When Transparency Backfires: Trust Repair Strategies in Shore-Based Maritime Autonomous Supervision

Zhichen Lu[1,2,3], Zhegong Shangguan[4†], Matthew Stephenson[2,3], Benoit Clement[1,3] and Adriana Tapus[1]

***Abstract*— As Maritime Autonomous Surface Ships (MASS) shift operational oversight to shore-based control centres, maintaining appropriate operator trust after system failures has become a key challenge. It remains unclear how different post-violation feedback strategies affect trust repair and cognitive cost. We conducted a between-subject experiment ($N = 30$) in a simulated shore-based autonomous ship monitoring task to examine how apology, causal explanation, and no feedback influenced trust repair and workload after a system failure. The results showed that apology produced higher post-repair situational trust than both explanation and no feedback, whereas explanation provided no advantage over no feedback. Across all conditions, restored system performance was the primary driver of trust recovery, while dispositional trust significantly predicted post-repair trust. These findings do not imply that transparency is generally harmful. Rather, they suggest that real-time technical explanations must be carefully calibrated to operator expertise, message complexity, timing, and task workload.**

I. Introduction

With the development of autonomous systems, the maritime industry is moving toward unmanned operations. Maritime Autonomous Surface Ships (MASS) are expected to navigate with increasing independence. Shore-based operators in Shore Control Centres (SCCs) (Figure 1) are shifting from traditional bridge watchkeeping to remote supervisory roles [1], [2]. They no longer stand on deck with direct sensory access to the environment. Instead, they monitor vessels through digital interfaces [3], separated by physical distance. This separation changes the nature of the operator's relationship with the system. Trust, rather than direct observation, becomes the foundation for effective collaboration [4].

In this context, trust and system transparency become particularly important for effective human-automation interaction [5]–[7]. Lack of trust leads to disuse, where operators override capable automation. Overtrust leads to misuse, where they fail to intervene when the system errs [8]. In safety-critical maritime operations, both failures carry serious consequences. The problem is compounded after a system malfunction. A single trust violation can trigger a persistent decline in operator confidence, one that undermines subsequent collaboration even after the system resumes reliable performance [9].

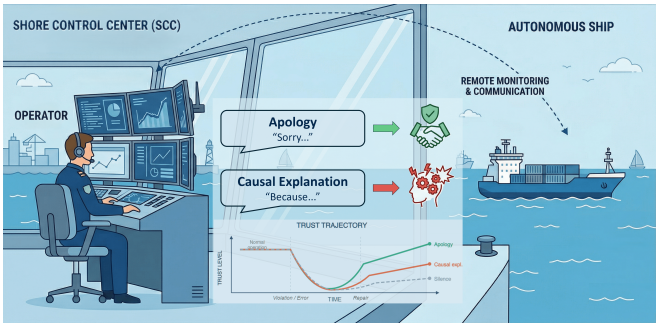

Fig. 1. (Left) The shore-based controller in the Shore Control Centres; (Right) the unmanned merchant vessel; (Center) delayed/incorrect execution by the automatic collision-avoidance navigation system.

A growing body of research in Human-Machine Teaming (HMT) has examined strategies for repairing trust after such violations, including apologies, explanations, denials, and promises [10]–[14]. A recent meta-analysis by Esterwood and Robert [15] found that apologies and explanations were the most commonly tested and among the most effective strategies, although overall effect sizes remained small. However, this literature has predominantly focused on social robots or simplified laboratory tasks, leaving the question of how trust repair strategies transfer to complex supervisory control settings largely unexplored.

The maritime domain presents several characteristics that may modulate the effectiveness of trust repair strategies. First, the operational context is safety-critical: failures in collision avoidance can result in loss of life, environmental damage, and economic harm. Second, shore-based operators are physically remote from the vessel, relying entirely on digital representations of the maritime situation. Third, the emerging SCC workforce, often comprising maritime academy graduates rather than experienced seafarers, may lack the deep domain expertise required to evaluate technical system explanations. This last characteristic is particularly relevant to the debate in the explainable AI (XAI) literature regarding whether increased transparency benefits or harms user trust [1], [16].

The present study addresses this gap by comparing three post-violation feedback conditions: apology, causal explanation, and no feedback, in a simulated shore-based autonomous ship monitoring task. The participants supervised an autonomous vessel performing a collision avoidance

[1]Authors are with Autonomous Systems and Robotics Lab/U2IS, ENSTA, Institut Polytechnique de Paris, Palaiseau, France {zhichen.lu, benoit.clement, adriana.tapus}@ensta.fr

[2]Authors are with Flinders University, Tonsley, Australia matthew.stephenson@flinders.edu.au

[3]Authors are with the French Australian Laboratory for Humans-Machine Teaming (IRL Crossing), Adelaide, Australia

[4]The author is with Cognitive Robotics Lab, Department of Computer Science, University of Manchester, Manchester M13 9PL, U.K. zhegong.shangguan@manchester.ac.uk

maneuver in two trials: a first trial in which the system's avoidance action was deliberately delayed (trust violation) and a second trial in which the system performed correctly (trust restoration opportunity). Between trials, participants in the experimental conditions received either a brief first-person apology or a technical causal explanation for the delayed action. In this study, we define a causal explanation as a post-violation message that identifies the perceived cause of the system failure and describes the corrective system response.

