Active Maritime Copilot for Ambiguity and Risk-Driven Decisions

Zhichen Lu
Autonomous Systems and Robotics Lab, U2IS, ENSTA
Paris Institut Polytechnique de Paris & CROSSING IRL
Palaiseau France

Matthew Stephenson
College of Science and Engineering, Flinders University &
CROSSING IRL
Adelaide Australia

Benoit Clement
ENSTA, Institut Polytechnique de Paris, Lab-STICC, CNRS
UMR 6285 & CROSSING IRL
Brest France

Adriana Tapus
Autonomous Systems and Robotics Lab, U2IS, ENSTA
Paris Institut Polytechnique de Paris
Palaiseau France

Abstract

Cross-modal conflicts in maritime navigation—where a vessel's verbal communication contradicts its physical maneuvers (e.g., promising to give way while maintaining speed) pose severe risks to safety. Current autonomous systems often process sensor data and linguistic inputs in isolation, failing to detect such discrepancies. We present a Multimodal Agentic Framework that serves as a "Watchful Copilot," using Retrieval-Augmented Generation (RAG) to cross-reference navigational dialogue with real-time kinematic data. To manage uncertainty, a Risk-Prioritized Interface employs progressive disclosure, escalating from a "Green" (Verified) state to a "Yellow" (Ambiguous) state, where the agent visualizes supporting evidence and requests human supervision for clarification. Preliminary validation in a 2D simulation benchmark ($N = 13$) provides initial evidence that this human-in-the-loop workflow may support reduced cognitive load and appropriate trust calibration in high-ambiguity scenarios, warranting further investigation.

CCS Concepts

• **Human-centered computing → Collaborative interaction**; *User interface design*; • **Computing methodologies → Intelligent agents**; • **Applied computing → Transportation**.

Keywords

Human-Agent Teaming, Maritime Human-Robot Interaction, Multimodal Agent, Intent Verification

ACM Reference Format:
Zhichen Lu, Matthew Stephenson, Benoit Clement, and Adriana Tapus. 2026. Active Maritime Copilot for Ambiguity and Risk-Driven Decisions. In *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI Companion '26), March 16–19, 2026, Edinburgh, Scotland, UK*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3776734.3794393

1 Introduction

As the maritime industry transitions toward Maritime Autonomous Surface Ships (MASS), the human operator's role is fundamentally shifting from manual execution to supervisory control [12][14]. Although automation now performs much of the trajectory planning, maritime safety still depends heavily on nuanced social coordination and intent sharing between vessels [6][8][13]. Statistics from the IMO (International Maritime Organization) indicate that most maritime accidents are caused by human error and incorrect interpretation of intentions during ship-to-ship negotiations [7]. In high-intensity maritime environments, operators encounter what we call "Cross-modal Conflicts", situations where verbal intent communicated via VHF contradicts kinematic observations from sensors. Unlike traditional multimodal fusion approaches that assume modality complementarity, our framework explicitly models the possibility of conflict, for instance, when a target vessel verbally declares a 'turn to port' while radar continuously tracks a steady course. These discrepancies can arise from misunderstanding, deception, or equipment failure, each demanding distinct operator responses. Resolving them typically requires manually integrating multiple data sources [4] [17], which can lead to cognitive overload and reduced situational awareness [10]. Resolving these conflicts is challenging because nautical communication is inherently ambiguous, creating potentially catastrophic risks in high-stakes situations [5] [16]. Context-dependent expressions often lack precise kinematic information, making reliance on Large Language Models (LLMs) alone insufficient for accurately inferring vessel intent [19] [9]. Moreover, current autonomous systems frequently treat sensing and communication separately, neglecting the crucial human component in mixed-traffic environments [2]. This opacity in system design where the agent acts as a "black box" without verifying verbal agreements, significantly undermines the efficacy of Human-Agent Teaming (HAT) and operator trust [18] [15].

