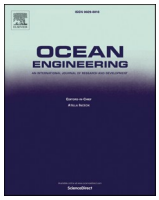
Ocean Engineering 325 (2025) 120753

Contents lists available at ScienceDirect

Ocean Engineering

journal homepage: www.elsevier.com/locate/oceaneng



Research paper

Potential of explanations in enhancing trust – What can we learn from autonomous vehicles to foster the development of trustworthy autonomous vessels?

Rohit Ranjan [b], Ketki Kulkarni [a,c], Mashrura Musharraf [a,*]

[a] Marine and Arctic Technology, Department of Mechanical Engineering, Aalto University, Finland
[b] Department of Computer Science, Indian Institute of Technology Kharagpur, India
[c] HUMLOG Institute, Supply Chain Management and Social Responsibility, Hanken School of Economics, Finland

ARTICLE INFO

Keywords:
Autonomous vessels
Explainable AI
Trustworthiness
Systematic literature review

ABSTRACT

The development of autonomous vessels presents a complex socio-technical challenge where AI and humans must coexist and cooperate. A crucial aspect of successfully deploying these systems is ensuring trust in the AI-powered autonomy. Our research aims to explore the potential of explanations in enhancing trust and its correlated metrics (such as preference, understanding, anxiety) in autonomous vessels. While the investigation of explainability and its role in increasing end-user trust is still at an elementary level for autonomous vessels, it has already been identified as a key requirement for successful adoption of self-driving cars and highly automated vehicles in general. We conducted a systematic literature review to investigate how the impact of explainability on trust and its correlated metrics has been studied in the domain of autonomous vehicles. We examined the diverse experimental setups employed to assess trust-building, exploring instruments, explanation modes, types, timings, and additional human factors influencing trust. The study scrutinizes prevalent data collection methods and commonly used questionnaires for measuring trust levels following explanations and examines the characteristics and theories integral to effective explanations for trust development. Review results indicate that explanations generally have a positive impact on trust and its correlated metrics *preference*, although this impact is not statistically significant in all cases. The effect of explanations on correlated metrics *understanding* was found to be statistically significant in all cases. For correlated metrics *anxiety*, a decrease was observed with the presence of explanations in most cases, even though this decrease wasn't always statistically significant. This study discusses how lessons learned from autonomous vehicles can be applied in the context of autonomous vessels, with the aim of fostering the development of trustworthy autonomous vessels.

1. Introduction

With the emergence of artificial intelligence (AI) technologies and their growing role in decision-making processes, it is important to understand the perceived risks and develop policies and interventions to ensure safety. Intelligent automated vehicles, which are one of the most transformative applications of AI, continue to face barriers such as distrust in the technology. Worldwide 65% of people feel unsafe being driven in self-driving cars (Foundation, 2021). It is anticipated that this perception would translate to the maritime domain as well (Mallam et al., 2020).

Autonomous vessels are defined as ships that operate without human intervention, either through remote control or by following pre-programmed routes and making real-time decisions based on data analysis (Łosiewicz and Mironiuk, 2020). The advancement of technology towards autonomous vessels is largely driven by AI, particularly Machine Learning (ML) and data-driven models. However, the introduction of data-driven decision support systems brings forth new challenges to assurance and safety of unmanned vessels (Brandsæter et al., 2020). Even though these systems show promising results, their applicability to real-life problems is reduced by the lack of understanding of how the underlying ML models such as neural networks make their decisions. The complexity of these models, characterized by numerous parameters and interconnections, renders them "black boxes" that are hard for humans to interpret (Gjærum et al., 2021a,b). This inherent opacity, coupled with the incompleteness of training datasets and the

* Corresponding author. Marine and Arctic Technology, Department of Mechanical Engineering, Aalto University, Finland.
E-mail addresses: rohit_ranjan@kgpian.iitkgp.ac.in (R. Ranjan), mashrura.musharraf@aalto.fi (M. Musharraf).

https://doi.org/10.1016/j.oceaneng.2025.120753
Received 17 March 2024; Received in revised form 2 December 2024; Accepted 20 February 2025
Available online 1 March 2025

potential for systematic biases (Salay et al., 2017), raises significant concerns about their suitability and assurance in unfamiliar maritime environments (Guidotti et al., 2018).

Overall, the awareness and understanding of the choices, values, and operating logic of the ML/AI reliant automated systems have been identified to be crucial for the practical implementation of autonomous vessels (Mallam et al., 2020). van de Merwe et al. (2024) conducted a comprehensive review on the human factor implications of automation transparency in safety-critical domains, demonstrating the positive impact of making an automation's inner workings observable on situation awareness and operator performance without adding to mental workload. Similarly, Houweling et al. (2024) leveraged Augmented Reality (AR) to enhance navigators' understanding of their environment, which improved operator performance.

A systematic way to better understand how the ML/AI models make decisions is with explainable AI (XAI), defined as a set of processes and methods focused on making AI systems more interpretable and understandable to humans. By explaining the models' choices, values, and operating logic the XAI methods can possibly foster confidence and trust (Gjærum et al., 2021a,b; Singh et al., 2023; Veitch and Alsos, 2021). Trust, in this context, is a multifaceted concept that involves several dimensions such as the system's perceived competence, safety, and reliability (Shafi, 2017; Chaal et al., 2023). Veitch and Alsos (2021) highlighted a human-centered approach to XAI, emphasizing the importance of making AI technology understandable and interpretable for different user groups to build trust in autonomous surface vessels. This involves representing the capabilities and constraints of the AI system in line with the interaction needs of developers, primary users, and secondary users. On the contrary, Glomsrud et al. (2019) contended that trustworthy AI in the context of autonomy is a broader concept than XAI and that trustworthiness depends on successfully matching the needs of different users with the various types of explanations required.

Although various XAI models have been proposed to enhance the trustworthiness of autonomous vessel systems (Chowdhury et al., 2022; Gjærum et al., 2021a,b), there is a notable lack of validation regarding how these explanations affect real-life trust building. Establishing concrete evidence on whether explanations effectively build trust in autonomous vessels within the maritime domain is crucial before widely adopting these XAI models. Addressing this gap, our research aims to investigate validated measures that assess the real-world impact of explainability on trust and its correlated metrics in autonomous systems. We scrutinize prevalent methods for the empirical measurement of trust and its correlated metrics to provide a standard for validation of explainability.

To achieve this, we draw upon advancements made in XAI within the context of autonomous vehicles, which have been studied more thoroughly than autonomous vessels, as evidenced by the significantly higher volume of research publications. Furthermore, autonomous vehicles have been available in the consumer domain for the past decade, offering a valuable opportunity to validate the effects of explainability on end-user trust. While there are unique requirements and challenges associated with the maritime domain and autonomous vessels, findings from the domain of autonomous vehicles offer valuable insights on methodological approaches that can be adapted to enhance our understanding of the real-life impact of explainability on trust and its correlated metrics in autonomous systems at sea. Given that 1) understanding the impact of explainability on trust and its correlated metrics is a fundamental research question for all autonomous vehicles, and 2) autonomous vessels are specific implementations of self-driving technology designed for waterborne transportation and can be considered a subclass of autonomous vehicles, our hypothesis is that knowledge transfer from the automotive to the maritime domain is both logical and practical.