Based on the trust repair literature and the experimental design, three hypotheses were formulated as follows:

- H1 (Trust Repair Effectiveness): A brief apology following the trust violation will produce higher post-repair situational trust than a causal explanation or no feedback.
- H2 (Cognitive Cost of Explanation): The causal explanation condition will impose greater cognitive workload than the apology and no-feedback conditions.
- H3 (Monitoring Behaviour Shift): Participants who receive a causal explanation will exhibit altered system monitoring behaviour in the subsequent trial.

This study makes three contributions. First, it provides empirical evidence on the relative effectiveness of apology versus causal explanation for trust repair in a safety-critical supervisory control context, extending the HMT trust repair literature beyond social robotics. Second, it documents the cognitive costs of technical explanations by demonstrating that these form of causal explanations can increase workload and reduce perceived informativeness, a finding with implications for XAI system design. Third, it highlights the role of dispositional trust as a moderator of trust recovery, reinforcing the need to account for individual differences in trust dynamics.

II. Related Work

A. Trust in Human-Machine Teaming

Trust in automation has been conceptualised as a multidimensional, dynamic attitude that determines an operator's willingness to rely on an automated system under conditions of uncertainty and vulnerability [5]. Mayer et al. [17] proposed an influential model identifying three antecedents of trust including ability, benevolence, and integrity, which, while originally formulated for interpersonal relationships, has been widely adapted to human-machine contexts. Lee and See [5] extended this framework by distinguishing between analytic, analogical, and affective processes through which trust is formed and updated, arguing that trust in automation is shaped not only by observed performance but also by surface-level cues and emotional responses.

Hoff and Bashir [6] advanced the field by proposing a three-layered model that differentiates dispositional trust, situational trust, and learned trust. This model provides a useful framework for understanding why trust responses to the same system event can vary substantially across individuals. Hancock et al. [18] conducted a meta-analysis confirming that robot performance is the strongest predictor of trust in HRI, consistent with the performance-primary hypothesis observed in the present study. In the present study, we operationalise dispositional trust through the TiA Propensity to Trust subscale [19] and situational trust through the Jian trust scale [20], measured at multiple time points to capture trust dynamics.

The Jian trust scale is the most widely used self-report instrument in the trust-in-automation literature, comprising 12 items that measure trust and distrust along a single bipolar dimension. While recent critiques have noted a potential positivity bias and questioned the unidimensionality assumption [21], the scale remains a standard in the field and facilitates comparison with prior work. In the present study, we additionally examine the trust and distrust subscales separately to provide finer-grained insight into the mechanisms of trust change.

B. Trust Violation and Repair in HMT

Trust violations occur when an automated system fails to meet the operator's expectations, leading to a decline in trust. The severity and persistence of trust decline depend on factors including the type of failure (competence- vs. integrity-based) [22], [23], its perceived cause, and the operator's prior trust level [10]. In HMT, several repair strategies have been investigated: apologies, which acknowledge the failure and express regret; explanations, which provide causal accounts for the failure; denials, which attribute the failure to external factors; and promises, which commit to improved future performance [24]. Lewicki and Brinsfield [25] provided a comprehensive taxonomy of trust repair strategies, distinguishing short-term verbal responses from longer-term structural rearrangements.

Sebo et al. [11] found that the effectiveness of repair strategies depends on the framing of the violation: in their competitive game paradigm, participants interacting with a robot that used integrity-based framing and denial showed increased behavioural retaliation. More recently, Esterwood and Robert [15] conducted a meta-analysis of 22 trust repair studies (N = 3,763) that included a no-repair control condition. Their analysis revealed three key findings: repair strategies have differential effectiveness; the overall impact on trust restoration is marginal, with a small effect size; and apologies and explanations are the most effective among tested strategies.

However, the existing literature has two notable gaps. First, most studies employ social robot paradigms or simplified decision-making tasks that do not capture the complexity and time pressure of safety-critical supervisory control. Second, the cognitive costs of trust repair strategies have received limited attention. While explanations are intuitively appealing because they address the informational deficit created by the violation, they also impose processing demands on the operator, demands that may compete with the ongoing monitoring task. The present study addresses both gaps by testing trust repair strategies in a maritime supervisory

control simulation and by incorporating cognitive workload (NASA-TLX [26]) as a dependent variable alongside trust.

C. Transparency, Explainability, and Complexity Tradeoff

The relationship between system transparency and user trust is not straightforwardly positive. While early work in explainable AI suggested that providing explanations enhances trust and acceptance [27], [28], subsequent research has revealed a more nuanced picture. Zhang et al. [16] demonstrated that confidence indicators and explanations can improve trust calibration in AI-assisted decision-making, but only when the explanations are interpretable to the user. When explanations exceed the user's expertise or deviate from their existing mental model, they can paradoxically increase uncertainty and reduce trust, a phenomenon described as the transparency-complexity tradeoff.

In the maritime autonomy context, Veitch et al. [2] argued for human-centred XAI approaches that account for the end user's expertise level, noting that shore-based operators may have different explainability needs than system developers. Ramos et al. [1] similarly emphasised the importance of matching information presentation to the operator's cognitive capacity in MASS operations. These perspectives motivate the present study's focus on whether a technical causal explanation, while informationally richer, may actually undermine trust and increase workload for operators with foundational but non-expert domain knowledge.