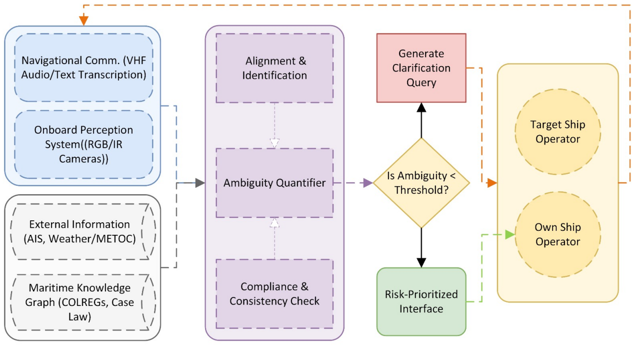

Navigational Comm. (VHF Audio/Text Transcription)
Onboard Perception System((RGB/IR Cameras))
External Information (AIS, Weather/METOC)
Maritime Knowledge Graph (COLREGs, Case Law)
Alignment & Identification
Ambiguity Quantifier
Compliance & Consistency Check
Generate Clarification Query
Is Ambiguity < Threshold?
Risk-Prioritized Interface
Target Ship Operator
Own Ship Operator

Figure 1: Human-Agent Collaborative Maritime Copilot Framework.

HRI Companion '26, Edinburgh, Scotland, UK
© 2026
ACM ISBN 979-8-4007-2321-6/2026/03
https://doi.org/10.1145/3776734.3794393

In this paper, we propose a Multimodal Agentic Framework designed to support human decision-making as a ”Watchful Copilot”. Unlike passive monitoring systems, our agent employs a RAG mechanism to actively cross-verify transcribed verbal intents against dynamic behavioral signals. Our core contribution is a Risk-Prioritized Interface that manages uncertainty through progressive disclosure: the system dynamically adjusts information granularity—escalating from a ”Green” (Verified) state to a ”Yellow” (Ambiguous) state, to facilitate multi-turn clarification under human supervision. We also introduce deceptive utterances to simulate challenging scenarios [3]. Preliminary results indicate that this approach helps align sensor data with linguistic intent, promoting calibrated trust in high-stakes maritime collaboration.

2 Methodology

The conceptual architecture of our Multimodal Agentic Framework is depicted in Figure 1. To address the challenge of cross-modal conflict, the system operates on a closed-loop workflow that verifies information through three interconnected stages: Simulated Perception, Agentic Reasoning Core, and a Human-Agent Interaction Loop.

2.1 Simulated Perception and Knowledge Grounding

To isolate the effects of linguistic-behavioral discrepancies from environmental noise, we implement a Simulated Perception Layer in our 2D benchmark. Rather than processing raw visual inputs, the agent uses high-fidelity Kinematic State Vectors (position, velocity, heading) that represent the vessel’s physical state. This information is synchronized with Navigational Communication (transcribed VHF text). Reasoning relies on two complementary knowledge sources: a Dynamic Database of real-time objective data (Automatic Identification System (AIS) readings, relative geometry) and a static Maritime Knowledge Graph encoding regulatory constraints.

2.2 Agentic Reasoning Core via RAG

The central processing unit employs a RAG mechanism to bridge the gap between semantic intent and physical action. Specifically, the pipeline operates in three stages:

(i) Intent Extraction: An LLM (Google Gemini 3 pro) parser processes transcribed VHF dialogue to extract structured navigational intent (e.g., action: alter_course, direction: starboard).

(ii) Knowledge Retrieval: The extracted intent triggers retrieval from two knowledge bases: (a) a Dynamic Database containing real-time objective facts fetched via APIs, including kinematic data of ships(AIS/visual estimates) and navigational status; and (b) Static Maritime Knowledge Graph: Contains encoded regulatory constraints, primarily the Convention on the International Regulations for Preventing Collisions at Sea (COLREGs) [1] and standard maneuver patterns.

(iii) Grounded Generation: The LLM does not directly compute trajectories. Instead, it extracts estimated structured parameters (intended action, direction, magnitude) from the text, which are then passed to a deterministic kinematic model that generates the Ghost Trajectory using standard equations of motion. The LLM serves as a semantic parser, not a trajectory planner.

