Journal of International Maritime Safety, Environmental Affairs, and Shipping

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tsea20

A state-of-the-art review of AI decision transparency for autonomous shipping

A. N. Madsen & T. E. Kim

To cite this article: A. N. Madsen & T. E. Kim (2024) A state-of-the-art review of AI decision transparency for autonomous shipping, Journal of International Maritime Safety, Environmental Affairs, and Shipping, 8:1-2, 2336751, DOI: 10.1080/25725084.2024.2336751

To link to this article: https://doi.org/10.1080/25725084.2024.2336751

Published online: 03 Apr 2024.

JOURNAL OF INTERNATIONAL MARITIME SAFETY, ENVIRONMENTAL AFFAIRS, AND SHIPPING
2024, VOL. 8, NOS. 1–2, 2336751
https://doi.org/10.1080/25725084.2024.2336751

Taylor & Francis
Taylor & Francis Group

A state-of-the-art review of AI decision transparency for autonomous shipping

A. N. Madsen[a] and T. E. Kim[b]

[a]Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology, Ålesund, Norway;
[b]Department of Technology and Safety, UiT The Arctic University of Norway, Tromsø, Norway

ABSTRACT

The introduction of maritime autonomous surface ships (MASSs) has prompted a significant shift in the role of navigators from traditional navigation to supervising artificial intelligence (AI) collision avoidance systems or managing operations from remote operation centers (ROCs). This may be problematic because the integration of AI technologies into collision avoidance systems may jeopardize safety if done in a way that reduces human control or leaves humans out of the loop. For onboard navigators or ROC operators who work with AI, it is important that the AI's "thinking" and decisions are transparent and that alternative decisions are easy to execute. Regarding navigation and collision avoidance, this issue can be defined as AI decision transparency. In this systematic review, we examined state-of-the-art research on traffic alerts and collision avoidance systems with respect to decision transparency for autonomous shipping. Through a thematic analysis, we identified three main groups of transparency in the reviewed literature: strategies, visualization, and technology, with respective subgroups.

ARTICLE HISTORY
Received 22 February 2024
Accepted 25 March 2024

KEYWORDS
MASS; TCAS; decision transparency; ROC; XAI

Introduction

Thousands of years ago, the Polynesian navigators voyaged across thousands of nautical miles of the Pacific Ocean in canoes, reaching most islands in the Polynesian Triangle. They navigated by observing the stars, birds, clouds, ocean swells, and wind patterns, relying on their senses and knowledge passed down through generations. Navigators in today's modern society use modern technology, such as radar and global navigation satellite systems, to determine ships' positions and collision hazards. However, navigators rely heavily on knowledge and experience to interpret traffic situations and make situational adaptations, as required by the Convention on the International Regulations for Preventing Collisions at Sea, 1972 (COLREGs). Since maritime autonomous surface ships (MASSs) have become a reality, artificial intelligence (AI) has the potential to support collision avoidance. Recent research have recently found that it is difficult for an AI system to make situational adaptations in the same manner as an experienced mariner, as supported by COLREGs (e.g (A. N. Madsen et al., 2022; Rutledal et al., 2020)).

Therefore, AI systems (e.g., for collision avoidance) must be designed for compatibility with human navigators, either onboard or in remote operation centers (ROCs). Navigators must find an AI's decisions transparent, interpretable, and accountable. This ensures that, when an unexpected incident occurs, the operators understand the AI's decision-making process and can verify its decisions. At present, it is difficult for AI systems to "explain" their ways of thinking to humans. Explainable artificial intelligence (XAI) has emerged as an AI subfield in computer science to address the complexities and nuances of the interpretability of AI models' interpretability. Stakeholders have different XAI needs, and for onboard navigators or ROC operators working with AI systems, it is important that AI systems' "thinking" and decisions are transparent and that alternative decisions are easy to execute. Regarding navigation and collision avoidance, this issue can be defined as AI decision transparency.

In this paper, we present the findings from a review of state-of-the-art research on traffic alerts and collision avoidance systems with respect to decision transparency for MASS. The goal of this review was to discover how researchers have treated the collaboration between humans and machines regarding MASS.

Method

MASSs represent a transformative step in maritime transportation, promising significant advantages for operational efficiency, safety, and environmental sustainability. However, these benefits are intertwined with the challenges of ensuring that autonomous technology- and AI-driven decisions are both interpretable and accountable. Hence, we conducted a systematic review to collate and analyze existing research

CONTACT A. N. Madsen andreas.madsen@ntnu.no Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology, Ålesund, Norway

regarding the transparency of AI decisions for autonomous ships. Guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines developed by (Page et al., 2021), we methodically extracted and assessed articles that directly or indirectly shed light on this emerging topic.

We searched three databases: Scopus, Web of Science, and EBSCO Academic Complete. This ensured broad indexing, although it resulted in many duplicates. We included articles published between 1 January 2012, and 18 September 2023, and selected papers according to the criteria shown in Figure 1. The figure is based on the PRISMA four-phase diagram inspired by (Veitch & Alsos, 2022). The first author of this paper carried out the initial search and identified 111 articles that were potentially eligible for screening by abstract.

As shown in Figure 1, we removed duplicate articles and then inspected the titles and abstracts to exclude articles unrelated to the topic. We scrutinized the full texts of the remaining articles for relevance. The inclusion criteria encompassed peer-reviewed English-