D. Human Factors in Maritime Autonomous Systems

The transition toward MASS operations is creating a new class of operators: shore-based supervisors who monitor multiple vessels remotely and intervene when autonomous systems encounter situations beyond their operational envelope [3], [29]. Research on human factors in MASS has examined situational awareness challenges [30], [31], interface design requirements for SCCs [29], and risk assessment frameworks that incorporate human-system interaction [1], [32]. Trust has been identified as a critical factor in this context: operators must trust the autonomous navigation system sufficiently to allow it to operate independently, while maintaining enough vigilance to detect and respond to failures [2].

However, empirical studies of trust dynamics in MASS supervisory control remain scarce. Most existing work has relied on interviews, surveys, or theoretical analyses rather than experimental paradigms that manipulate trust-relevant variables. The present study contributes to this emerging field by providing controlled experimental evidence on how different communication strategies following a system failure affect trust, workload, and monitoring behaviour in a simulated SCC environment.

III. Methodology

A. Participants

A total of 30 participants were enrolled in the present study, comprising 11 females and 19 males, with a mean age of 23.8 years (SD = 1.27). All participants had received formal education in basic maritime theory and had prior experience with ship bridge simulator operations. Moreover, all had completed at least one semester of coursework related to seafaring or ship operation. Thus, they were expected to have a basic level of domain knowledge relevant to navigational tasks, including the COLREGs, collision-avoidance practices, and ECDIS interpretation. However, none of the participants held a professional seafarer certificate. This sample profile is consistent with the domain-informed but non-expert participant category commonly adopted in maritime human-robot interaction research. Participation was entirely voluntary. Prior to the experiment, all participants read and signed an informed consent form and were explicitly informed that they could withdraw from the study at any time without penalty. Each participant completed a single individual experimental session lasting approximately 20 minutes. Before arrival at the laboratory, participants were randomly assigned to one of three between-subject experimental conditions, with 10 participants in each condition.

B. Experimental System

The study employed a real-time maritime encounter simulation system developed in Unity. The experiments were conducted on a machine equipped with an Intel Core i9-9900K processor, an NVIDIA GeForce RTX 2080 Ti, and a 34-inch monitor (Figure 2). The encounter scenario was a crossing-from-starboard situation, in which the target vessel approached from the own vessel's starboard bow. Under COLREGs, the own vessel was designated the give-way vessel and was required to execute a collision avoidance manoeuvre. Vessel motion was governed by a deterministic kinematic controller compliant with COLREGs, which initialised the encounter at a fixed crossing geometry and executed rule-based avoidance behaviour, thereby ensuring consistent encounter geometry across participants and conditions. Navigation behaviour was held constant across conditions to isolate the effect of interface design on human perception. With the exception of own-ship manoeuvre timing, all visual elements, including vessel geometry, sea state, lighting conditions, and target vessel trajectory, were identical across the two video stimuli. Own-ship manoeuvre timing thus constituted the sole visual manipulation variable.

The system comprised four integrated components:

a) Bridge View Display: Presented a first-person perspective from the ship's bridge.

b) ECDIS: Continuously displayed and updated the positions and heading vectors of both vessels. When participants activated the optional explanation overlay via a toggle switch, the interface additionally showed the 40-second predicted tracks of both vessels (with and without yaw rate), a Closest Point of Approach (CPA) marker, and a line connecting the predicted positions of the two vessels at the moment of closest approach. The overlay remained available throughout the scenario and could be enabled or disabled at any time.

c) Agent Explanation Panel: A structured text panel that continuously updated its analysis of the current encounter situation across four subsections: (1) scene classi-

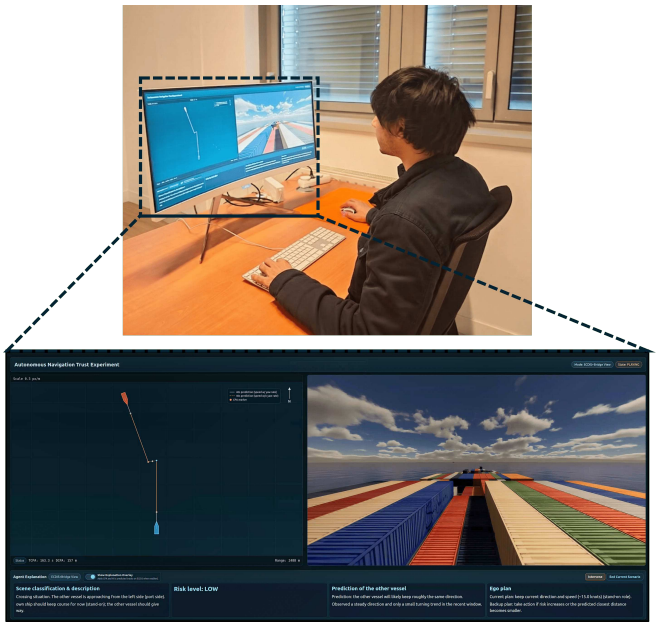

Fig. 2. (Top) Experimental setup of the system; (Bottom) Screenshot of the capture interface.