A Compliance & Consistency Check module compares this Ghost Trajectory against the actual vessel vector. An Ambiguity Quantifier calculates $S_{amb} \in [0, 1]$ as: $S_{amb} = 0.5 \times V_{semantic} + 0.5 \times D_{kinematic}$ (equal weighting was selected as a neutral prior for this formative validation phase)

where $V_{semantic}$ captures linguistic vagueness, reflecting the absence of precise action verbs or parameters in navigational dialogue and $D_{kinematic}$ measures kinematic divergence, the geometric discrepancy between the predicted Ghost Trajectory and the vessel’s real-time motion. State thresholds are defined as Green ($S_{amb} < 0.3$), Yellow ($0.3 - 0.7$), and Red (> 0.7). We let $V_{semantic}$ be quantified via a LLM employing a few-shot prompting strategy, calibrated to output a scalar score $\in [0, 1]$ based on the presence of specific actionable parameters such as action verb, value, reference) in the transcribed VHF message. We also let $D_{kinematic}$ be computed as a time-averaged, normalized divergence between the observed vessel state vector $\mathbf{S}_{actual}$ and the predicted ghost state vector $\mathbf{S}_{ghost}$ over a sliding window T. It is defined as the weighted sum of normalized heading divergence (s_θ) and speed divergence (s_v):

$$D_{kinematic} = \frac{1}{N} \sum_{t \in T} [w_\theta \cdot \min(\frac{|\Delta\theta_t|}{\theta_{max}}, 1) + w_v \cdot \min(\frac{|\Delta v_t|}{v_{max}}, 1)]$$

where N is the number of time steps in the sliding window T, $\Delta\theta_t$ and Δv_t are instantaneous discrepancies, and θ_{max} and v_{max} are empirically defined thresholds representing significant deviation.

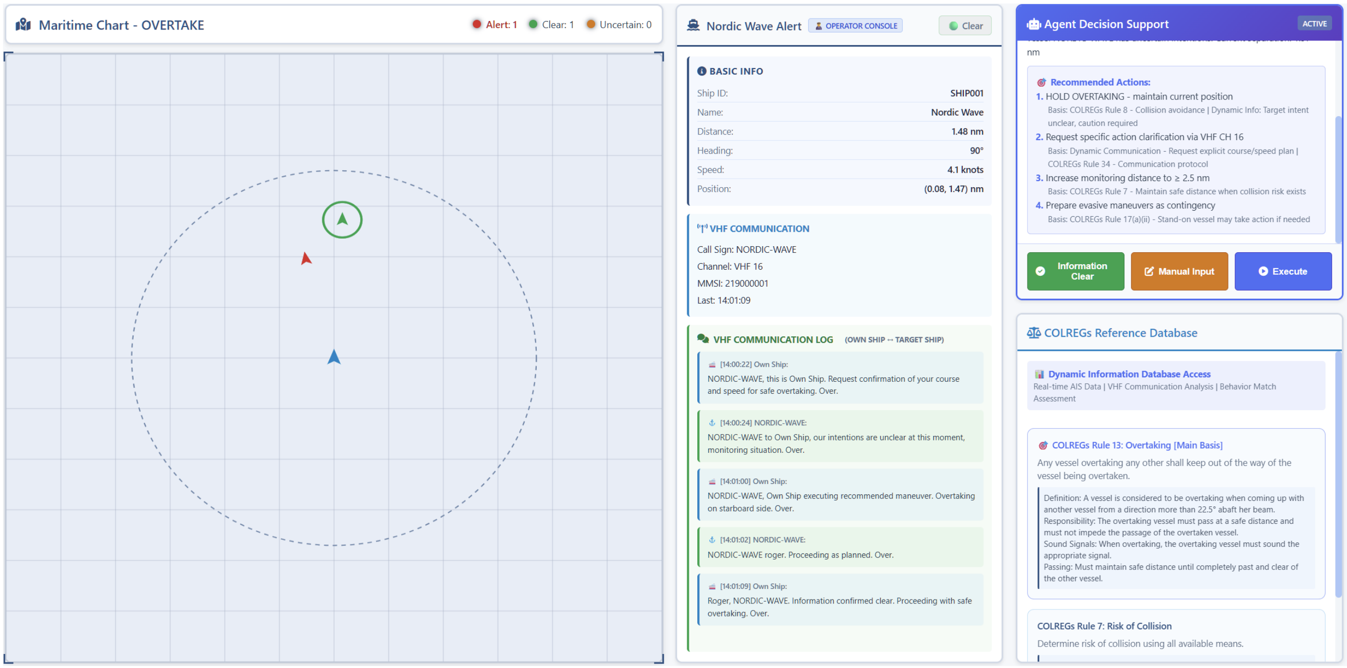
2.3 Operational Workflow and Risk-Prioritized Interface