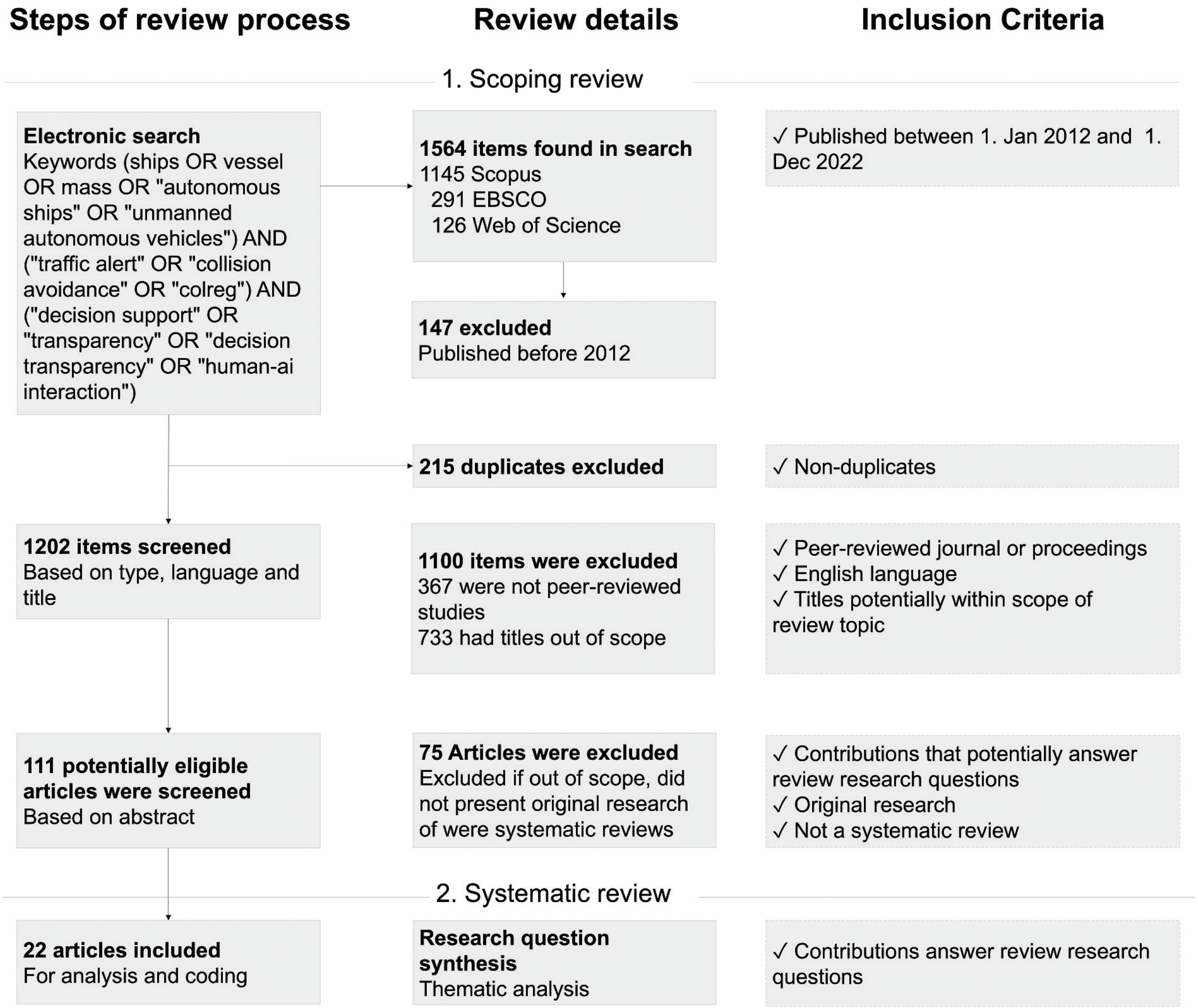
Steps of review process
Review details
Inclusion Criteria
1. Scoping review
Electronic search
Keywords (ships OR vessel OR mass OR "autonomous ships" OR "unmanned autonomous vehicles") AND ("traffic alert" OR "collision avoidance" OR "colreg") AND ("decision support" OR "transparency" OR "decision transparency" OR "human-ai interaction")
1564 items found in search
1145 Scopus
291 EBSCO
126 Web of Science
✓ Published between 1. Jan 2012 and 1. Dec 2022
147 excluded
Published before 2012
215 duplicates excluded
✓ Non-duplicates
1202 items screened
Based on type, language and title
1100 items were excluded
367 were not peer-reviewed studies
733 had titles out of scope
✓ Peer-reviewed journal or proceedings
✓ English language
✓ Titles potentially within scope of review topic
111 potentially eligible articles were screened
Based on abstract
75 Articles were excluded
Excluded if out of scope, did not present original research of were systematic reviews
✓ Contributions that potentially answer review research questions
✓ Original research
✓ Not a systematic review
2. Systematic review
22 articles included
For analysis and coding
Research question synthesis
Thematic analysis
✓ Contributions answer review research questions

Figure 1. Flowchart of the review process for the systematic review.

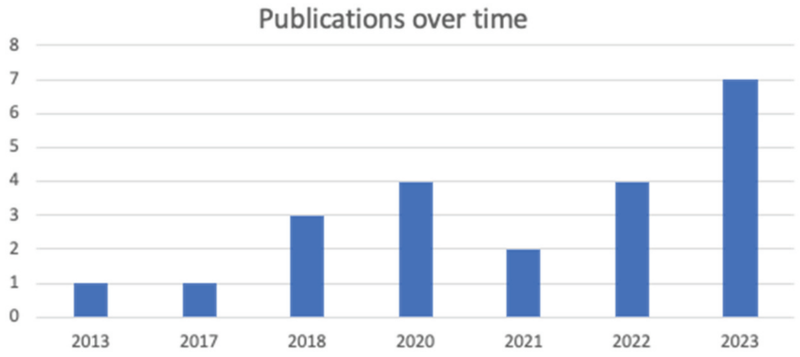

Figure 2. Number of included papers published over time.

Table 1. Groups identified in thematic analysis.

Main groups	Subgroups	
A	1. Human factors	*n = 4*
Strategies	2. Risk assessment	*n = 4*
n = 13	3. Design principles	*n = 8*
B	4. Color coding	*n = 6*
Visualization	5. Bounding boxes	*n = 3*
n = 8	6. Route displays	*n = 3*
C	7. System SA	*n = 3*
Technology	8. Route exchange	*n = 2*
n = 4	9. Identification	*n = 2*

language articles that focused on AI transparency, decision-making interpretability, or accountability in MASS contexts. We then subjected relevant articles to an in-depth review and data extraction.

Thereafter, we performed independent reviews of the papers and cross-verified our results in a workshop. We resolved any disagreements and noted reasons for excluding papers. Of the full-text sample of 111 articles, 89 were excluded for reasons relating to the predefined criteria (see Table 1). A final dataset of 22 full-text publications remained for the qualitative analysis: 17 journal articles and 5 conference papers.

We then performed thematic analysis of the publications. Thematic analysis in a literature review involves identifying recurring themes and patterns across a body of literature, making it possible to categorize, summarize, and synthesize existing knowledge. This approach allowed us to extract meaningful insights from the data and uncover key groups of how research have treated the collaboration between humans and machines with respect to MASS.