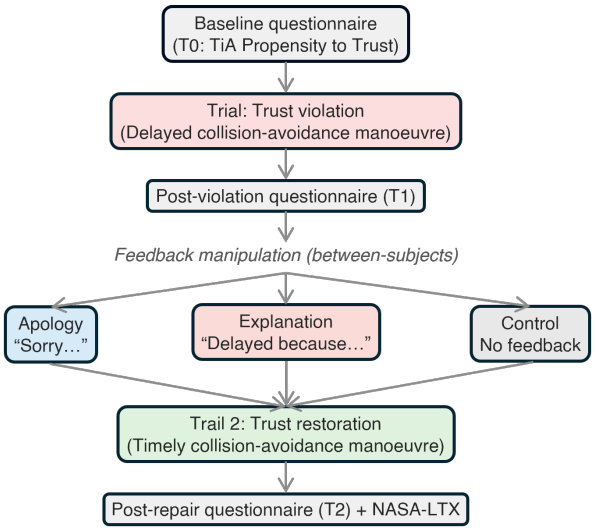

Fig. 3. Experimental Procedure

fication and role assignment based on COLREGs (give-way or stand-on); (2) current risk level (LOW, MED, or HIGH) computed from DCPA and TCPA thresholds; (3) prediction of target vessel behaviour derived from observed yaw rate; (4) the own vessel's current plan and contingency action.

d) Intervene Button: Allowed participants to press at any moment they judged the vessel's avoidance turn to be insufficiently timely, thereby recording an explicit operator takeover intent event.

C. Conditions

The scenario content experienced under the three experimental conditions was identical. The between-group manipulation concerned the feedback presented by the system interface at the moment when the own ship began to turn after the participant pressed the Intervene button. Specifically, a modal overlay appeared on the ECDIS panel. The feedback message was phrased in the first person, consistent with the narrative style used in the agent explanation panel, and disappeared automatically after 10 seconds. The 10-second duration was selected to ensure that participants had sufficient time to read the message while limiting obstruction of the ECDIS display during the ongoing monitoring task.

a) Control: The system initiated the turning maneuver after confirming the participant's button press, but no feedback message was displayed.

b) Apology: The following message was presented: Sorry, I initiated the avoidance maneuver later than I should have. I have now resumed normal operation.

c) Explanation: The following message was presented: I delayed the avoidance maneuver because the AIS reports from the target vessel were delayed, which led me to underestimate the closing speed. I have now switched to radar priority mode and resumed normal operation. This message attributed the delay to an external technical cause, namely AIS signal latency.

The three-condition design was chosen to isolate the effects of two theoretically distinct repair strategies against a baseline without feedback: a concise affective acknowledgment and a causal technical account. We did not include a combined apology-plus-explanation condition in the present experiment because such a hybrid message would introduce an additional compound mechanism and increase message length, making it difficult to attribute effects to apology, explanation, or their combination. Future work should explicitly compare hybrid repair messages with the same length and complexity.

D. Procedure

At the beginning of the experiment, the experimenter provided a standardized verbal introduction to the study background and explained that participants would assume the role of a shore-based operator responsible for monitoring the navigation of an autonomous vessel. Participants then completed a practice trial that was not included in the formal experiment. During this stage, the experimenter introduced the different interface components and answered any questions concerning the display. After the practice trial, a preparation page was presented, and participants initiated the formal experiment themselves by clicking the start button and first completing the baseline questionnaire. After each experimental scenario, participants immediately completed the corresponding post-scenario questionnaire before proceeding to the next trial. After both scenarios had been completed, participants filled out the NASA-TLX scale and the final questionnaire (Figure 3).

The formal experiment consisted of two starboard crossing scenarios presented in a fixed order. The experiment included two trials. In the first trial, the system delayed the own ship's avoidance maneuver. In the second trial, the timing of the turning maneuver was determined by selecting a random

moment within a 10-second window before or after the time at which the participant had pressed the Intervene button in the first trial. This order was kept identical for all participants in order to preserve the trust repair manipulation and to align with the present study's design logic concerning the trajectory of trust change. The entire experimental procedure lasted approximately 20 minutes.

IV. Results

A. Preliminary Checks

All multi-item scales (Figure 4) demonstrated acceptable internal consistency (Jian trust scale: α = .86-.92; NASA-TLX: α = .70; TiA Propensity to Trust: α = .74). A one-way ANOVA on the TiA Propensity to Trust composite (T0) yielded a marginally non-significant omnibus effect, $F(2, 27) = 3.05$, p = .064, with the apology group reporting numerically higher dispositional trust (M = 5.55) than the control (M = 4.81) and explanation groups (M = 4.57). T0 was therefore entered as a covariate in subsequent analyses. Post-violation situational trust (T1), measured before any repair manipulation, did not differ across groups, $F(2, 27) = 2.27$, p = .123, confirming that the three groups experienced the delayed-manoeuvre scenario equivalently.