Through a systematic literature review encompassing studies on explainability and trust in autonomous vehicles we aim to answer the following research questions (RQ).

RQ1: What experimental designs have been employed for the evaluation of trust-building in autonomous vehicles? This leads to the subsequent questions.

- What are the prevalent experimental setups - including the instruments, explanation modes, types, and timings - in these studies?
- What additional human factors correlate with trust and act as surrogate metrics in the study of autonomous vehicles?
- What data is collected and what are the prevalent data collection methods? Which questionnaires have been commonly used to measure trust levels in autonomous vehicles following explanations? What are the prevalent timings of administering these questionnaires?

RQ2: What is the impact of explainability on trust and its correlated metrics in autonomous vehicles?

RQ3: What characteristics and theories form the constituents of good explanations? This leads to the subsequent questions.

- What explanation types and timings contribute to building trust in autonomous vehicles?
- What theoretical frameworks have been proposed for effective explanations in the context of autonomous vehicles?

Our work contributes to the existing literature by addressing the limitations of previous studies, such as those by Omeiza et al. (2021)a, b , c and Zhang et al. (2021). Omeiza et al. (2021) reviewed explanations in autonomous driving and categorized the approaches used to develop and evaluate them. Their work briefly categorizes many explanations under "Unvalidated Guidelines," highlighting that most explanations proposed for increased trust and transparency have been based on authors' experiences without substantial validation for trust building. Our study exclusively reviews literature on validated measures that assess the real-world impact of explainability on trust and the relevant empirical methods. We define "validated evaluation" of trust as the adoption of established research instruments previously utilized to measure trust, while "empirical evaluation" of trust refers to methods specifically crafted for the reviewed paper, based solely on the authors' empirical guidelines. Zhang et al. (2021) discuss the impacts of explainability on autonomy in terms of trust and acceptance but do not differentiate between results from validated and empirical evaluations of trust. In contrast, we review the impact of explanations on trust with a focus on how the trust levels were evaluated. By considering validated guidelines, evaluating realistic effects, and analyzing a comprehensive range of studies, our review provides a rigorous assessment of the relationship between explainability and trust in autonomous vehicles. This research offers valuable insights into the design, evaluation, and practical implications of explanations for building trust in autonomous vessels.

In summary, the goal of this paper is to systematically review literature that uses some validated measures to examine the use of explanations to enhance trust and its correlated metrics in autonomous vehicles. We then see if and how the lessons learned from autonomous vehicles can be applied in the context of autonomous vessels. Different aspects of the literature such as diverse experimental setups (including experimental instruments, explanation modes, types, and timing), correlated trust metrics, and data collection methods are reviewed. We examine the characteristics and theories integral to effective explanations for trust development. The paper provides a critical review of the statistical analysis (e.g., significance testing) done by the authors of the reviewed papers when available. However, performing our own statistical analysis where they were unavailable or insufficient was out of the scope of the paper.

The remainder of the article is structured as follows. Section 2 describes the methodology used in conducting the systematic literature review including the guidelines and keywords used. Section 3 discusses the results of the three research questions and related sub-questions. Finally, Section 4 summarizes the key takeaways for autonomous

vessels and concludes the article.

2. Methodology

This systematic literature review was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to ensure a systematic and transparent approach, promoting reproducibility. It focuses on validated guidelines for explanations in the context of autonomous driving systems to promote trust building. The scope is limited to explanations related to the task of autonomous driving and autonomous vehicle features.

The review process involved two independent reviewers to ensure the robustness and reliability of the literature analysis. The research questions identified in Section 1 guided our review of the literature. Additionally, we established preliminary frameworks for expected answers, which we refined as the review progressed. One reviewer conducted a comprehensive review and interpretation of all the articles. In cases where the findings were difficult to interpret or ambiguous, the second reviewer was consulted to provide additional insights and resolve any uncertainties. For the organization and analysis of the literature, we utilized Excel spreadsheet to facilitate consistent and efficient data management.

The literature search was conducted using two scientific databases, Scopus and Web of Science Core Collection. These databases offer comprehensive coverage of scholarly articles across various disciplines. The databases included numerous AI and Autonomous Vehicles related academic journals such as the Journal of Intelligent and Connected Vehicles, the International Journal of Vehicle Autonomous Systems, and the International Journal of Intelligent Systems. The type of publication was limited to peer-reviewed journal articles and conference papers and did not include preprints such as arXiv. This review considers papers published between January 2015 and January 2023. This is done to focus on the latest notions of levels of autonomy and keep our findings about explanations up to date with the latest autonomous vehicle capabilities. The Society of Automotive Engineers (SAE) automation categorization (SAE International, 2018) is the predominant categorization model used by automotive engineers and has been adhered to in our work. SAE's levels of autonomy range from Level 0, where there is no driving automation, to Level 5, where vehicles achieve full driving automation. Each level represents a progressive degree of automation with specific criteria defining the extent of vehicle control and the role of the driver in the driving process.

The search strategy involved the use of appropriate keywords and search terms derived from the research questions. These were "explanation", "trust", "effect", and "autonomous vehicle". The final search string as shown in Fig. 1, was applied to search on paper title, abstract, keywords and citations if supported by the database. Fig. 2 shows the various stages of the search procedure.

Using the Scopus database, the initial search retrieved a total of 2933 articles. Refinement filters for publication year, publication type and keywords reduced the articles to 1384 for further screening. The Web of Science database yielded 37 articles after applying similar search criteria. During the title screening phase, each article was evaluated for relevance to explainable AI, explanations, trust, and autonomous vehicles. Abstracts were examined in cases of ambiguity. Only articles specifically focusing on the impact of explanations or explainable AI on trust in the context of autonomous vehicles were included. Articles exploring trust in other domains or examining different aspects of autonomous vehicles unrelated to trust were excluded.

After the title screening and duplicate removal, a total of 47 articles remained. These articles underwent abstract analysis and light paper examination to further refine the selection to 23 articles. Additionally, citation chaining was performed for these articles. Finally, 25 articles were deemed relevant for inclusion in the literature review. These articles were thoroughly analyzed to extract key information, including the year and type of publication, level of autonomy based on the SAE classification, tested explanation types and timings, explanation mode, hypothesized theory of effective explanation, and experimental methodology.

3. Review results and discussion

This section presents the results and discussion based on the analysis of the selected articles.

3.1. Overview

Of the reviewed articles, 44% were published in scientific journals, and 56% were in conference proceedings. The distribution of publications per year is shown in Fig. 3. The studies predominantly follow the pattern of having users/participants experience autonomous driving with and without explanations and then answer questionnaires to determine levels of trust and its correlated metrics. The median sample size for the autonomous driving experiment among the reviewed articles is 40. Goldman and Bustin (2022) had the highest sample size of 2586 participants, followed by Avetisyan et al. (2022) with 340 participants. The distribution of the level of autonomy used to test explanations is shown in Fig. 4. Twelve out of the 25 papers reviewed study explanations with fully autonomous vehicles as described by Level Five of SAE International, 2018 guidelines. Out of these, five papers explicitly state their tested autonomy level to be at Level five, whereas the rest (shown as 5* in Fig. 4) have been inferred from the autonomous capabilities supported in test setup.