For the analysis, we employed (Braun & Clarke, 2012) six-phase approach. In the first phase, we familiarized ourselves with the data by reading the texts and noting information that seemed relevant to the research topic. In the second phase, we conducted initial coding by reading the articles again and highlighting passages of text that were relevant to the topic. In the third phase, we reviewed the coded data using an iterative process to construct themes based on how decision transparency is treated in the articles.

In the fourth phase, we conducted a quality check, reviewing the themes in relation to the coded data and the entire dataset. This led us to the fifth phase, in which we named the themes and divided them into three groups: strategies, visualization, and technology. Each of the coded articles was allocated to the most appropriate group, since some of the articles discussed themes that fit into multiple groups, and we later defined subgroups, as mentioned in the next section. The sixth and final phase involved writing this article.

Results

We identified three main groups of decision transparency in the papers: strategies (addressed in 13 different

Table 2. Articles in each group.

Main group	Article ID numbers corresponding to table in Appendix A																					
Sub group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Strategies			x	x	x			x		x	x	x	x	x			x	x			x	
Human factors				x	x						x			x								
Risk Assesment										x							x	x				
Design principles			x	x	x			x				x	x								x	x
Visualization	x					x	x								x	x			x		x	
Color Coding	x					x									x	x					x	x
Bounding boxes	x																				x	x
Route Display							x								x				x			
Technology		x							x											x		
System SA		x							x													x
Route Exchange									x											x		
Identification									x											x		

reviewed studies), visualization (addressed in 8 different reviewed studies), and technology (discussed in 4 different reviewed studies). These main groups were further split into subgroups, as shown in Table 1.

A table summarizing these articles is presented in Appendix A. The graph in Figure 2 shows the publication year of the included articles, while Table 2 is a matrix showing each article and its corresponding group(s).

Strategies

The first group of articles focused on the intricacies of developing and implementing AI decision-making processes for autonomous ships. The role of human factors was evident within this purview, with many research studies emphasizing the importance of integrating human cognitive processes into AI system designs to ensure synergy between human operators and AI. Another vital theme was risk assessment. The literature indicated that AI systems, when integrated into dynamic risk evaluations, especially those involving real-time environmental and maritime traffic data analysis, tended to make safer and better-informed decisions. Furthermore, many of the discussions in this category revolved around specific design principles, particularly modularity, transparency, and fault tolerance, as foundational elements in designing robust and reliable AI systems for the challenges of maritime navigation.

This section presents the findings regarding strategies for ensuring decision transparency and addresses human factors, risk assessment, and design principles.

Human factors

The first group of strategies for achieving AI decision transparency related to human factors. A key element of such transparency is ensuring the situational awareness (SA) of agents interacting with a system. Although the articles highlighted the importance of SA, they failed to explain how the authors viewed or defined it. When exploring human – machine interactions and their influence on SA, it is important to define SA and to identify measures that can build and maintain operators' SA. Compatibility between human and system behaviors is important (Man et al., 2018; Van de Merwe et al., 2022), but it depends on understanding the underlying decision-making processes of operators (Lynch et al., 2022; Van de Merwe et al., 2022). Due to the novelty of MASSs and the lack of real-world case studies, considering other industries, such as aviation, is a potential approach for understanding operators' new roles and, in turn, discovering the means to achieve optimal levels of decision transparency (Lynch et al., 2022). On a conventional ship, the navigator's role and tasks are clearly defined (International Chamber of Shipping, 2022), and it might be helpful to conduct in-depth studies of the workflows of manned ships, to identify the tasks that new systems are meant to support. However, designers should not assume that technological design should match existing practices since technology created for onboard ship handling purposes may induce "human errors" when used in a different context, such as a ROC (Man et al., 2018). Another key component of decision transparency is trust. Operators must trust an AI system, and transparent decision-making processes foster this trust by enabling operators to gain insight into the system's limitations and to know when to intervene or override system decisions. Van de Merwe et al. (2022) found that navigators trust an AI system if it behaves as they would. An effective strategy can therefore be to ensure that a system mimics (or at least reflects) human behavior, resulting in a human-like machine. Wu et al. (2021) proposed using prospect theory (Kahneman et al., 1979) to systematically identify possible hazards involving human cognition and interactions, and to create intelligent collision avoidance systems that consider individual navigators' risk appetite (RA) and maneuvering behavior. Taking an RA approach can support both efficiency and transparency since a system is "tailored" to each navigator.

The research reviewed for this section emphasized the importance of SA, the compatibility of human and system behaviors, and the role of trust in fostering transparency. It also suggested drawing insights from other industries, such as aviation, and highlighted the need to understand operators' decision-making processes. Operators must trust the systems they use, and transparency in decision-making processes may foster this trust. Considering human behavior in system design can lead to the development of machines that mimic human decision-making.

Risk management

In this section, we present the findings regarding the potential of risk management systems to contribute to decision transparency for autonomous ships. The integration of proactive and reactive risk management strategies, as well as the consideration of human – system interactions, have been identified as crucial factors for enhancing the safety and reliability of MASS operations.

A fairly new method of risk modeling is system-theoretic process analysis (STPA), which is a safety analysis process used to identify and mitigate potential hazards in complex systems. Unlike traditional methods that focus on specific components, STPA is used to analyze how an entire system functions, emphasizing interactions between components and human operators. It systematically considers unsafe control actions, inadequate feedback, and flawed system designs, with the aim of uncovering vulnerabilities during the early stages of a development process. STPA provides a holistic view of system safety, enabling engineers to design safer systems by addressing underlying systemic issues rather than merely symptoms. Utne et al. (2020) proposed an online risk model to provide decision support for the control systems of autonomous ships subject to environmental and operational conditions and constraints, both proactively and reactively. Here, proactively means early warnings on possible violations of an autonomous ship's operating envelope based on safety constraints, and reactively means that human operators and supervisors are given more time for efficient responses and crisis intervention through predictions of possible outcomes. Such an approach may enable a system to perform risk management and enhance its intelligence. When operators have the opportunity to tap into this capability, it may lay the foundation for decision transparency. Cheng et al. (2023). further developed STPA with a focus on the safety of interactions between humans and systems and proposed a system theoretic approach to safety analysis for human – system collaboration. The novelty of this approach lies in defining the MASS operational context and integrating human cognitive modeling into STPA (i.e., STPA-Cog). This involves the systematic identification of possible hazards involving human cognition and interactions, which can be used to improve the design of ROCs for MASSs.