B. Apology Produced the Highest Post-Repair Trust

H1 was partially supported. Shapiro-Wilk tests indicated that the apology group departed significantly from normality on the trust-positive subscale ($W = .78, p = .009$), while the control and explanation groups did not ($Ws \geq .92, ps \geq .34$); non-parametric pairwise tests were therefore used to corroborate parametric comparisons. The omnibus ANOVA on post-repair situational trust (T2 Jian) was significant, $F(2, 27) = 3.69, p = .039, \eta^2 = .22$, confirmed by non-parametric tests. The apology group reported the highest trust ($M = 6.14, SD = 0.84$), significantly exceeding both the control ($M = 5.21; U = 14.0, p = .009$) and explanation groups ($M = 5.03; U = 85.0, p = .005$). Crucially, the control and explanation groups did not differ ($p = .758$), indicating that the technical explanation provided no trust repair advantage over receiving no feedback at all. This effect was driven by the trust-positive subscale ($F = 4.99, p = .015$); the distrust subscale did not reach significance.

However, when dispositional trust (T0) was entered as a covariate, the condition effect was attenuated to non-significance, $F(2, 26) = 1.26, p = .302$, while T0 was a significant predictor ($F = 6.62, p = .016, \beta = 0.48$). The direction of adjusted means nonetheless remained consistent with the unadjusted pattern. Dispositional trust correlated more strongly with T2 ($r = .57, p = .001$) than with T1 ($r = .38, p = .044$), suggesting that its influence grew as participants moved from a violation-only to a violation-plus-recovery experience.

Paired-samples t-tests revealed that trust increased significantly from T1 to T2 in all conditions: control ($t(9) = 3.83, p = .005, d = 1.28$), apology ($d = 0.82$), and explanation ($d = 0.70$, marginally significant). The magnitude of this within-participant gain did not differ across groups ($p = .617$), suggesting that restored system performance in Trial 2 was the primary driver of trust recovery, with the repair message playing a secondary, modulatory role. H1 was thus partially supported: apology outperformed both other conditions in absolute trust, but the uniform within-group recovery indicates that behavioural evidence of system competence mattered most.

C. Explanation Imposed Substantial Cognitive Cost

H2 was strongly supported. Shapiro-Wilk tests revealed significant departures from normality in the apology group ($W = .57, p < .001$) and the explanation group ($W = .81, p = .019$); the control group was approximately normal ($W = .91, p = .346$). Given violations in two of three groups, a Kruskal-Wallis test was conducted alongside the ANOVA to verify robustness. The ANOVA on overall NASA-TLX was significant with a large effect, $F(2, 27) = 7.88, p = .002, \eta^2 = .38$. The explanation group reported the highest workload ($M = 3.75, SD = 1.05$), significantly exceeding both the control ($M = 2.72; p = .019$) and apology groups ($M = 2.38; p = .019$). This pattern was consistent across all six subscales, with frustration ($F = 11.07, p < .001$) and mental demand ($F = 6.76, p = .004$) showing the most pronounced differences. The explanation group also rated their own performance lowest ($M = 4.60$ vs. control 6.00 and apology 6.40).

Converging evidence emerged from the post-experiment questionnaire. The explanation group rated information sufficiency significantly lower than both other groups ($F = 10.56, p < .001$; explanation $M = 3.20$ vs. apology $M = 6.10$), and perceived the system's status information as less helpful ($F = 11.16, p < .001$)-a noteworthy reversal, given that this group received strictly more information than the others.

D. Partial Evidence for Altered Monitoring Behaviour

H3 was not statistically supported and should be interpreted as exploratory. Shapiro-Wilk tests indicated that the explanation group's ECDIS overlay viewing duration departed significantly from normality ($W = .65, p < .001$), while the control and apology groups did not ($Ws \geq .91, ps \geq .29$); a Kruskal-Wallis test was therefore conducted alongside the ANOVA. ECDIS overlay viewing duration in Trial 2 showed a trend toward reduced usage in the explanation group ($M = 36.9s, SD = 54.8$) compared with the control ($M = 90.7s$) and apology groups ($M = 77.7s$), though this difference only approached significance, $F(2, 27) = 2.78, p = .080$. Overlay toggle counts did not differ across conditions in either trial. Intervene-button presses during Trial 1 showed a descriptive trend toward higher counts in the explanation group ($M = 3.50$ vs. control 0.33 and apology 0.20), but this was driven by two outlying participants and did not reach significance ($H = 3.39, p = .184$). Therefore, the behavioural monitoring results should be interpreted as exploratory rather than confirmatory evidence for altered monitoring behaviour.