The workflow starts with continuous kinematic monitoring. When a potential collision risk is detected, the agent activates its **Tiered Response Strategy**, employing a **Progressive Disclosure** mechanism to manage operator trust and reduce cognitive load:

- *Green (Verified):* When the verbal intent aligns with physical maneuvers and complies with COLREGs (low S_{amb}), the system logs the event as ”Clear” without interrupting the operator.
- *Yellow (Ambiguous - Supervisor-in-the-Loop):* If S_{amb} exceeds a moderate threshold (e.g., vague language like ”I will keep clear” or minor trajectory deviation), the interface highlights the target in yellow. Crucially, the agent **visualizes the reasoning rationale** (citing specific COLREGs as basis) and requests human authorization to initiate a specific clarification query.
- *Red (Conflicting):* In cases of critical cross-modal conflict (e.g., a verbal ”turn port” contradicted by a visual straight-line vector) the agent escalates to a red alert and prepares an emergency query to resolve the immediate danger.

3 Experimental Design

Given that the proposed architecture is a highly coupled system integrating multi-turn human interaction, multimodal information processing, and RAG, it is important to verify system completeness and evaluate the interaction design prior to engaging professional human evaluators. A significant challenge in experimental design lies in the inherent ambiguity of COLREGs [1]. Concepts such as ”safe distance” lack precise quantitative definitions, and scenario



BASIC INFO

Ship ID:	SHIP001
Name:	Nordic Wave
Distance:	1.48 nm
Heading:	90°
Speed:	4.1 knots
Position:	(0.08, 1.47) nm

VHF COMMUNICATION

Call Sign: NORDIC-WAVE
Channel: VHF 16
MMSI: 219000001
Last: 14:01:09

VHF COMMUNICATION LOG (OWN SHIP ↔ TARGET SHIP)

[14:00:22] Own Ship:
NORDIC-WAVE, this is Own Ship. Request confirmation of your course and speed for safe overtaking. Over.

[14:00:24] NORDIC-WAVE:
NORDIC-WAVE to Own Ship, our intentions are unclear at this moment, monitoring situation. Over.

[14:01:00] Own Ship:
NORDIC-WAVE, Own Ship executing recommended maneuver. Overtaking on starboard side. Over.

[14:01:02] NORDIC-WAVE:
NORDIC-WAVE roger. Proceeding as planned. Over.

[14:01:09] Own Ship:
Roger, NORDIC-WAVE. Information confirmed clear. Proceeding with safe overtaking. Over.

nm

Recommended Actions:

1. HOLD OVERTAKING - maintain current position
   Basis: COLREGs Rule 8 - Collision avoidance | Dynamic Info: Target intent unclear, caution required
2. Request specific action clarification via VHF CH 16
   Basis: Dynamic Communication - Request explicit course/speed plan | COLREGs Rule 34 - Communication protocol
3. Increase monitoring distance to ≥ 2.5 nm
   Basis: COLREGs Rule 7 - Maintain safe distance when collision risk exists
4. Prepare evasive maneuvers as contingency
   Basis: COLREGs Rule 17(a)(ii) - Stand-on vessel may take action if needed

COLREGs Reference Database

Dynamic Information Database Access
Real-time AIS Data | VHF Communication Analysis | Behavior Match Assessment

COLREGs Rule 13: Overtaking [Main Basis]
Any vessel overtaking any other shall keep out of the way of the vessel being overtaken.

Definition: A vessel is considered to be overtaking when coming up with another vessel from a direction more than 22.5° abaft her beam.
Responsibility: The overtaking vessel must pass at a safe distance and must not impede the passage of the overtaken vessel.
Sound Signals: When overtaking, the overtaking vessel must sound the appropriate signal.
Passing: Must maintain safe distance until completely past and clear of the other vessel.

COLREGs Rule 7: Risk of Collision
Determine risk of collision using all available means.