Putting such systems to the test, (Dugan et al., 2023) demonstrated the use of STPA to analyze system behavior and identify test cases for system operation. They found that requirements for the verification of critical systems fall into two categories: failure handling and integration testing. The STPA was applied to develop a decision support system for collision avoidance based specifically on vessel stability and the calculation of collision avoidance maneuvers.

Implementing proactive risk management strategies, such as early warnings on safety violations, can enhance decision transparency by allowing stakeholders to anticipate potential issues before they occur. Transparency in the form of clear warnings and alerts may enable human operators to understand a system's assessment of risks, thereby fostering trust and comprehension. The integration of STPA-Cog into risk management processes can enhance decision transparency by making AI decision-making processes more understandable to human operators. When human cognition and interactions are systematically considered in risk analysis, they provide a rationale for an AI system's decisions, enhancing transparency, and may reduce the "black box" effect. Reactive risk management, coupled with efficient crisis response mechanisms, can ensure that human operators have adequate time to respond to emerging situations. Transparent communication regarding a system's predictions of possible outcomes may enable operators to make swift, informed decisions. Understanding an AI system's predictions enhances decision transparency by allowing human operators to comprehend the basis for crisis interventions. Integrating risk management strategies directly into decision-making processes, such as collision avoidance, may provide transparency by making decision criteria explicit. When AI-driven decisions are based on predefined risk management protocols, operators can understand the factors considered in decision-making, promoting transparency and predictability.

Incorporating STPA and similar methodologies facilitates the systematic identification of hazards and the formulation of test cases. This structured approach, in turn, enhances system explainability by providing clear reasoning for the identification and management of risks. Clear and comprehensible explanations may build trust in AI systems, thereby increasing decision transparency.

Design principles

Visual feedback is important for helping MASS operators acquire and maintain satisfactory SA. However, (Lynch et al., 2022) found that it is necessary to consider how visual feedback is implemented and to ensure that it does not cause distraction. It may be necessary to combine visual feedback with audio alerts to ensure that visual information noticed on the human – machine interface. It is vital for operators to have the necessary knowledge of a ship's automated system to allow them to make appropriate decisions. Lynch et al. (2022) suggested that appropriate simulator training may facilitate this.

When designing a system for a distributed context, such as a ROC, (Man et al., 2018) stated that work tasks can determine the design requirements for the interface, but the way the interface is designed also suggests possibilities for new work practices. Furthermore, an important element of system design for a new work practice may not be how closely it matches the current work practice, but how well it can support possible future work practices and actions. (Man et al., 2018). also argued that a system must reflect constraints in the work domain and support user – environment coupling so that users can directly understand what is going on. Veitch et al. (2022) results revealed a discrepancy between designers' constructions of human – AI collaboration and navigators' own accounts in the field. Collaboration with AI systems largely depends on rendering computational activities more visible to align with the needs of human collaborators by displaying the AI system's actions transparently. Veitch & Alsos. (2021) pointed out that design practices are often oriented toward end-user interactions without fully formulated conceptualizations of what design practices entail. They proposed the formation and definition of human-centered XAI for interaction design and underlined the need for different visualizations for different groups due to their different XAI needs.

Van de Merwe et al. (2023) investigated how an information processing model and cognitive task analysis could be used to drive the development of transparency concepts. They suggested a layered approach to transparency to enable remote operators to observe the different facets of a system's input parameters, reasoning, decisions, and actions pertaining to collision situations. They found that the information needed to understand a system depended on the type of situation, the degree of human oversight, the complexity of the situation, and/or the time available to intervene. Pietrzykowski. (2018) found that the ship domain differed significantly in port approach areas and suggested that systems must be developed to function appropriately in port approach areas where traffic is much heavier. Systems should be designed to enable operators to intervene when necessary, combining the complementary strengths of human operators and machines. Huang et al. (2020) proposed an AI-based collision avoidance system to support cooperation between humans and systems. The system displayed its optimal decisions and highlighted dangerous solutions. This helped the human operator validate the system's solutions or propose and validate new solutions if discrepancies arose between the operator's and the system's solutions.

The focus on design principles, especially decision transparency, marks a shift in human – machine collaboration. By focusing design efforts on making AI system decisions visible, understandable, and tailored to human cognition, these principles not only enhance the efficiency of operations but may also contribute significantly to the safety and effectiveness of MASS and other complex systems. Designers should also be aware of MASS operating areas and distinguish and maintain separation between ships in open waters, coastal areas, and inland waters.

Visualization

As the previous section has shown, several of the identified strategies for decision transparency involve the visualization of AI systems' decisions. This section presents the types of visualization proposed in the included publications.

Color CodING

One of the significant approaches explored in recent research is color coding, which provides a visual means of conveying complex information to operators. Several studies have investigated the application of color coding to AI decision transparency. Ban & Kim. (2023) conducted an evaluation of a prototype software tool for marine traffic with vessel traffic service (VTS) operators. The tool aimed to help the decision-making process of the VTS operators and reduce their workload. The results of which may be transferrable to the roles of remote operators. The results indicated that the operators needed multiple color schemes to visualize the spatial data in charts, emphasizing the importance of tailoring color coding to the specific needs of operators. Abu-Tair & Naeem. (2013) used color coding for a decision support system based on information obtained from an object detection system. The system displayed a sector 180º (line of sight in front of an unmanned surface vehicle) in real time based on the object detection system, and objects not within this sector had predicted position of targets observed earlier. These were displayed in red and as query marks, whereas objects in the line of sight were marked in green.

Ożoga & Montewka. (2018) proposed color coding that would function as an overlay on either Automatic Radar Plotting Aid (ARPA) or Electronic Chart Display Information System (ECDIS). They proposed a color scheme as a function of Time to Closest Point of Approach and Closest Point of Approach (CPA), with five different colors indicating statuses from no hazard to extreme danger. The displayed color coding illustrated safe sectors for desired maneuvers. Pietrzykowski et al. (2017) presented a rosette of color coded sectors indicating safe and dangerous zones, similar to (Ożoga & Montewka. 2018), but not displayed on a map but as an addition to the map, on an information bar on top of the screen. This resembles the transparency layers conceptualized by (Van de Merwe et al., 2023), which functioned as maneuverability indicators to enable a ship to maneuver within a vector length, with traffic light color coding based on risk evaluation. Zhao et al. (2023) color coded each target ship according to its collision risk and further illustrated the position of possible collisions, with a red circle representing the ship domain.