3.2. Experimental design

This section addresses RQ1 and analyzes the experimental designs of the reviewed studies. This includes the discussion of 1) experimental setups, 2) additional human factors correlating with trust, and 3) types and timings of questionnaire use.

3.2.1. Experimental setup

This section primarily focuses on the variety of experimental instruments used including driving simulators, first-person videos, virtual reality, and real-world driving scenarios. The discussion covers the incorporation of secondary non-driving tasks. Besides the instruments,

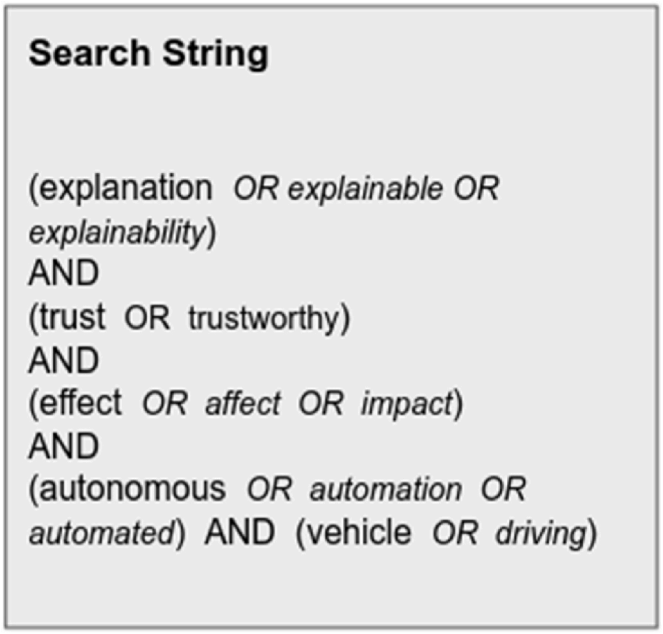


Fig. 1. The search string.

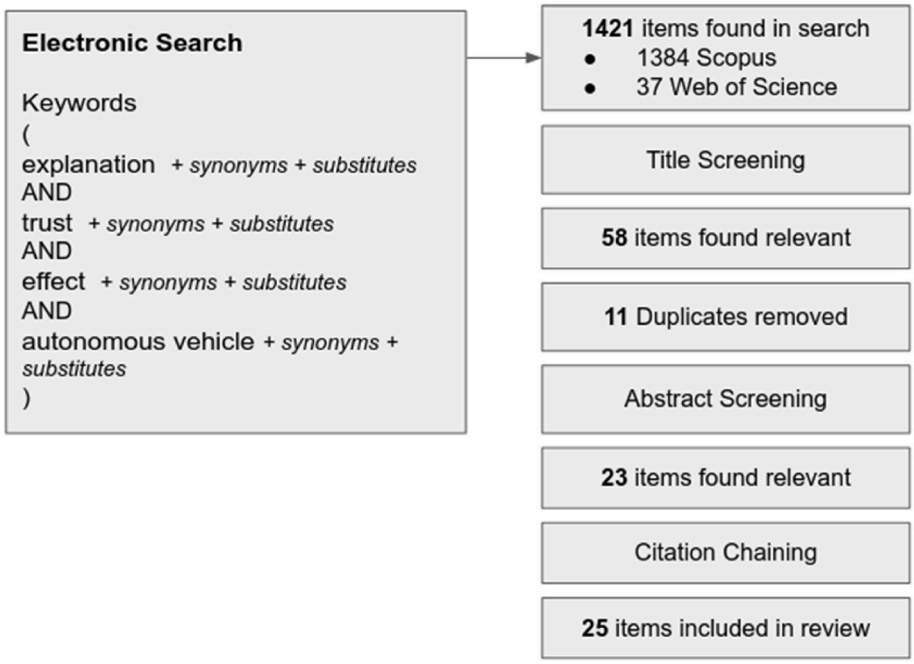


Fig. 2. Procedure for constructing dataset of relevant literature.

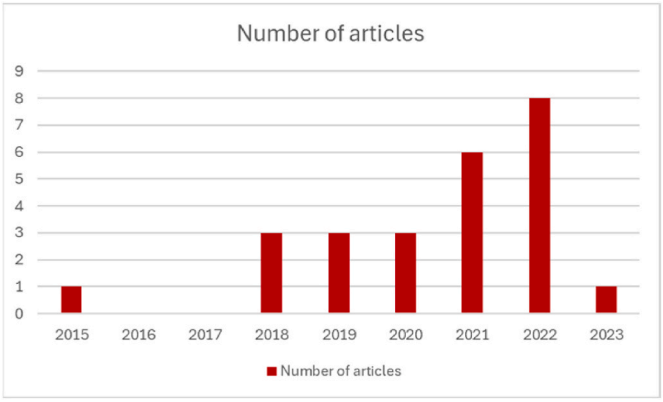


Fig. 3. Distribution by year of publication.

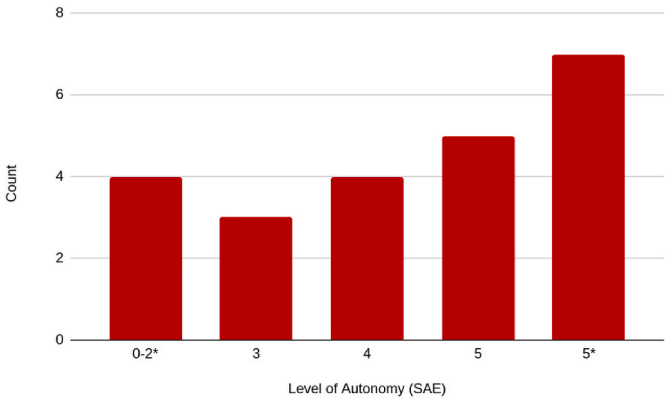


Fig. 4. Distribution of level of autonomy.

the experimental settings such as explanation modes, types, and timings are also discussed in this section.

3.2.1.1. Experimental instrument. Fig. 5 shows the variety of experimental instruments used to simulate autonomous driving for participants. Four types of instruments were used: full-fledged driving simulators, first-person driving videos, sequences of driving scene images, and virtual reality. Full-fledged driving simulators (14 papers) seek to simulate the complete driver experience in a vehicle. Within this category, Du et al. (2019) utilized a simulator built inside a real car while screens surrounding the car simulated traffic scenarios. Others like Petersen et al. (2019) used commercially available driving rigs for gaming and training purposes which consist of a steering rack, accelerator/brake pedals, and surrounding screens. In first person driving videos (eight papers) participants were exposed to videos displayed on a screen, providing a simulated experience of being inside an autonomous vehicle. Videos were gathered either from real-world driving or by recreating traffic scenarios using simulator software. This approach allowed participants to experience autonomous driving in an online setting, as seen in Schneider et al. (2023), resulting in larger sample sizes. In two papers, participants were presented with sequences of driving scene images, accompanied by explanatory captions. In two

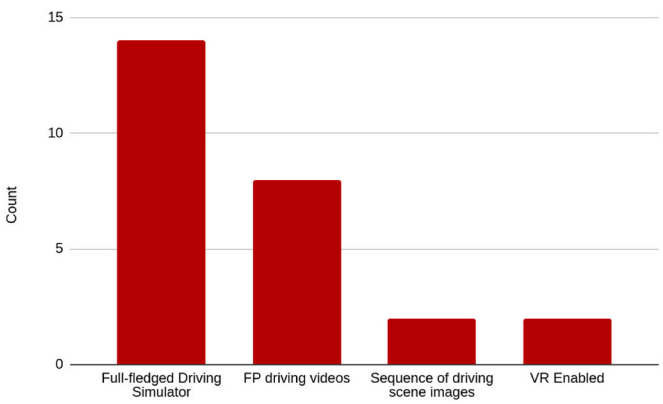


Fig. 5. Distribution of experimental instruments.

other papers, virtual reality headsets were used to create a complete 360-degree autonomous driving experience. Meanwhile, Omeiza et al. (2022) implemented an experimental setup where a professional driving instructor conducted real-life traffic sessions while offering think-aloud explanations. Notably, Lundberg et al. (2022) opted to survey participants without providing explicit autonomous driving experiences.