Figure 2: Risk-Prioritized Interface: (left) Maritime chart with vessel positions; (middle) Real-time status and VHF log; (top-right) RAG evidence panel showing retrieved COLREGs and kinematic data; (bottom-right) Operator authorization controls.

classification is often equivocal; for instance, a Target Ship (TS) approaching from the port quarter could be interpreted as either a "crossing" or an "overtaking" situation. To rigorously assess the agent's robustness in handling such multimodal ambiguity, we developed a simulation benchmark library comprising selected typical scenarios. This experiment employs a 2D planar simulation paradigm specifically to isolate environmental noise (e.g., wave dynamics and lighting conditions), thereby focusing the evaluation on the agent's capability to cross-verify physical spatial relationships against semantic communication content (Figure 2).

Each experimental scenario is defined along two dimensions. *Physical Situation* specifies the encounter geometry and kinematic parameters based on COLREGs, while *Communication Congruency* measures how well the preset VHF communication aligns with the actual physical situation, serving as the key independent variable.

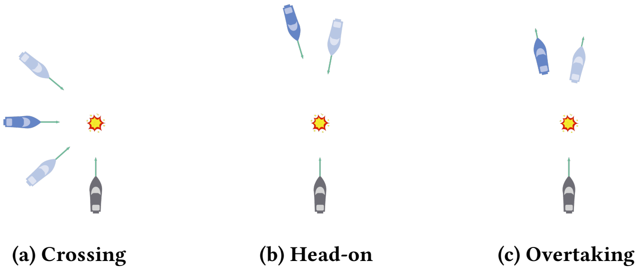

(a) Crossing **(b) Head-on** **(c) Overtaking**

Figure 3: COLREGs-based classification of encounter situations. (Star symbol indicates the predicted point of collision.)

3.1 Physical Base Layer

We designed a set of physical scenarios covering the main COLREGs situations (Figure 3):

- Head-on Situation (1 scenario): Two vessels on reciprocal courses, creating a direct collision risk.
- Crossing Situations (2 scenarios): One case with the target vessel approaching from the port side (own ship is the stand-on vessel) and one from the starboard side (own ship is the give-way vessel).
- Overtaking Situations (2 scenarios): One case where the own ship overtakes the target vessel and one where the target vessel overtakes the own ship.

Note: The number in parentheses indicates the number of distinct scenario instances created for each encounter type.

3.2 Communication & Congruency Layer

To evaluate the agent's iterative clarification mechanism, we assigned specific VHF communication presets to each physical scenario, defining three Ground Truth Conditions.

Condition A: Consistent/Clear (Expected State: Green) represents cases where communication is unambiguous and perfectly aligned with physical movements. For instance, in a high-risk starboard crossing where the target ship is faster and should give way, the visual feed shows the target vessel clearly turning starboard, accompanied by the preset message, "Own ship is altering course to starboard to pass astern of you." In such cases, the agent identifies low ambiguity and triggers no intervention.

Condition B: Vague/Ambiguous (Expected State: Yellow) captures situations where potential physical risk exists but communication is non-specific and lacks actionable parameters. For example, in a port-side crossing with similar vessel speeds, the preset message, “I see you; will keep clear,” does not specify how the target ship will maneuver. Here, the agent detects semantic ambiguity and triggers a supervisory query.

Condition C: Conflicting/Discrepant (Expected State: Red) involves scenarios where verbal statements directly contradict sensor observations. In a head-on encounter with collision risk, the visual feed may show the target vessel maintaining a straight course despite a message stating, “We are altering course to starboard now for port-to-port passing.” In such cases, the agent identifies a critical cross-modal conflict and initiates an autonomous emergency query.

4 Results

We recruited $N = 13$ participants (8 HRI researchers and 5 engineering graduate students) to evaluate the 2D simulation prototype using a Think-Aloud protocol [11]. Participants assumed the role of Officer of the Watch (OOW), tasked with monitoring evolving traffic scenarios and authorizing agent interventions. We collected decision response times, behavioral logs, and qualitative feedback in Figure 4.