In summary, these studies have collectively underscored the importance of nuanced, adaptive color-coding systems for visualizing AI decision transparency. By considering the specific needs of operators and integrating quantitative data, these color-coding approaches may contribute to enhanced SA, ultimately contributing to safer and more efficient operations.

Bounding boxes

In computer visualization and object detection, bounding boxes are used to outline and identify objects within images or video frames. (A. Madsen et al. 2023) found that operators need to know which targets an AI system considers when calculating and making a decision for collision avoidance. Helping operators visualize detected objects may allow them to understand what an AI system has detected and base their decisions on that information. We discovered differences between the bounding boxes identified in this review: (1) bounding boxes in object detection systems (e.g., camera/video feeds), and (2) bounding boxes that provided an overview (e.g., ECDIS or radar). Abu-Tair & Naeem. (2013) used bounding boxes around target vessels based on objects successfully detected by the camera of an object detection system, allowing operators to focus on what the system saw. Once an obstacle was detected and its state estimated, the information was displayed on a virtual map to assist the operators. Van de Merwe et al. (2023) used color coded shapes on a radar screen to highlight ships that the system consider as risk for collision. Although these approaches enhance transparency, some implications have not yet been fully explored. For instance, a bounding box alone does not convey a system's evaluation of an identified object. It seems that a holistic approach is needed to combine bounding boxes with additional layers of transparency. Integrating bounding boxes into comprehensive information frameworks could provide operators with a more nuanced understanding of situations and facilitate more informed decision-making processes.

Route display

An obvious component of decision transparency is the decision itself. In collision avoidance, such decisions typically relate to changes in trajectories, routes from the present position to the next waypoint, or speed changes (A. Madsen et al., 2023; Martelli et al., 2023; Pietrzykowski et al., 2017; Zhao et al., 2023). Pietrzykowski et al., (2017) summarized research on navigational decision (NAVDEC) support systems. The NAVDEC have been positively verified under laboratory conditions and in field tests. The NAVDEC system displays collision avoidance maneuvers as routes from the present position to the next waypoint and allows ships to pass other targets at an assumed CPA. (A. Madsen et al., 2023) termed this layer of transparency "minimal explainability" and found that the effect of displaying system-recommended routes/maneuvers to the operators enhanced their understanding of the system's intentions according to circumstances. In simple scenarios, a displayed route is often self-explanatory, but in more complex situations, operators may not always comprehend a system's reasoning, and they may thus intervene inappropriately to override the system's decisions. (Martelli et al., 2023). further emphasized the need for comprehensive graphical and textual displays on system interfaces. Based on the review research, displayed routes should be accompanied by more information about how a system has reached its decision, while at the same time ensuring operator SA and avoiding information overflow.

Technology

The final transparency group relates to findings directly involving technology to ensure decision transparency.

System situational awareness

Hansen et al. (2020) developed a framework for autonomous SA grounded on discrete event system (DES) theory and integrated Endsley's three-level SA model. The authors proposed deterministic automata to separate a collision avoidance system's understanding of a situation from its anticipation of the near future. Such a division may provide a clear foundation for communicating a system's intentions to operators, and it aligns with the findings of Madsen et al (A. Madsen et al. 2023), who highlighted that a system should provide information about its SA processes. Using a three-level model to separate a system's SA may allow transparency levels to be determined for each SA level. Although (A. Madsen et al., 2023) suggested a trial function for fast-forwarding a recommended maneuver or an on-demand animation of such, (Porathe, 2022) emphasized the importance of an AI system sharing its SA and intentions with the surrounding agents (ships) and suggested that such information obtained from sensors could be broadcast through a web portal, a live map, or similar. Such integration of transparency into decision-making models and dynamic visualizations of systems' SA is prompting a transformative shift in human – machine interactions. These advances may not only enhance operators' understanding of the systems decisions and the actual situation, but also pave the way for more collaborative autonomous systems.

Route exchange

As pointed out by Porathe (Porathe, 2022), there is a need to make MASS intentions transparent to surrounding agents. A possible solution is route exchange (Porathe, 2022; Rødseth et al., 2023), by which ships could exchange their routes and intended deviations of routes through VHF data exchange systems (VDESs), allowing humans and machines to observe others' intentions and incorporate them into their decision-making. This approach seems reasonable and technologically feasible, since it aligns with one of the objectives of the International Hydrographic Office's new S-100 standard for charts (International Hydrographic Organization).

Signaling autonomy

Besides route exchange, there may be ways to enhance interactions between manned and unmanned ships. (Porathe, 2022) emphasized that humans tend to attribute human traits, emotions, and/or intentions to nonhuman entities, meaning that there is a risk of navigators on manned ships expecting MASSs to behave like humans. This aligns with the findings of (Van de Merwe et al., 2022), who found that navigators trust AI systems more if they behave as they would themselves. However, humans are less predictable than machines, and if a MASS strictly follows the rules, the ship's behavior is transparent. To signal to manned ships, an MASS should be able to send an individual identification signal, such as via a lantern or an Automatic Identification system (AIS) icon (Porathe, 2022; Rødseth et al., 2023).

Conclusions and recommendations on further research

Regarding the future of autonomous navigation, decision transparency emerges as a crucial bridge between historical navigational wisdom and modern technological advances. Based on this review, we discovered that AI decision transparency for autonomous ships is developing in terms of both theoretical clarity and practical utilization, and the topic is attracting considerable attention. The results indicated a need to continue using the established techniques of design theory in the quest for optimal human – AI collaboration. However, the literature review also revealed a surprising lack of focus on onboard decision transparency, since the research focus seems to be mostly on ROCs rather than on manned MASSs. It is reasonable to assume that AI for collision avoidance would be equally, if not more, effective for manned ships. We categorized the review findings regarding decision transparency into three main groups: strategies, visualization, and technology. Strategies involve human factors, risk assessment, and design principles, highlighting the importance of understanding operators' roles and building trust in AI systems. Visualization techniques, including color coding, bounding boxes, and route displays, aim to bridge the gap between complex AI decisions and human comprehension. Technology solutions, such as SA frameworks and route exchange mechanisms, contribute to advancing collaboration between humans and autonomous systems.

Although many of the publications included in this review used the term SA, few discussed or provided clear definitions of it. Moreover, none of the studies identified in this review evaluated how proposed strategies, visualization, or technology affect systems' or individual operators' SA. Further research should concentrate on discovering how the approaches identified in this review affect operators' and AI systems' SA in distributed contexts.