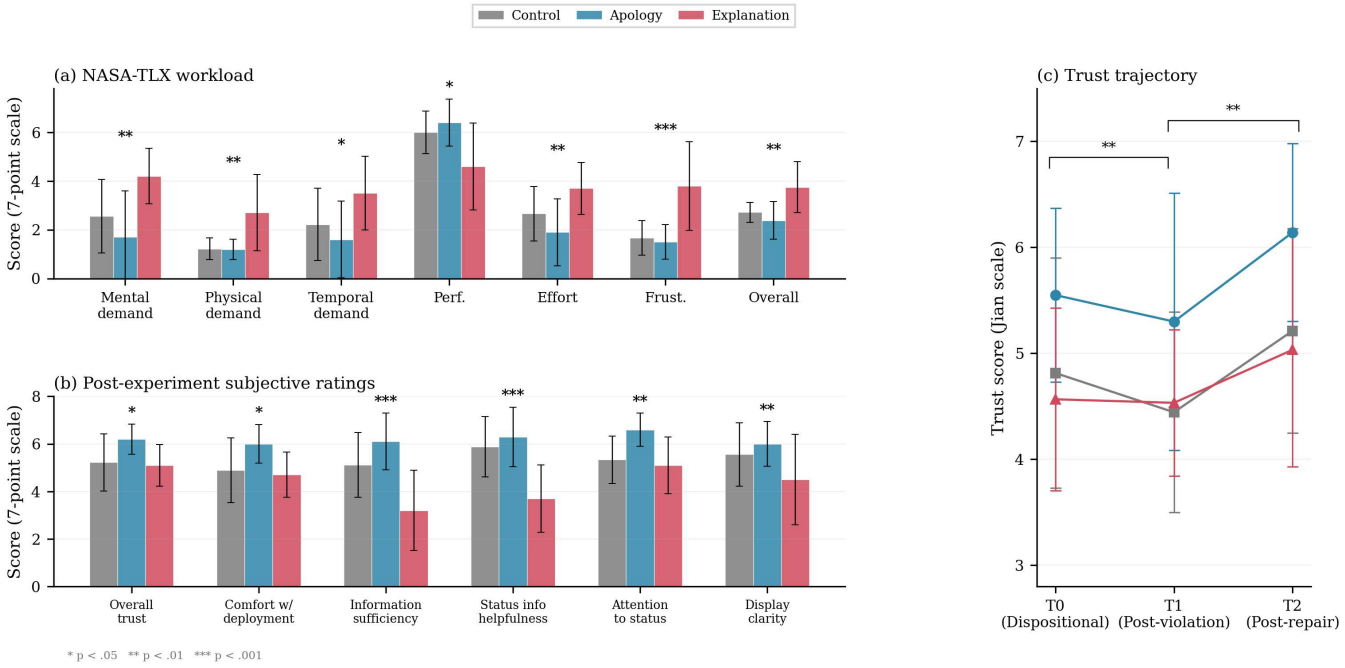

Fig. 4. Group-level comparisons across the three feedback conditions. (a) NASA-TLX workload ratings after the trust-repair trial, (b) post-experiment subjective evaluations of the system, and (c) trust trajectory from dispositional trust (T0) to post-violation trust (T1) and post-repair trust (T2). Bars and points indicate condition means, and error bars represent standard deviations. Asterisks denote significant between-group differences (*$p < .05$, **$p < .01$, ***$p < .001$). The apology condition generally yielded higher post-repair trust and more favorable subjective evaluations than the explanation and control conditions.

V. Discussion

This study examined whether post-violation feedback contentapology, causal explanation, or no feedback-modulated trust repair, cognitive workload, and monitoring behaviour in a shore-based autonomous ship supervisory control scenario. Three principal findings emerged, each having implications for the design of human-autonomy communication in safety-critical maritime operations.

A. The Simplicity Advantage of Apology

The most salient finding was that a brief, first-person apology produced the highest post-repair situational trust, significantly exceeding both the causal explanation and the no-feedback control, which did not differ from each other. This ordering contradicts the common theoretical expectation-rooted in the attribution-based trust repair framework [17] that providing a causal account should be more effective because it reduces uncertainty about the failure and signals corrective capacity.

Two factors may explain the apology's advantage. First, the causal explanation introduced domain-specific technical content (AIS signal latency, radar-priority switching) that participants-who possessed foundational maritime knowledge but lacked professional operational experience-may have found difficult to evaluate. Rather than reducing uncertainty, the explanation may have created new uncertainty about system components the participant was not previously aware of. This interpretation aligns with the transparency-complexity tradeoff identified in the explainable AI literature: explanations that exceed the recipient's evaluative capacity can paradoxically undermine trust calibration [16], [27].

Second, the apology succeeded precisely because of its simplicity. By acknowledging the failure without requiring the participant to construct a new mental model of the system's internal reasoning, the apology preserved cognitive resources for the monitoring task itself. This is consistent with the distinction drawn by Lee and See [5] between calibration-oriented trust (requiring deep understanding) and affectively grounded trust (requiring social cues of reliability). In a time-pressured supervisory setting, affective repair may be more ecologically appropriate than analytical repair. The finding also aligns with the recent meta-analysis by Esterwood and Robert [15], which identified apologies as among the most effective trust repair strategies, although with a small overall effect size, suggesting that even effective verbal strategies have limited reach compared to demonstrated system competence.

B. The Cognitive Cost of Real-Time Technical Explanation

The workload findings deserve particular emphasis. The explanation group reported significantly higher workload across every NASA-TLX subscale, with an overall effect size of $\eta^2 = .38$, the largest effect observed in this study. Frustration was the most discriminating subscale, suggesting that the explanation was not merely effortful to process but actively aversive. The explanation group also rated the system's status information as less helpful and the provided information as less sufficient-a striking reversal, given that this group received strictly more information than the others.

This pattern is consistent with information overload theory: when additional information exceeds the operator's processing capacity or deviates from their existing schema [33], it degrades rather than enhances perceived informativeness. The result carries a specific design lesson: explanation modality and timing matter as much as content. A modal overlay that interrupts the monitoring workflow and introduces unfamiliar technical concepts may detract from the operator's situational awareness rather than enhancing it. These findings resonate with the broader call in the HMT community for calibrated transparency-ensuring that system explanations are matched to the recipient's expertise and the demands of the operational context [9].