Four publications incorporated secondary non-driving tasks. Participants were asked to complete these tasks on an additional screen while experiencing autonomous driving as a measure of attention required towards the road. Körber et al. (2018) and Petersen et al. (2019) used versions of the Surrogate Reference Task, ISO 2012, in which participants are required to identify a visual target item amid a field of distractors. The task imposes a controllable level of cognitive load and resembles an ordinary activity like interacting with an infotainment system or smartphone. In Wiegand et al. (2019), participants were shown a video, and they counted people therein. In Du et al. (2021), participants had to chat with the experimenters via typing on smartphones and carry on a daily conversation.

3.2.1.2. Explanation modes. In the reviewed papers, explanation modes are categorized into Language (20 papers), AR Visualization (three papers), and Human Machine Interfaces (HMI) (four papers). The Language mode involves explanations presented in natural language, delivered through speech or as text captions overlaid on video. Considering both speech (audio) and text (visual) delivery as forms of natural language explanations allows for an emphasis on the framing and knowledge components of explanations. The AR Visualization mode utilizes augmented reality overlays on driving scenarios to highlight environmental components that affect the decision-making process of autonomous vehicles, such as obstacles on the road. Meanwhile, the HMI mode utilizes visual and auditory interfaces, including blinking LED lights or sound cues (e.g., beeps), to convey intent or provide explanations.

3.2.1.3. Explanation types. Predominantly, explanation types have been framed as answers to different styles of questions. Causal explanations (16 papers) address the "why" question and provide reasons for the autonomous vehicle's decision-making process. Contrastive explanations (two papers) explain the "why not" question, clarifying why the vehicle did not continue its previous trajectory and required a new decision. Action explanations (four papers) answer the "what/how" question by describing the actions the autonomous vehicle has taken or will take. Rule-based explanations (two papers) frame explanations around road rules guiding the autonomous vehicle's decisions. Plan-next explanations (two papers) answer the "what will" question by specifying the vehicle's next course of action.

Some studies propose new explanation types based on psychological theories. These types are not framed as answers to specific questions but incorporate elements from multiple simple explanation types. Avetisyan et al. (2022) and Petersen et al. (2019) utilize situational awareness (SA) to design their explanations. Avetisyan et al. (2022) bases their explanations on Endsley's (1995) three levels of information processing: Level 1 SA involves perceiving elements in the environment, Level 2 SA entails comprehending the current situation, and Level 3 SA encompasses projecting the future status to remain updated in a dynamic environment. Petersen et al. (2019) designs two types of explanations: one that promotes high situational awareness and another that promotes low situational awareness. Wiegand et al. (2019, 2020) focus on designing explanations that align with the user's mental model of the autonomous vehicle and driving scenarios. These explanations consider the user's perception and mental model to provide appropriate and tailored information.

3.2.1.4. Explanation timings. The reviewed papers can be divided into three categories based on the timing of the explanations: immediately prior to or during the decision-making process (15 papers), prior to with decision making capability (five papers), and post-event (eight papers). In the first category, explanations are provided either a few seconds before the event or in real-time as the event unfolds. These scenarios are grouped together due to the ambiguity of timing boundaries in some cases. The second category involves explanations given before the event, allowing participants to decide whether to accept the autonomous vehicle's proposed action based on the provided explanation. Throughout the studies, participants are typically given five to 10 s to make their decision. In the post-event category, explanations are given after the event has occurred, with the autonomous vehicle explaining its decision-making process.

3.2.2. Correlated trust metrics

Among the reviewed papers, 14 papers directly examined the effects of explanations on trust while 22 papers examined ensembles of other indirect metrics that are correlated with trust in autonomous vehicles. While the primary focus is on understanding the effects of explanations on trust-building, studying the effects on these correlated metrics allows us to examine trust-building effects indirectly. These metrics serve as surrogates that enable us to indirectly assess the trust levels of participants using potentially more objective measures. Although trust itself is a subjective construct, studying these correlated metrics provides tangible indicators and measurable outcomes that are associated with participants' perceptions and interactions with autonomous vehicles.

The popular choices for correlated metrics are Preference (seven papers), Understanding (six papers) and Anxiety of Participants after experiencing autonomous driving scenarios in the presence and absence of explanations (five papers). Preference measures participants' inclination or favorability towards autonomous driving technology. Understanding assesses how well participants understand the workings and decisions of autonomous vehicles. Anxiety measurements aim to capture participants' anxiety levels in the presence and absence of explanations.

Besides the three most popular choices mentioned above, the other correlated metrics that have been used are situational awareness, cognitive workload, user experience, and manual takeover desire. Situational awareness, studied by Wiegand et al. (2019) and Petersen et al. (2019), measured participants' environmental awareness in the presence of explanations. Cognitive workload, investigated by Du et al. (2019) and Avetisyan et al. (2022), assessed the mental workload caused by the presence of explanation mechanisms. User experience, studied by Schneider et al. (2021, 2023), explored participants' overall experience with autonomous driving events and explanations. Manual takeover desire, studied in Goldman and Bustin (2022), explored the desire of participants to stop autonomous driving and take over manual control.

In addition to these metrics, some studies also utilized objective quantitative measures to assess participants' responses to the driving environment. Unlike questionnaires that rely on self-assessment, these measures provide direct, observable data. Liu et al. (2018) utilized reaction time and steering angles to analyze drivers. Petersen et al. (2019) studied the monitoring ratio using eye trackers in a driving simulator as a surrogate test for trust in autonomous vehicles, under the hypothesis that monitoring the driving situation would hinder performance in secondary tasks. M. Faas et al. (2021) measured crossing onset time, which represents the time taken by a pedestrian to start crossing the road since first seeing the vehicle appear. The vehicle uses external HMIs to provide explanations regarding its intentions.

3.2.3. Data collection

The review reveals that the primary means of data collection is questionnaires, with only a handful of studies collecting physiological data as well. All but one of the reviewed papers use questionnaires to study trust and other correlated metrics in the participants. Omeiza et al. (2022) does not use questionnaires and instead conducts a think-aloud study and then subjectively studies the explanations produced. The following subsection provides more details of the questionnaire types,

their designs, and the timing of questionnaire administration. Four studies collected physiological data during the autonomous driving experience. Petersen et al. (2019) used eye tracker and heart monitor to collect physiological data. Physiological data provided these studies with indicators of attention, anxiety, and other factors, allowing for measures that are less influenced by self-report biases.