Disclosure statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by the Research Council of Norway: Center for Research-based Innovation AutoShip [project number 309230]

ORCID

A. N. Madsen http://orcid.org/0000-0003-3997-2096
T. E. Kim http://orcid.org/0000-0002-9339-2933

References

Abu-Tair, M., & Naeem, W. (2013). A decision support framework for collision avoidance of unmanned maritime vehicles. *Communications in Computer and Information Science*, *355*, 549–557. https://doi.org/10.1007/978-3-642-37105-9_61

Ban, H., & Kim, H. J. (2023). Analysis and visualization of vessels' relative motion (REMO). *International Journal of Geo-Information*, *12*(3). https://doi.org/10.3390/ijgi12030115

Braun, V., & Clarke, V. (2012). Thematic analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), APA Handbook of Research Methods in psychology vol. 2: Research designs: Quantitative, qualitative, neuropsychological, and biological (Vol. 2, pp. 57–71). American Psychological Association.

Cheng, T., Utne, I. B., Wu, B., & Wu, Q. (2023, March). A novel system-theoretic approach for human-system collaboration safety: Case studies on two degrees of autonomy for autonomous ships. *Reliability Engineering & System Safety*, *237*, 109388. https://doi.org/10.1016/j.ress.2023.109388

Dugan, S., Skjetne, R., Wrobel, K., Montewka, J., Gil, M., & Utne, I. B. (2023). Integration test procedures for a collision avoidance decision support system using STPA. *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, *17*(2), 375–381. https://doi.org/10.12716/1001.17.02.14

Hansen, P. N., Papageorgiou, D., Blanke, M., Galeazzi, R., Lutzen, M., Mogensen, J., Bennedsen, M., & Hansen, D. (2020). Colregs-based situation awareness for marine vessels: A discrete event systems approach. *IFAC-Papers Online*, *53*(2), 14501–14508. https://doi.org/10.1016/j.ifacol.2020.12.1453

Huang, Y., Chen, L., Negenborn, R. R., & van Gelder, P. H. A. J. M. (2020). A ship collision avoidance system for human-machine cooperation during collision avoidance. *Ocean Engineering*, *217*, 107913. [Online]. Available. https://doi.org/10.1016/j.oceaneng.2020.107913

International Chamber of Shipping. (2022). *Bridge procedures Guide* (6th ed.). International Chamber of Shipping Publications. https://www.ics-shipping.org/publications/single-product.php?id=58

"International Hydrographic Organization" [Online]. Available: https://registry.iho.int/productspec/view.do?idx=185&product_ID=S-421&statusS=5&domainS=ALL&category=product_ID&searchValue=

Kahneman, D., Tversky, A., & Tversky', A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–292. https://doi.org/10.2307/1914185

Lynch, K. M., Banks, V. A., Roberts, A. P. J., Radcliffe, S., & Plant, K. L. (2022). Maritime autonomous surface ships: Can we learn from unmanned aerial vehicle incidents using the perceptual cycle model? *Ergonom*, *66*(6), 772–790. https://doi.org/10.1080/00140139.2022.2126896

Madsen, A. N., Aarset, M. V., & Alsos, O. A. (2022, January). Safe and efficient maneuvering of a maritime autonomous surface ship (MASS) during encounters at sea: A novel approach. *Maritime Transport Research*, *3*, 100077. https://doi.org/10.1016/J.MARTRA.2022.100077

Madsen, A., Brandsæter, A., & Aarset, M. V. (2023). Decision transparency for enhanced human-machine collaboration for autonomous ships. *Human Fact Robot, Drones Unman System*, *93*, 76–84. https://doi.org/10.54941/ahfe1003750

Man, Y., Weber, R., Cimbritz, J., Lundh, M., & MacKinnon, S. N. (2018, April). Human factor issues during remote ship monitoring tasks: An ecological lesson for system design in a distributed context. *International Journal of Industrial Ergonomics*, *68*, 231–244. https://doi.org/10.1016/j.ergon.2018.08.005

Martelli, M., ŽuŁkin, S., Zaccone, R., and Rudan, I. (2023). A COLREGs-compliant decision support tool to prevent collisions at sea. *TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation*, *17*(2), 347–353. https://doi.org/10.12716/1001.17.02.11

Ożoga, B., & Montewka, J. (2018, April). Towards a decision support system for maritime navigation on heavily trafficked basins. *Ocean Engineering*, *159*, 88–97. https://doi.org/10.1016/j.oceaneng.2018.03.073

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E. … Whiting, P. (2021, March). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ: British Medical Journal*, *372*, n71. https://doi.org/10.1136/BMJ.N71

Pietrzykowski, Z., Magaj, J., and Wielgosz, M. (2018). Navigation decision support for sea-going ships in port approach areas. *Scientific Journals of the Maritime University of Szczecin*, *126*(54), 75–83. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=131649089&site=ehost-live

Pietrzykowski, Z., Wołejsza, P., & Borkowski, P. (2017). Decision support in collision situations at sea. *Journal of Navigation* [[Online]. Available], *70*(3), 447–464. https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=122423031&site=ehost-live

Porathe, T. (2022). Safety of autonomous shipping: COLREGs and interaction between manned and unmanned ships maritime unmanned navigation through intelligence in networks view EffcienSea 2 view project. *European Safety and Reliability Conference (ESREL) Dublin*. https://doi.org/10.3850/978-981-11-2724-3_0655-cd

Rødseth, Ø. J., Wennersberg, L. A. L., & Nordahl, H. (2023, June). Improving safety of interactions between conventional and autonomous ships. *Ocean Engineering*, *284*, 115206. https://doi.org/10.1016/j.oceaneng.2023.115206

Rutledal, D., Relling, T., & Resnes, T. (2020, November). It's not all about the COLREGs: A case-based risk study for autonomous coastal ferries. *IOP Conference Series: Materials Science and Engineering*, *929*(1), 929 012016. https://doi.org/10.1088/1757-899X/929/1/012016

Utne, I. B., Rokseth, B., Sørensen, A. J., & Vinnem, J. E. (2020). Towards supervisory risk control of autonomous ships. *Reliability Engineering & System Safety*, *196*, 106757. https://doi.org/10.1016/j.ress.2019.106757