C. System Performance as the Primary Driver of Trust Recovery

All three groups showed significant or near-significant trust increases from T1 to T2, with medium-to-large effect sizes within the group (d = 0.70-1.28). Crucially, the magnitude of this gain did not differ across conditions. This suggests that the primary driver of trust recovery was the system's restored behavioural performance in Trial 2, i.e., timely execution of the avoidance manoeuvre rather than the verbal content of the repair message. This is consistent with the performance primary hypothesis: in safety-critical domains, operator trust is anchored to observed system competence, and verbal communication plays a secondary, modulatory role [9]. The apology's advantage over the other conditions should therefore be understood not as a demonstration that words outweigh actions, but rather as evidence that the right words can amplify the trust-restorative effect of restored competence, while the wrong words (or overly complex ones) can dampen it.

D. The Moderating Role of Dispositional Trust

The ANCOVA revealed that dispositional trust (T0) was a significant predictor of post-repair situational trust (β = 0.48), and that after controlling for it, the between-group condition effect was attenuated. The strengthening correlation from T0-T1 (r = .38) to T0-T2 (r = .57) suggests that dispositional trust's influence grew as participants moved from a violation only to a violation-plus-recovery experience. This is consistent with multilevel trust models [6], in which dispositional trust serves as a stable moderator of how situational events are interpreted. Participants who entered the study with higher trust propensity may have been more willing to give the system a second chance after the violation, regardless of the specific repair message. Methodologically, this finding highlights the importance of measuring and controlling for individual differences in automation trust propensity in small sample studies, where random assignment may not fully eliminate baseline imbalances.

E. Implications for Maritime Autonomy Design

From a design standpoint, the results suggest that autonomous ship systems should favour concise, empathic acknowledgements of error over detailed technical justifications when communicating with shore-based supervisors. This recommendation is especially pertinent for operators who possess foundational domain knowledge but are not deeply versed in system internals, a profile likely to characterise many early-career shore-based controllers as the maritime industry transitions toward autonomous operations. Detailed causal explanations might be reserved for post-incident debriefings or experienced engineers, rather than delivered in real time during active monitoring.

F. Limitations and Future Directions

Several limitations constrain the generalisability of these findings. The sample size (N = 30) provided limited statistical power, particularly for detecting moderate interaction effects. The marginal baseline imbalance in dispositional trust, although controlled via ANCOVA, may have contributed to the attenuation of the condition effect. Future studies should employ larger samples and stratified randomisation on dispositional trust. Additionally, no comprehension check was administered to verify whether participants in the explanation condition correctly understood the technical content of the causal explanation. The absence of such a manipulation check limits our ability to determine whether the explanation's ineffectiveness was due to information overload, misunderstanding, or a lack of engagement with the message content.

A further limitation concerns the non-equivalence of the apology and explanation stimuli. The explanation message differed from the apology not only in repair strategy but also in length, technical specificity, semantic complexity, and likely processing demand. Therefore, the present design does not allow us to determine why the explanation condition underperformed, whether due to the causal explanation itself, the amount of information provided, the use of technical terminology, or its presentation as a real-time message during the monitoring task. The results should therefore be interpreted as evidence against this specific form of real-time technical explanation for non-expert operators, rather than against explanations or transparency more broadly.

Although all participants experienced the same failure-then-recovery sequence, the between-group comparisons among apology, explanation, and control remain interpretable. However, the within-participant increase from T1 to T2 should not be attributed solely to trust repair. It likely reflects a combination of restored system performance, increased familiarity with the task and interface, and possible learning effects between the first and second formal trial. Participants were maritime students rather than professional seafarers; while representative of the emerging shore-based controller workforce, findings may not generalise to expert operators whose richer domain models could enable deeper processing of causal explanations. Finally, only one type of causal explanation was tested. Alternative framings-simpler language, combined apology plus explanation, or visual rather than textual explanations remain to be explored.

VI. Conclusion

In a simulated shore-based autonomous ship monitoring task, a brief apology following a delayed collision-avoidance manoeuvre produced higher post-repair trust than either a technical causal explanation or no feedback. The causal explanation, while intended to enhance transparency, instead elevated cognitive workload, increased frustration, and reduced perceived informativeness. Meanwhile, the uniform within-group trust recovery observed across all conditions underscores that demonstrated system competence remains the primary foundation for trust repair, with verbal strategies serving a modulatory function. Dispositional trust emerged as a significant predictor of trust outcomes, reinforcing the importance of individual differences in trust dynamics. These findings contribute to the growing evidence base for Human-Machine Teaming design and suggest that transparency interventions must be calibrated to the recipient's expertise and the demands of the operational context.

References

[1] M. A. Ramos, C. A. Thieme, I. B. Utne, and E. Bjørnsen, "Human-system interaction in the context of maritime autonomous surface ships," *Safety Science*, vol. 142, p. 105383, 2021.

[2] E. Veitch and O. A. Alsos, "Human-centered explainable artificial intelligence for marine autonomous surface vehicles," *Journal of Marine Science and Engineering*, vol. 9, no. 11, p. 1227, 2021.

[3] T. Porathe, J. Prison, and Y. Man, "Situation awareness in remote control centres for unmanned ships," *Proceedings of Human Factors in Ship Design & Operation*, 2014.