3.2.3.1. Questionnaires. We examine the prevalence of questionnaire types and designs, including validated, empirical, and partially validated approaches. Questionnaires found across the reviewed studies are analyzed for their dimensions and applications. The questionnaires employed by the papers included in this review can be classified into three categories.

1. Validated (11 papers),
2. Empirical (seven papers),
3. Partially validated (six papers).

Based on Omeiza et al. (2021), questionnaires are considered validated when they use established research instruments that have been previously utilized for measuring various human factors, including trust. An empirical questionnaire is specifically crafted for the reviewed paper based solely on the author's empirical guidelines. Questionnaires categorized as partially validated combine elements of pre-existing researched questionnaires with additional questions tailored to the specific research context.

Likert scales (Likert, 1932) were commonly employed in these questionnaires with 22 papers using them. Likert scales typically involve a range of ordinal response categories, allowing respondents to indicate their level of agreement or disagreement with specific statements or items pertaining to the research inquiry.

Table 1 lists the most widely used validated questionnaires for trust and its correlated metrics as found in the studies. The choice of the questionnaire used across studies is primarily driven by two factors - how widely accepted the corresponding metric is and how well the dimensions are aligned with the notion of trust specified in the study. For instance, papers focused on improving trust using situational awareness use a questionnaire with that specific dimension.

Table 1
Metrics and relevant validated Questionnaires.

Metric	Questionnaires	Dimensions
Trust	Muir Scale (Muir, 1987)	6 - Competence, predictability, dependability, responsibility, reliability, faith
	Psychometric Trust Scale (Hoffman et al., 2018)	7 - Rule-abiding behavior, predictability, reliability, safety, efficiency, effectiveness, adoptability
	Situational Trust Scale for Automated Driving (STS-AD) (Holthausen et al., 2020)	6 - Trust, performance, engagement, perceived risks, judgment, reaction
	Trust in Automation Scale (Jian et al., 2000)	1 - Trust and Distrust as polar opposites along a single dimension
Preference	Driver Attitude Questionnaire (CHIME lab at Stanford, 2008)	–
	Laan et al. (1997)	–
	Technology Acceptance Model (1989)	–
	Autonomous Vehicle Acceptance Model Questionnaire (AVAM) (Hewitt et al., 2019)	–
Understanding	Situational Awareness Rating Technique (Taylor, 2017)	–
Anxiety	Driver Attitude Questionnaire (CHIME lab at Stanford, 2008)	–
	Naas et al. (2005)	–
Cognitive Workload	NASA-TLX Questionnaire	–

Four questionnaires were predominantly used for direct trust measurement. The Muir scale (Muir, 1987), employed in four papers, is an automation trust scale consisting of six dimensions. Here, competence, predictability, dependability, responsibility, reliability, and faith responses are measured on a seven-point Likert scale. The Psychometric trust scale (Hoffman et al., 2018) was utilized in two papers. The scale assessed users' perceptions of autonomous vehicles' rule-abiding behavior, predictability, reliability, safety, efficiency, effectiveness, and adaptability. It also includes open-ended questions to gather participants' thoughts on trust, safety, reliability, and related aspects. The measurements focused on the system aspect of trust and utilized a Likert scale. The Situational Trust Scale for Automated Driving (STS-AD) (Holthausen et al., 2020), was employed in three papers. The questionnaire included six items designed to measure trust, performance, engagement, perceived risks, judgment, and reaction. The STS-AD scale specifically targets the context of automated driving and includes both general trust and situational trust as measurements. Another trust questionnaire was the Trust in Automation Scale by Jian et al. (2000), which was used in two papers.

For the correlated trust metrics, the choice of questionnaires was different. The questionnaires used for preference included the Driver Attitude Questionnaire developed by the CHIME lab at Stanford in 2008, the questionnaire by Van Der Laan et al. (1997), the Technology Acceptance Model by Davis & others (1989), and the Autonomous Vehicle Acceptance Model Questionnaire (AVAM) by Hewitt et al. (2019). To measure understanding, the Situational Awareness Rating Technique by Taylor (2017) was utilized. Anxiety levels were assessed using the Driver Attitude Questionnaire by the CHIME lab at Stanford from 2008 and the questionnaire by Nass et al. (2005). Cognitive Workload was measured using the NASA-TLX questionnaire.

3.2.3.2. Timing of questionnaire administration. The timing of questionnaire administration was categorized into three phases: before the explanation event (reported in six papers), after the explanation event but before the experiment concluded (reported in 14 papers), and after the entire experiment (reported in 16 papers). The "before explanation event" phase measured pre-levels of trust and other factors. The "after explanation event" phase captured immediate post-explanation trust levels, while the "after experiment" phase assessed trust and factors after exposure to multiple driving scenarios and explanations.

3.3. Impact of explanations

In this section, we address RQ2 and examine whether and to what extent the reviewed studies found that explanations impacted trust. The effects of explanations on correlated metrics and other quantitative measures such as reaction times are also presented.

Examining the effects of explanations on trust, the results varied across the reviewed papers as shown in Table 2. Notably, five studies by Petersen et al. (2019), Zhang et al. (2021), Avetisyan et al. (2022), Lundberg et al. (2022) and Shen et al. (2020) reported a significant increase in trust following the provision of explanations. Another two studies, Ha et al. (2020) and Du et al. (2019) observed a significant increase. However, they found that the effect of explanations on trust was contingent upon the perceived levels of risk and explanation timing, respectively. Six studies including Haspiel et al. (2018), Wintersberger et al. (2021), Omeiza et al. (2021), M. Faas et al. (2021), Du et al. (2021) and Zhang et al. (2023) showed that even though there was some increase in trust following the provision of explanation, the impact was not statistically significant. Körber et al. (2018) reports that explanations had no effect on trust. Among the studies, none found a decrease in trust in autonomous vehicles following explanations. Based on the reviewed studies, it can be concluded that while the impact of explanations on

Table 2
Explanation effects on Trust.

Paper	Questionnaire	Sample Size	Significance Analysis Method	Significance Analysis Result	Effect
Körber et al., 2018	Validated	40	Bayes Factor	BF = 0.32 (Before explanation) and BF = 0.31 (After explanations).	No effect
Haspiel et al. (2018)	Empirical	8	No significance analysis done.	Authors claim non-significance.	Not Significant
Du et al. (2019)	Validated	32	Friedman Test	p (Trust effects) = 0.008. P (Apriori vs Post-Hoc effects) = 0.018. Threshold = 0.05	Increase but timing dependent
Petersen et al. (2019)	Validated	38	Mixed Linear Models	p = 0.02. Threshold = 0.05.	Increase
Ha et al. (2020)	Validated	48	ANOVA	p (Trust effects) = 0.038 p (Perceived Risk effects) = 0.01 Threshold = 0.05	Increase but perceived risk dependent
Wintersberger et al., 2021	Validated	77	Cronbach's Alpha	No significant correlation found between general trust (M = 4.52, SD = 1.06, Cronbach's α = 0.858) or distrust (M = 3.35, SD = 1.22, Cronbach's α = 0.858).	Not Significant
Zhang et al. (2021)	Validated	40	Mixed Linear Models	No p-value reported but p value threshold of 0.05 reported to be satisfied.	Increase
Omeiza et al. (2021)	Validated	101	ANOVA	p = 0.71. Threshold = 0.1	Not Significant
M. Faas et al. (2021)	Validated	60	ANOVA	F (3, 63) = 1.28, p = 0.289 Threshold = 0.05	Not Significant
Du et al. (2022)	Validated	118	Mixed Linear Models	F (2, 289) = 2.548, p = 0.08 Threshold = 0.05	Not Significant
Lundberg et al. (2022)	Empirical	30	ANOVA	No p-value reported but p value threshold of 0.05 reported to be satisfied.	Increase
Avetisyan et al. (2022)	Validated	340	ANOVA	p = 0.040 Threshold = 0.05	Increase
Zhang et al. (2023)	Validated	30	Mixed Linear Models	p = 0.118 Threshold = 0.1	Not Significant
Shen et al., 2020	Empirical	18	Friedman Test	No p-value reported but p value threshold of 0.05 reported to be satisfied.	Increase