Van de Merwe, K., Mallam, S. C., Engelhardtsen, O., & Nazir, S. (2022). Exploring navigator roles and tasks in transitioning towards supervisory control of autonomous collision avoidance systems. *Journal of Physics Conference Series*, *2311*(1), 012017. https://doi.org/10.1088/1742-6596/2311/1/012017

Van de Merwe, K., Mallam, S., Engelhardtsen, Ø., & Nazir, S. (2023). Operationalising automation transparency for maritime collision avoidance. *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, *17* (2), 333–339. https://doi.org/10.12716/1001.17.02.09

Veitch, E., & Alsos, O. A. (2021). Human centered explainable artificial intelligence for marine autonomous surface vehicles. *Journal of Marine Science and Engineering*, *9*(11), 1227. https://doi.org/10.3390/jmse9111227

Veitch, E., & Alsos, O. A. (2022). A systematic review of human–AI interaction in autonomous ship systems. *Safety Science*, *152*, 105778. [Online]. Available. https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=157031110&site=ehost-live&scope=site.

Veitch, E., Dybvik, H., Steinert, M., & Alsos, O. A. (2022). Collaborative work with highly automated Marine navigation Systems, no. 2007. Springer. https://doi.org/10.1007/s10606-022-09450-7

Wu, X., Liu, K., Zhang, J., Yuan, Z., Liu, J., & Yu, Q. (2021). An optimized collision avoidance decision-making system for autonomous ships under human-machine cooperation situations. *Journal of Advanced Transportation*, *2021*, 1–17. [Online]. Available. https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=152124374&site=ehost-live.

Zhao, L., Sunilkumar, S. R. T., Wu, B., Li, G., & Zhang, H. (2023). Toward an online decision support system to improve collision risk assessment at sea. *IEEE Intelligent Transportation Systems Magazine*, *15*(2), 137–148. https://doi.org/10.1109/MITS.2022.3190965

Appendix A

ID	Authors	Year	Title	Country	Design	Outcomes	Main Group Subgroup
1	M. Abu-Tair and W. Naeem	2013	"A decision support framework for collision avoidance of unmanned maritime vehicles"	United Kingdom	Hardware testing	The article discusses an AI system that employs color coding to differentiate targets within its 180° real-time line of sight using an object detection system. Identified vessels appear in green, while those not in the line of sight are marked in red using bounding boxes derived from the camera's perspective	**Visualization** Color coding Bounding boxes
2	P. N. Hansen, et al.	2020	"COLREGs-based situation awareness for marine vessels: A discrete event systems approach"	Denmark	Simulation	The authors created an autonomous Situational Awareness (SA) framework based on discrete event system (DES) theory and integrating Endsley's three-level model of SA. They introduced deterministic automata to distinguish the system's understanding of a situation from its anticipation of the near future.	**Technology** System SA
3	Y. Huang, L. Chen, R. R. Negenborn and P. H. A. J. M van Gelder	2020	"A ship collision avoidance system for human-machine cooperation during collision avoidance"	China	Simulation	The authors suggested a collision avoidance system for human-system cooperation. It presented what the system considered as optimal optimal decisions, flagged risky solutions, enabling human operators to confirm system recommendations or propose/validate alternatives in case of discrepancies	**Strategies** Design principles
4	K. M. Lynch, V. A. Banks, A. P. J. Roberts, S. Radcliffe and K. L. Plant	2022	Maritime autonomous surface ships: Can we learn from unmanned aerial vehicle incidents using the perceptual cycle model?"	UK	Perceptual cycle model, case study	The authors utilized a perceptual cycle model from the aviation domains and extended the insights to autonomous ships. They demonstrated that, for effective collision avoidance, a well-designed human-machine interface is crucial. It should enable operators to interpret information accurately, receive collision alerts, and take necessary actions	**Strategies** Human factors Design principles
5	Y. Man, R. Weber, J. Cimbritz, M. Lundh and S. N. MacKinnon	2018	"Human factor issues during remote ship monitoring tasks: An ecological lesson for system design in a distributed context"	Sweden	Scenario-based simulation	The study revealed challenges when operators utilize widely available navigation and collision avoidance technologies in diverse settings for remote supervisory control tasks. Without adapting tools to the specific domain constraints of shore-based remote monitoring and control, operators face difficulties making timely decisions and performing tasks reliably.	**Strategies** Human factors Design principles

(*Continued*)

(Continued).

ID	Authors	Year	Title	Country	Design	Outcomes	Main Group Subgroup
6	B. Ożoga and J. Montewka	2018	"Towards a decision support system for maritime navigation on heavily trafficked basins"	Poland	Model development	The paper outlined a MARPA system through case studies, featuring an algorithm identifying hazards and suggesting maneuvers for ships encountering others. The authors proposed a color scheme based on TCPA and CPA, using five colors to denote statuses from no hazard to extreme danger. These were displayed via overlay on ARPA or ECDIS, employing color coding to emphasize safe sectors for each maneuver	**Visualization** Color coding
7	Z. Pietrzykowski, P. Wołejsza and P. Borkowski	2017	Decision support in collision situations at sea"	Poland	Software module development	The authors condensed research on Navigational Decision (NAVDEC) support systems. The NAVDEC system shows avoidance maneuvers as routes from the current position to the next waypoint, facilitating passing other ships at an assumed CPA. A color-coded rosette on the top bar of NAVDEC depicts safe sectors for desired maneuvers.	**Visualization** Color coding Route display
8	Z. Pietrzykowski, J. Magaj and M. Wielgosz	2018	"Navigation decision support for sea-going ships in port approach areas"	Poland	Conceptual paper	The authors investigated the necessary attributes for decision support system for port approach areas, compared to those on the market designed for the open sea. They noted significant differences in the ship domain in these zones. They emphasized the need for systems capable of operating effectively in port approach areas with heavier traffic.	**Strategies** Design principles
9	T. Porathe	2020	"Safety of autonomous shipping: COLREGs and interaction between manned and unmanned ships"	Norway	Conceptual paper	Assuming anthropomorphism, the author argued that MASS should send individual identification signals using lanterns or AIS icons. An AI system should share information about its situational awareness (SA) and intentions, with data from sensors broadcast through a web portal or live map. They proposed the exchange of sailing routes among ships.	**Technology** System SA Route exchange Identification
10	I. B. Utne B. Rokseth A. J. Sørensen and J. E. Vinnem	2020	"Towards supervisory risk control of autonomous ships"	Norway	Conceptual paper	The authors suggested an online risk model for decision support in autonomous ship control systems, addressing environmental and operational conditions. Proactively, it offers early warnings on potential violations of safety constraints in the approved operation. Reactively, it provides human operators and supervisors with additional time for efficient responses and crisis interventions, relying on predictions of potential outcomes.	**Strategies** Risk management

(*Continued*)

(Continued).