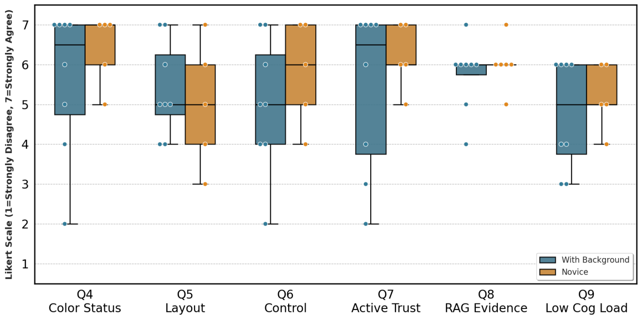


Figure 4: User experience comparison

Validation of Risk-Prioritized Disclosure: The quantitative results show high median scores (≥ 6) for Interface Color Status ($Q4$) across both groups, validating that the Red-Yellow-Green transition effectively communicated risk levels. Behavioral observations indicated that the interface mitigated the ”startle effect” often seen in high-stakes alarms. Participants reported their immediate reaction was to ”read the warning and see the suggestion” rather than panic, because the agent provided actionable context. However, the variance in initial reactions highlights a reliance solely on visual cues, with participants suggesting that ”adding sound may make it more alerting” for immediate awareness.

Trust Calibration via RAG Evidence: A critical success of the framework is the establishment of trust through explainability. The ”With Background” group showed remarkably high consensus on RAG Evidence ($Q8$), with the interquartile range collapsing around a score of 6. This indicates that HRI experts unanimously agreed on the value of structured dynamic/static evidence. We recorded an average decision response time of 4.5 seconds ($SD = 1.2$) in the ”Yellow” state. Behavioral observations confirmed this latency was not due to confusion, but to active cognitive verification: participants utilized this window to cross-reference the RAG evidence against the map. This deliberation fostered trust, as noted by one participant: ”If it just turned green automatically, I might doubt it. But since it showed me the ambiguity and asked for my approval, I trust the final result much more.”

Interface Friction and Future Iteration: While the workflow logic was validated, specific bottlenecks explain the wider variance observed in expert ratings for Layout ($Q5$) and Control ($Q6$). Participants noted that parsing textual COLREGs under pressure was cognitively demanding, suggesting that ”the font for important info could be larger” and VHF logs should be ”inversed” to show recent messages first. Most critically, the confusion regarding manual input suggests a need to transition from text-based to graphical explainability. Future iterations will overlay a ”Ghost Trajectory” directly onto the map to facilitate rapid visual validation of the agent’s proposed maneuvers.

5 Conclusion

To address the critical challenge of ”Cross-modal Conflicts” in maritime HRI, as a proof of concept, we introduced a Multimodal Agentic Framework functioning as a ”Watchful Copilot.” By integrating a RAG-based reasoning core with active multi-turn human-agent interaction and a Risk-Prioritized Interface, the system shows the potential to bridge the gap between sensor data and linguistic intent. Our key contribution lies in the application of Progressive Disclosure: instead of operating as a ”black box,” the agent visualizes its reasoning rationale, engaging the operator in an iterative clarification loop. Initial validation ($N = 13$) confirms that this mechanism shows preliminary promise in mitigating cognitive overload and, crucially, fosters appropriate trust calibration by explicitly acknowledging system uncertainty. Limitations include the abstracted 2D simulation, non professional participants, and predetermined scenarios, which bound generalizability to real-world maritime operations.

Future work will focus on improving ecological validity by moving from the current simulated kinematic layer to raw perception using Vision-Language Models (VLMs). While validated in maritime contexts, the core framework—cross-modal conflict detection via RAG and progressive disclosure for trust calibration—generalizes to other safety-critical human-agent teams, including air traffic control, surgical robotics, and autonomous vehicle supervision, where verbal coordination and physical behavior must align.

References

[1] 2022. Convention on the International Regulations for Preventing Collisions at Sea, 1972 (COLREGs) - COLREGs Operator Guidance Framework. https://www.rasgateway.com.au/Code-of-Practice-COLREGS-framework.pdf.