BF = Bayes Factor; M = Mean; SD = Standard Deviation.

trust overall follows a positive trend, the impact may not be of statistical significance in all cases and hence must be interpreted with caution.

We further examine the questionnaires used and significance testing performed in the studies as summarized in Table 2. It is important to note that Lundberg et al. (2022), Shen et al. (2020), and Haspiel et al. (2018) report effects on trust using empirical questionnaires, while the other 11 studies use validated questionnaires. Thirteen out of the 14 studies performed some analysis to test the significance of the effect of explanations on trust. Sample size, an essential aspect of significance testing, is reported for each study in Table 2. For studies that included significance testing, sample size ranged from 18 to 340, with a mean of 75 and a median of 40. Five studies conducted Analysis of Variance (ANOVA), four studies employed Mixed Linear Models, and two studies applied the Friedman Test - all reporting significance based on p-values compared to a predefined threshold. One study uses the Bayes Factor hypothesis testing, which serves as a Bayesian alternative to traditional significance testing, and reports significance based on Bayes Factor instead of p-values. Multiple studies have relied on Cronbach's Alpha test to prove measurement reliability but one also uses the test result to infer significance. Further details of the experiments such as effect size and power estimation were missing from most of the reviewed studies and hence not reported here. More discussions on the design of experiments and how they can be further improved are presented in Section 3.5.

We further analyze the impact of explanations on trust through correlated metrics. As shown in Table 3, six out of seven studies reported an increase in participants' preference for autonomous driving technology after exposure to explanations, with four out of these six having reported a significant increase. Only one of the seven studies utilized empirical questionnaires. For significance analysis, three studies conducted ANOVA, one used the Friedman Test, one employed Mixed Linear Models, and one used Bayes Factor. Understanding of autonomous vehicles also showed a positive trend, with all six studies indicating a significant increase in participants' comprehension following explanations as shown in Table 4. Five out of these six studies reported impact on understanding using empirical questionnaires. The significance analysis included ANOVA, Mixed Linear Models, Pearson's chi-square test, and Bayes Factor, with all but one study performing

Table 3
Explanation effects on correlated metric - Preference.

Correlated Metric - Preference					
Paper	Questionnaire	Sample Size	Significance Analysis Method	Significance Analysis Result	Effect
Koo et al. (2015)	Validated	64	ANOVA	F (1, 60) = 4.79, p < 0.05	Increase
Körber et al., 2018	Validated	40	Bayes Factor	BF = 0.36 (Before explanation) and BF = 0.42 (After explanations).	No effect
Haspiel et al. (2018)	Empirical	8	Not available.	Authors claim non-significance.	Increase, not significant
Du et al. (2019)	Validated	32	Friedman Test	p = 0.027 Threshold = 0.05	Increase
Du et al. (2022)	Validated	118	ANOVA	p = 0.019 Threshold = 0.05	Increase
Zhang et al. (2023)	Validated	30	Mixed Linear Models	p = 0.205 Threshold = 0.05	Increase, not significant
Schneider et al. (2023)	Validated	121	ANOVA	No p-value reported but p value threshold of 0.05 reported to be satisfied.	Increase

Table 4
Explanation effects on correlated metric - Understanding.

Correlated Metric - Understanding					
Paper	Questionnaire	Sample Size	Significance Analysis Method	Significance Analysis Result	Effect
Körber et al. (2018)	Empirical	40	Bayes Factor	BF(Measured only after explanation) = 14.71	Increase
Liu et al. (2018)	Empirical	30	Pearson's chi-square test	No p-value reported but p value threshold of 0.05 reported to be satisfied.	Increase
Petersen et al. (2019)	Validated	38	Mixed Linear Models	p = 0.03 Threshold = 0.05	Increase
Omeiza et al. (2021)	Empirical	101	ANOVA	No p-value reported but p value threshold of 0.001 reported to be satisfied.	Increase
Goldman and Bustin, 2022	Empirical	2586	Not available.	Not available.	Increase
Lundberg et al. (2022)	Empirical	30	ANOVA	No p-value reported but p value threshold of 0.01 reported to be satisfied.	Increase

Table 5
Explanation effects on correlated metric - Anxiety.

Correlated Metric - Anxiety					
Paper	Questionnaire	Sample Size	Significance Analysis Method	Significance Analysis Result	Effect
Haspiel et al. (2018)	Empirical	8	No significance analysis done.	Authors claim non-significance.	Decrease, not significant
Du et al. (2019)	Validated	32	Friedman Test	p = 0.666 Threshold = 0.05	Decrease, not significant
Zhang et al. (2021)	Validated	40	Mixed Linear Models	No p-value reported but p value threshold of 0.05 reported to be satisfied.	Decrease
Du et al. (2022)	Validated	118	Mixed Linear Models	p = 0.786 Threshold = 0.05	Increase, not significant
Zhang et al. (2023)	Validated	30	Mixed Linear Models	p = 0.422 Threshold = 0.05	Decrease, not significant

significance testing. In terms of anxiety, four out of five studies observed a significant decrease in anxiety levels with explanations, while Du et al. (2021) found an insignificant increasing effect, as shown in Table 5. Only one out of these five studies utilized empirical questionnaires. For significance analysis, one study used the Friedman Test, and three employed Mixed Linear Models.

Among the studies that used other metrics, Wiegand et al. (2019) and Petersen et al. (2019) reported a significant increase in Situational Awareness. Du et al. (2019) found no significant effect of explanations on cognitive workload, while Avetisyan et al. (2022) reported a significant increase. Schneider et al. (2021) found that user experience increased from negative to neutral when explanations were provided, while Schneider et al. (2023) reported a positive correlation between explanations and a positive user experience. Perceived safety and perceived feeling of control examined by Schneider et al. (2023), indicated an increase in perceived feeling of control when explanations were provided. However, the effect on perceived safety was not significant.

Liu et al. (2018) reports that participants who received explanations about the limitations of autonomous vehicle systems exhibited better reaction times and safer steering angles when the system failed to act. Petersen et al. (2019) found that the monitoring ratio decreased with explanations and improved secondary task performance. A significant negative correlation between trust in the vehicle providing explanations and crossing onset time was noted by M. Faas et al. (2021), indicating that explanations influenced pedestrians' trust and their decision to start crossing earlier.

