica Sebillo, Giuliana Vitiello, and Angela Vozella. 2021. Improving human ground control performance in unmanned aerial systems. *Future Internet* 13, 8 (2021), 188.

[14] Zhenyi He, Ruofei Du, and Ken Perlin. 2020. Collabovr: A reconfigurable framework for creative collaboration in virtual reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 542–554.

[15] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 624–635.

[16] Antino Kim, Mochen Yang, and Jingjing Zhang. 2023. When algorithms err: Differential impact of early vs. late errors on users' reliance on algorithms. *ACM Transactions on Computer-Human Interaction* 30, 1 (2023), 1–36.

[17] Richard Kirby. 2009. Development of a real-time performance measurement and feedback system for alpine skiers. *Sports Technology* 2, 1-2 (2009), 43–52.

[18] Spencer C. Kohn, Ewart J. de Visser, Eva Wiese, Yi-Ching Lee, and Tyler H. Shaw. 2021. Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Frontiers in Psychology* 12 (Oct. 2021). doi:10.3389/fpsyg.2021.

604977

[19] Riikka Koulu. 2020. Human control over automation: EU policy and AI ethics. *Eur. J. Legal Stud.* 12 (2020), 9.

[20] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency.* 29–38.

[21] Min Kyung Lee and Katherine Rich. 2021. Who is included in human perceptions of AI?: Trust and perceived fairness around healthcare AI and cultural mistrust. In *Proceedings of the 2021 CHI conference on human factors in computing systems.* 1–14.

[22] David Lindlbauer, Anna Maria Feit, and Otmar Hilliges. 2019. Context-aware online adaptation of mixed reality interfaces. In *Proceedings of the 32nd annual ACM symposium on user interface software and technology.* 147–160.

[23] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[24] Yidu Lu and Nadine Sarter. 2019. Eye tracking: a process-oriented method for inferring trust in automation as a function of priming and system reliability. *IEEE Transactions on Human-Machine Systems* 49, 6 (2019), 560–568.

[25] Zhuoran Lu and Ming Yin. 2021. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–16.

[26] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* 1–19.

[27] P. Madhavan and D. A. Wiegmann. 2007. Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science* 8, 4 (July 2007), 277–301. doi:10.1080/14639220500337708

[28] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121.

[29] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring bias affects mental model formation and user reliance in explainable AI systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces.* 340–350.

[30] Raja Parasuraman and Dietrich H. Manzey. 2010. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 52, 3 (June 2010), 381–410. doi:10.1177/0018720810376055

[31] Samir Passi and Mihaela Vorvoreanu. 2022. Overreliance on AI literature review. *Microsoft Research* (2022).

[32] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2024. Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* 1–17.

[33] Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. *arXiv preprint arXiv:2204.06916* (2022).

[34] Sebastian Schirmer, Christoph Torens, Johann C. Dauer, Jan Baumeister, Bernd Finkbeiner, and Kristin Y. Rozier. [n. d.]. *A Hierarchy of Monitoring Properties for Autonomous Systems.* doi:10.2514/6.2023-2588 arXiv:https://arc.aiaa.org/doi/pdf/10.2514/6.2023-2588

[35] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. 2024. Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* 1–18.

[36] Jakob Schoeffer, Johannes Jakubik, Michael Voessing, Niklas Kuehl, and Gerhard Satzger. 2023. On the interdependence of reliance behavior and accuracy in AI-assisted decision-making. In *HHAI 2023: Augmenting Human Intellect.* IOS Press, 46–59.

[37] Sihan Sun. 2022. *Exploring the Interface to Aid the Operator's Situation Awareness in Supervisory Control of Multiple Drones.* Master's thesis. https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-323073

[38] Kostas M Tsiouris, Vassilios D Tsakanikas, Dimitrios Gatsios, and Dimitrios I Fotiadis. 2020. A review of virtual coaching systems in healthcare: closing the loop with real-time feedback. *Frontiers in Digital Health* 2 (2020), 567502.

[39] Yinsu Zhang, Aakash Yadav, Sarah K Hopko, and Ranjana K Mehta. 2024. In Gaze We Trust: Comparing Eye Tracking, Self-report, and Physiological Indicators of Dynamic Trust during HRI. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction.* 1188–1193.

Received November 2024; revised February 2025; accepted March 2025