ID	Authors	Year	Title	Country	Design	Outcomes	Main Group Subgroup
11	G. K. Van de Merwe, S. C. Mallam, O. Engelhardtsen and S. Nzir	2022	"Exploring navigator roles and tasks in transitioning towards supervisory control of autonomous collision avoidance systems"	Norway	Case study, interviews, and cognitive task analysis	The authors employed a systematic task analysis, integrating COLREGs, procedures, navigator input, and collision avoidance observations to define maneuvers and performance requirements. They discovered that navigators trust systems mirroring their behavior. The authors argued that establishing "compatibility" between human and system behaviors is crucial for building trust in AI systems. They also highlighted the importance of system transparency to help operators comprehend AI's information processing, decision-making, and future actions.	**Strategies** Human factors
12	(A) Veitch and O. A. Alsos	2021	"Human-centered explainable artificial intelligence for marine autonomous surface vehicles"	Norway	Conceptual paper	The authors emphasized the development of end-user interaction designs for autonomous ships lacking a fully formulated conceptualization of design practices. They introduced the concept of human-centered Explainable Artificial Intelligence (XAI) for interaction design to advance Autonomous Surface Vehicle (ASV) design.	**Strategies** Design principles
13	(A) Veitch. H. Dybvik, M. Steinert and O. A. Alsos	2022	"Collaborative work with highly automated marine navigation systems"	Norway	Empirical study, interviews, and field observations	This study revealed disparities between designers' constructs of human-AI collaboration and navigators' real-world experiences. The authors identified that effective collaboration with AI systems relied on enhancing the visibility of computational activities.	**Strategies** Design principles
14	X. Wu, K. Liu, J. Zhang, Z. Yuan, J. Liu and Q. Yu	2021	"An optimized collision avoidance decision-making system for autonomous ships under human-machine cooperation situations"	China	Model development and case study	In a case study, the authors demonstrated the incorporation of operators' risk preferences into a Risk-Appetite Collision Avoidance Decision-Making System (RA-CADMS). They applied prospect theory to systematically explain human behavior and integrate it into the decision-making system model.	**Strategies** Human factors
15	L. M. Zhao, S. R. T. Sunilkumar, B. H. Wu, G. Y. Li and H. X. Zhang	2022	"Toward an online decision support system to improve collision risk assessment at sea"	Norway	Software development	The authors developed and tested a decision support system using simulators with nautical students. The design incorporated color-coding each target ship based on collision risk and highlighted potential collisions with red circles representing the ship domain.	**Visualization** Color coding Route display

(*Continued*)

(Continued).

ID	Authors	Year	Title	Country	Design	Outcomes	Main Group Subgroup
16	H. Ban H. and H.-J. Kim	2023	"Analysis and visualization of vessels' relative motion (REMO)"	USA, Korea	Software development	The authors evaluated a maritime traffic control system with VTS operators, with results applicable to remote operators. Findings showed that operators required multiple color schemes for effective visualization of spatial data on a chart. This underscores the importance of customizing color coding to meet the specific needs of operators.	**Visualization** Color coding
17	T. Cheng, I. B. Utne, B. Wu and Q. Wu	2023	"A novel system-theoretic approach for human-system collaboration safety: Case studies on two degrees of autonomy for autonomous ships"	Norway	Case study	The authors proposed a System-Theoretic Process Analysis (STPA) concentrating on safe human-system interactions. They suggested a system theoretic approach to safety analysis for human-system collaboration, aiming to systematically identify hazards related to human cognition and interactions. The goal is to enhance the design of shore control centers for autonomous ships.	**Strategies** Risk management
18	S. A. Dugan, R. Skjetne, K. Wróbel, J. Montewka, M. Gil and I. B. Utne	2023	"Integration test procedures for a collision avoidance decision support system using STPA"	Norway	Model testing	This study showcased the application of System-Theoretic Process Analysis (STPA) to analyze system behavior and pinpoint test cases for system operation. Verification requirements for critical systems were categorized into failure handling and integration testing. The method was specifically applied to a decision support system for collision avoidance, focusing on stability monitoring.	**Strategies** Risk management
19	M. Martelli, S. Žuškin, R. Zaccone and I. Rudan	2023	"A COLREGs-compliant decision support tool to prevent collisions at sea"	Italy, Croatia	System architecture	The authors assessed a decision support system architecture designed to recommend actions upon detecting potential collisions. The system displayed safety assessments through a graphical user interface in both graphical and textual formats.	**Visualization** Route display
20	Ø. J. Rødseth, L. A. L. Wennersberg and H. Nordahl H	2023	"Improving safety of interactions between conventional and autonomous ships"	Norway	Case study	The authors focused on interactions between manned and unmanned ship and limits the study to four categories of problems in these interactions; Information acquisition, situation assessment, other ship predictions and planning and execution of actions.	**Technology** Route exchange Identification

(*Continued*)

(Continued).

ID	Authors	Year	Title	Country	Design	Outcomes	Main Group Subgroup
21	K. van de Merwe, S. Mallam, Ø. Engelhardtsen and S. Nazir	2023	"Operationalizing automation transparency for maritime collision avoidance"	Norway	Empirical study and interviews	The authors concentrated on interactions between manned and unmanned ships, narrowing the study to four problem categories: information acquisition, situation assessment, predictions about other ships, and planning and execution of actions.	**Strategies** Design principles **Visualization** Color coding Bounding boxes
22	A. Madsen, A. Brandsæter, and V. Aarset Magne	2023	"Decision transparency for enhanced human-machine collaboration for autonomous ships"	Norway	Empirical study, interviews, and simulator observation	In a simulator study, the authors showcased system-recommended maneuvers as routes on an ECDIS. The adequacy of ensuring Situational Awareness (SA) depended on circumstances. They recommended the system to convey how it built its SA and suggested a fast-forward option for recommended maneuvers. Additionally, the authors emphasized the importance of the system indicating targets considered using bounding boxes.	**Strategies** Design principles **Visualization** Color coding Bounding boxes **Technology** System SA