[4] T. B. Sheridan, *Telerobotics, Automation, and Human Supervisory Control*. Cambridge, MA: MIT Press, 1992.

[5] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[6] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human Factors*, vol. 57, no. 3, pp. 407–434, 2015.

[7] S. Le and S. Zhegong, "The moment that the driver takes over: Examining trust in full self-driving in a naturalistic and sequential approach," in *Proceedings of the 22nd European Conference on Computer-Supported Cooperative Work*, no. 10.48340/ecscw2024_ep04, 2024.

[8] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors*, vol. 39, no. 2, pp. 230–253, 1997.

[9] E. J. de Visser, M. M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx, "Towards a theory of longitudinal trust calibration in human–robot teams," *International Journal of Social Robotics*, vol. 12, pp. 459–478, 2020.

[10] C. Esterwood and L. P. Robert, "Having the right attitude: How attitude impacts trust repair in human–robot interaction," *Computers in Human Behavior*, vol. 148, p. 107881, 2023.

[11] S. S. Sebo, P. Krishnamurthi, and B. Scassellati, ""I don't believe you": Investigating the effects of robot trust violation and repair," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2019, pp. 57–65.

[12] Y. Liu, Z. Shangguan, A. Tapus, S. Safin, F. Détienne, and E. Lecolinet, "Silicone-based haptic interfaces: Enhancing multimodal interactions through pneumatic tactile feedback," in *2024 16th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*. IEEE, 2024, pp. 140–143.

[13] Z. Shangguan, Y. Liu, L. Song, T. Li, and A. Tapus, "Using a pneumatic tactile steering wheel to enhance the multi-modal takeover request in smart vehicle," in *International Conference on Social Robotics*. Springer Nature Singapore Singapore, 2024, pp. 122–132.

[14] Y. Liu, Z. Shangguan, A. Tapus, S. Safin, F. Détienne, and E. Lecolinet, "Enhancing safety and user experience in automated driving: A multimodal comparison of pneumatic and vibrotactile haptic feedback takeover scenarios," in *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2025, pp. 622–628.

[15] C. Esterwood and L. P. Robert, "Repairing trust in robots?: A meta-analysis of HRI trust repair studies with a no-repair condition," in *Proceedings of the 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2025)*, Melbourne, Australia, 2025, march 4–6, 2025.

[16] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making," *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 295–305, 2020.

[17] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Academy of Management Review*, vol. 20, no. 3, pp. 709–734, 1995.

[18] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors*, vol. 53, no. 5, pp. 517–527, 2011.

[19] M. Körber, "Theoretical considerations and development of a questionnaire to measure trust in automation," in *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*. Springer, 2019, pp. 13–30.

[20] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International Journal of Cognitive Ergonomics*, vol. 4, no. 1, pp. 53–71, 2000.

[21] R. S. Gutzwiller, E. K. Chiou, S. D. Craig, C. M. Lewis, G. J. Lematta, and C.-P. Hsiung, "Positive bias in the "trust in automated systems survey"? an examination of the Jian et al. (2000) scale," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 63, no. 1, 2019, pp. 217–221.

[22] P. H. Kim, D. L. Ferrin, C. D. Cooper, and K. T. Dirks, "When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence- vs. integrity-based trust violation," *Organizational Behavior and Human Decision Processes*, vol. 94, no. 2, pp. 49–65, 2004.

[23] S. Tolmeijer, A. Weiss, M. Hanheide, F. Lindner, T. M. Powers, C. Dixon, and M. L. Tielman, "Taxonomy of trust-relevant failures and mitigation strategies," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2020, pp. 3–12.

[24] K. Rogers, R. J. A. Webber, and A. Howard, "Lying about lying: Examining trust repair strategies after robot deception in a high-stakes HRI scenario," in *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, pp. 706–710.

[25] R. J. Lewicki and C. Brinsfield, "Trust repair," *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 4, pp. 287–313, 2017.

[26] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," in *Advances in Psychology*. Elsevier, 1988, vol. 52, pp. 139–183.

[27] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[28] Z. Shangguan, X. Hei, F. Li, C. Yu, S. Song, J. Zhao, A. Cangelosi, and A. Tapus, "Learning from human conversations: A seq2seq based multi-modal robot facial expression reaction framework in hri," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 7261–7268.

[29] H. Dybvik, E. Veitch, and M. Steinert, "Design for resilient human-system interaction in autonomy: The case of a shore control centre for unmanned ships," in *Proceedings of the International Conference on Engineering Design (ICED21)*, Gothenburg, Sweden, 2021.

[30] S. C. Mallam, S. Nazir, and S. K. Renganayagalu, "Rethinking maritime education, training, and operations in the digital era: Applications for emerging immersive technologies," *Journal of Marine Science and Engineering*, vol. 8, no. 11, p. 848, 2020.

[31] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human Factors*, vol. 37, no. 1, pp. 32–64, 1995.

[32] Z. Shangguan, X. Han, Y. E. Mrhasli, N. Lyu, and A. Tapus, "Factors influencing emotional driving: examining the impact of arousal on the interplay between age, personality, and driving behaviors," *Frontiers in Psychology*, vol. 16, p. 1487493, 2025.

[33] M. J. Eppler and J. Mengis, "The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines," *The Information Society*, vol. 20, no. 5, pp. 325–344, 2004.