3.4. Characteristics and theories for effective explanations

In this section, we address RQ3 and examine the characteristics, such as explanation types and timings, that positively contribute to trust and its correlated metrics. Within the reviewed studies, various explanation types and timings have been proposed and tested. However, because the experimental instruments (including the level of immersion and the complexity of the assigned tasks), data collection methods, and timing of data collection varied widely across the studies, investigating explanation types and timings across different studies was not deemed meaningful. Instead, to compare the effects of different explanation types, we consider studies that concurrently employ two or more explanation types. We use the same approach for explanation timing.

Regarding explanation type, three studies meet this criterion. In Koo et al. (2015), Omeiza et al. (2021), and Du et al. (2021), both Action and Causal explanations are considered. Koo et al. (2015) and Omeiza et al. (2021) found causal explanations to have a greater impact on trust-related metrics, while Du et al. (2021) suggested that a combination of Action and Causal explanations yields the best results. Omeiza et al. (2021) investigated both contrastive and causal explanation types, with contrastive explanations leading to better trust metrics. Although no singular explanation type unequivocally prevails across all studies, particular types, such as Causal and Contrastive explanations, exhibit potential for enhancing trust. Combining different explanation types may also offer superior results.

Furthermore, additional studies in the review demonstrate how different explanation constructs are preferred by various age groups and demographics. This consideration is particularly relevant for autonomous vessels since seafarers are predominantly from certain age groups and demographics. Zhang et al. (2021) found older drivers to prefer explanation types where permission was sought for undertaking the explained action significantly more than younger demographics. Goldman and Bustin (2022) observed that Infrequent Drivers (58%) and those Under Age 40 (56%) indicated more comfort with an explanation type displaying a top view of the vehicle's surroundings, as opposed to no explanation or a general textual explanation. Shen et al. (2020) indicate that there is no universally preferred explanation type, as individuals with different levels of trust in autonomous vehicle technology exhibit preferences for different types of explanations.

Looking into studies that use explanation types based on psychological theories, Avetisyan et al. (2022) find that Level 2 SA explanations significantly outperform Level 1 and Level 3 explanations in situational trust. Petersen et al. (2019) confirm the hypothesis that explanation types with better situational awareness increase drivers' trust and lead to improved secondary task performance. Wiegand et al. (2019) identify

a target mental model that enhances the user's understanding by incorporating key components from expert mental models. The results indicate that displaying detected objects and their predicted motion is crucial for comprehending a situation.

To compare the effects of different explanation timings, we consider only studies that concurrently employ two or more timing strategies. Five studies fulfill this criterion, allowing for a comparison of the most effective timing based on the trust metrics utilized in these studies. The compiled results, presented in Table 6, indicate that both "immediately prior to or during" and "prior to with decision-making capability" timings are more effective in trust-building compared to post-event timings. However, it is unclear which of the two advance timing types is superior. Notably, Haspiel et al. (2018) found that providing participants with decision-making capability increased their cognitive workload.

3.5. Limitations in design of experiments

A limitation of the studies reviewed is the difficulty in accurately replicating real-life driving risks even with advanced simulators. While simulators offer controlled environments for studying trust in autonomous vehicles, they may not fully capture the complexity and unpredictability of real-world driving scenarios. As a result, trust measurements obtained in simulated environments, where there is no actual risk, may differ from trust experiences in real-world driving situations. Thus, findings from simulator-based studies should be interpreted with caution regarding their applicability to real-world driving contexts. Another limitation is the potential bias introduced by participant recruitment. Six studies recruited participants from technical institutes, leading to high levels of pre-existing trust in autonomous vehicle technology even before experiencing driving events accompanied by explanations. This pre-existing trust could potentially be a contributing factor to the observed increase in trust metrics. Individuals with technical knowledge or familiarity with autonomous systems may inherently possess higher levels of trust due to their understanding of the underlying technology. Consequently, the generalizability of findings to the broader population, including individuals with different educational backgrounds or technological literacy levels, may be limited. To address these limitations, future studies should strive for diverse participant recruitment strategies to capture a broader range of perspectives and attitudes towards autonomous vehicles and vessels.

4. Key takeaways for autonomous vessels and conclusion

AI-powered automation is rapidly advancing in both land vehicles and the maritime industry. This technological disruption brings new challenges in adoption, due to a lack of transparency and understanding of its functioning. Existing literature on explainability in autonomous vessels and autonomous vehicles reveals that while both domains emphasize navigational control and safety, research on autonomous vessels lags that on land vehicles. Although the paradigm of trust in autonomous vessels must be approached differently given the distinct operational context, understanding and addressing the impact of explainability on trustworthiness of AI-powered automation is essential for both domains. There is room for learning from autonomous vehicles to foster the development of trustworthy autonomous vessels.

Certain findings discussed in this paper offer valuable insights for guiding the development of trustworthy autonomous vessels. Beyond trust, two major issues—awareness and understanding, and control—have been highlighted in the adoption of autonomous vessels (Mallam et al., 2020). Subject matter experts (SMEs) interviewed in the study emphasized the critical need for understanding and being aware of an autonomous vessel's capabilities, strengths, weaknesses, and most importantly, its decision-making process (Mallam et al., 2020). The findings from our literature review show that while the impact of explanation on trust was not always found to be significant for autonomous vehicles, the impact on the correlated metrics "understanding" was unanimously positively significant. This is a key finding since in the maritime domain, the concept of "trust" in an autonomous system is perceived to be closely associated with the awareness and understanding of a system and its decision making (Mallam et al., 2020).

Control of autonomous vessels, identified as another major issue from the perspective of SMEs, can also benefit from the findings of our literature review. The discovery that providing explanations about the limitations of autonomous vehicles led to better reaction times in abnormal situations suggests potential utility in enhancing the sense of control for seafarers in autonomous vessels. Considering the characteristics and theoretical underpinnings of effective explanations, the approach to provide explanations that seek permissions before undertaking actions is anticipated to enhance trust. This would involve explaining the action, then asking seafarers for their approval. Additionally, timing such explanations before significant events can enhance their effectiveness. By allowing seafarers the option to accept or reject the autonomous system's decisions, it is possible to design a system that gives them a stronger sense of control in autonomous vessels.

Similar to autonomous vehicles, preferences regarding the level of detail in explanations and the mode of explanation (text versus visual display of vessel surroundings) are anticipated to vary between naïve and experienced seafarers. These considerations underscore the importance of tailoring the design of autonomous vessel systems to accommodate the diverse needs and preferences of seafarers with varying levels of experience.

The identified shortcomings in the experimental designs to measure the impact of explanation on trust in autonomous vehicles also serve as valuable lessons for autonomous vessels. Designing high-fidelity scenarios which encompass abnormal out-of-the-box situations within simulators and ensuring the recruitment of suitable participants are equally crucial in the context of autonomous vessels. By addressing these aspects, researchers and developers can better simulate real-world conditions and gather meaningful data to guide the design and implementation of trustworthy autonomous vessel systems.