[2] Anas S Alamoush and Aykut I Ölçer. 2025. Maritime autonomous surface ships: architecture for autonomous navigation systems. *Journal of Marine Science and Engineering* 13, 1 (2025), 122. doi:10.3390/jmse13010122

[3] Damien Anderson, Matthew Stephenson, Julian Togelius, Christoph Salge, John Levine, and Jochen Renz. 2018. Deceptive Games. In *Applications of Evolutionary Computation*, Kevin Sim and Paul Kaufmann (Eds.). Springer International Publishing, Cham, 376–391. doi:10.1007/978-3-319-77538-8_26

[4] Erik Blasch, Tien Pham, Chee-Yee Chong, Wolfgang Koch, Henry Leung, Dave Braines, and Tarek Abdelzaher. 2021. Machine learning/artificial intelligence for sensor data fusion–opportunities and challenges. *IEEE aerospace and electronic systems magazine* 36, 7 (2021), 80–93. doi:MAES.2020.3049030

[5] PA Carson and CJ Mumford. 2011. Communication failure and loss prevention. *Loss Prevention Bulletin* 218 (2011).

[6] Christine Chauvin. 2011. Human factors and maritime safety. *The Journal of Navigation* 64, 4 (2011), 625–632. doi:10.1016/j.trpro.2019.07.183

[7] Christine Chauvin, Salim Lardjane, Gaël Morel, Jean-Pierre Clostermann, and Benoît Langard. 2013. Human and organisational factors in maritime accidents: Analysis of collisions at sea using the HFACS. *Accident Analysis & Prevention* 59 (2013), 26–37. doi:10.1016/j.aap.2013.05.006

[8] Linda de Vries. 2017. Work as done? Understanding the practice of sociotechnical work in the maritime domain. *Journal of Cognitive Engineering and Decision Making* 11, 3 (2017), 270–295. doi:10.1177/1555343417707664

[9] George Gabedava and Yanming Hu. 2025. Enhancing maritime safety through linguistic analysis: a case study of communication failures in maritime accidents. *WMU Journal of Maritime Affairs* (2025), 1–15. doi:10.1007/s13437-025-00371-y

[10] Nermin Hasanspahić, Srđan Vujičić, Vlado Frančić, and Leo Čampara. 2021. The role of the human factor in marine accidents. *Journal of Marine Science and Engineering* 9, 3 (2021), 261. doi:10.3390/jmse9030261

[11] C. H. Lewis. 1982. *Using the "Thinking Aloud" Method in Cognitive Interface Design*. Technical Report RC-9265. IBM.

[12] M Nardo, Daniel Forino, and Teresa Murino. 2020. The evolution of man–machine interaction: The role of human in Industry 4.0 paradigm. *Production & manufacturing research* 8, 1 (2020), 20–34. doi:10.1080/21693277.2020.1737592

[13] Zhegong Shangguan, Yang Liu, Le Song, Tingcheng Li, and Adriana Tapus. 2024. Using a Pneumatic Tactile Steering Wheel to Enhance the Multi-Modal Takeover Request In Smart Vehicle. In *International Conference on Social Robotics*. Springer, 122–132. doi:10.1007/978-981-97-8963-4_12

[14] Thomas B Sheridan. 2021. Human supervisory control of automation. *Handbook of human factors and ergonomics* (2021), 736–760. doi:10.1002/9781119636113.ch28

[15] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International journal of human-computer studies* 146 (2021), 102551. doi:10.1016/j.ijhcs.2020.102551

[16] Ana Lucia Tavares Monteiro. 2019. *Reconsidering the measurement of proficiency in pilot and air traffic controller radiotelephony communication: From construct definition to task design.* Ph. D. Dissertation. Carleton University.

[17] Simon M Taylor and Marc De Leeuw. 2021. Guidance systems: from autonomous directives to legal sensor-bilities. *Ai & Society* 36, 2 (2021), 521–534. doi:10.1007/s00146-020-01012-z

[18] Jonas Wanner, Lukas-Valentin Herm, Kai Heinrich, and Christian Janiesch. 2022. The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study. *Electronic Markets* 32, 4 (2022), 2079–2102. doi:10.1007/s12525-022-00593-5

[19] Jingbo Yin, Rafi Ullah Khan, Muhammad Afzaal, Hamed M. Almalki, Mohamad Ahmad Saleem Khasawneh, and Saleh Al Sulaie. 2025. Quantitative risk assessment of speech acts and lexical factors in maritime communication failures and accidents. *Safety Science* 191 (2025), 106968. doi:10.1016/j.ssci.2025.106968

Received 2025-12-08; accepted 2026-01-12