However, when comparing autonomous vehicles to autonomous ships, the paradigm of trust presents some unique aspects. In the realm of autonomous vessels, trust operates within a team-based navigation and maneuvering environment, unlike the single operator scenarios often seen in autonomous vehicles. This necessitates tailored approaches to studying and fostering trust within multi-person crews, where group dynamics and collective decision-making play a pivotal role. Another major contrast is that while autonomous vehicles cater to a broad spectrum of users, the audience for autonomous vessels is considerably narrower, predominantly consisting of seafarers with specialized training and expertise in maritime operations. Unlike the diverse age groups, demographics, and technological literacy observed among users of autonomous vehicles, seafarers utilizing autonomous vessels are

Table 6
Preferred explanation timing.

Explanation Timing with respect to decision-making				
Paper	Immediately prior to or during (T1)	Prior to with decision making capability (T2)	Post-event (T3)	Timing that induce most trust
Haspiel et al. (2018)	Yes	Yes	Yes	T2
Du et al. (2019)	Yes	Yes	Yes	T1
Zhang et al. (2021)	Yes	Yes	Yes	T1
Du et al. (2022)	Yes	Yes	Yes	T1, T2
Zhang et al. (2023)	Yes	Yes		T2

expected to exhibit a more homogeneous profile, with generally lower variability in their technological proficiency and a narrower demographic range. This can influence the perception of both explainability and trust, as well as their relation.

Moving forward, future research should investigate more factors beyond explainability that may influence trust in autonomous vessels. Investigating these factors and their potential to enhance trust, alongside systematic studies of explanation types, timing, and modes, will provide clearer insights. Additionally, while leveraging knowledge from autonomous vehicles, researchers must account for the fundamental differences in operational practices, environments, and the implications of system failures to ensure the relevance and applicability of findings to maritime systems.

CRediT authorship contribution statement

Rohit Ranjan: Writing – original draft, Visualization, Investigation, Formal analysis, Data curation. **Ketki Kulkarni:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization. **Mashrura Musharraf:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

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.

Acknowledgements

This work was supported by the Research Council of Finland project: Towards human-centered intelligent ships for winter navigation (Decision number: 351491) and the Aalto Science Institute International Summer Research Programme.

References

Avetisyan, L., Ayoub, J., Zhou, F., 2022. Investigating explanations in conditional and highly automated driving: the effects of situation awareness and modality. Transport. Res. F Traffic Psychol. Behav. 89, 456–466.

Brandsæter, A., Smefjell, G., van de Merwe, K., Kamsvåg, V., 2020. Assuring safe implementation of decision support functionality based on data-driven methods for ship navigation. E-proceedings of the 30th European Safety and Reliability Conference and 15th Probabilistic Safety Assessment and Management Conference (ESREL2020 PSAM15).

Chaal, M., Ren, X., BahooToroody, A., Basnet, S., Bolbot, V., Banda, O.A.V., Van Gelder, P., 2023. Research on risk, safety, and reliability of autonomous ships: a bibliometric review. Saf. Sci. 167, 106256.

Chowdhury, R., Navsalkar, A., Subramani, D., 2022. GPU-accelerated multi-objective optimal planning in stochastic dynamic environments. J. Mar. Sci. Eng. 10 (4), 533.

Davis, F.D., 1989. Technology Acceptance Model: TAM. *Al-Suqri, MN, Al-Aufi, AS: Information Seeking Behavior And Technology Adoption*. others, pp. 205–219.

Du, N., Haspiel, J., Zhang, Q., Tilbury, D., Pradhan, A.K., Yang, X.J., Robert Jr, L.P., 2019. Look who's talking now: implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. Transport. Res. C Emerg. Technol. 104, 428–442.

Du, N., Robert, L.P., Yang, X.J., 2022. Cross-cultural investigation of the effects of explanations on drivers' trust, preference, and anxiety in highly automated vehicles. Transp. Res. Rec. 2677, 554–561.

Du, N., Zhou, F., Tilbury, D., Robert, L.P., Yang, X.J., 2021. Designing alert systems in takeover transitions: the effects of display information and modality. 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, pp. 173–180.

Endsley, M.R., 1995. A taxonomy of situation awareness errors. Human Factors in Aviation Operations 3 (2), 287–292.

Foundation, L.R., 2021. World risk poll 2021 – report 3: a digital world: perceptions of risk from AI and misuse of personal data. https://wrp.lrfoundation.org.uk/LRF_2021_report_a-digtial-world-ai-and-personal-data_online_version.pdf.

Gjærum, V.B., Rørvik, E.-L.H., Lekkas, A.M., 2021a. Approximating a deep reinforcement learning docking agent using linear model trees. 2021 European Control Conference (ECC), pp. 1465–1471.

Gjærum, V.B., Strümke, I., Alsos, O.A., Lekkas, A.M., 2021b. Explaining a deep reinforcement learning docking agent using linear model trees with user adapted visualization. J. Mar. Sci. Eng. 9 (11), 1178.

Glomsrud, J.A., Ødegårdstuen, A., Clair, A.L.S., Smogeli, Ø., 2019. Trustworthy versus explainable AI in autonomous vessels. Proceedings of the International Seminar on Safety and Security of Autonomous Vessels (ISSAV) and European STAMP Workshop and Conference (ESWC), p. 37.

Goldman, C.V., Bustin, R., 2022. Trusting explainable autonomous driving: simulated studies. IEEE Intelligent Vehicles Symposium (IV), 1255–1260, 2022.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D., 2018. A survey of methods for explaining black box models. ACM Comput. Surv. 51 (5), 1–42.

Ha, T., Kim, S., Seo, D., Lee, S., 2020. Effects of explanation types and perceived risk on trust in autonomous vehicles. Transport. Res. F Traffic Psychol. Behav. 73, 271–280.

Haspiel, J., Du, N., Meyerson, J., Robert Jr, L.P., Tilbury, D., Yang, X.J., Pradhan, A.K., 2018. Explanations and expectations: trust building in automated vehicles. Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction 119–120.

Hewitt, C., Politis, I., Amanatidis, T., Sarkar, A., 2019. Assessing public perception of self-driving cars: the autonomous vehicle acceptance model. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 518–527.

Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J., 2018. Metrics for explainable AI: challenges and prospects. arXiv Preprint arXiv:1812.04608.

Holthausen, B.E., Wintersberger, P., Walker, B.N., Riener, A., 2020. Situational trust scale for automated driving (STS-AD): development and initial validation. 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, pp. 40–47.

Houweling, K.P., Mallam, S.C., van de Merwe, K., Nordby, K., 2024. The effects of augmented reality on operator situation awareness and head-down time. Appl. Ergon. 116, 104213.

Jian, J.-Y., Bisantz, A.M., Drury, C.G., 2000. Foundations for an empirically determined scale of trust in automated systems. Int. J. Cognit. Ergon. 4 (1), 53–71.

Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., Nass, C., 2015. Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. Int. J. Interact. Des. Manuf. 9, 269–275.

Körber, M., Prasch, L., Bengler, K., 2018. Why do I have to drive now? Post hoc explanations of takeover requests. Hum. Factors 60 (3), 305–323.

Likert, R., 1932. A technique for the measurement of attitudes. Arch. Psychol. 22 (140), 55–55.

Liu, T., Zhou, H., Itoh, M., Kitazaki, S., 2018. The impact of explanation on possibility of hazard detection failure on driver intervention under partial driving automation. IEEE Intelligent Vehicles Symposium (IV), 150–155, 2018.

Łosiewicz, Z., Mironiuk, W., 2020. Critical areas of the autonomous seagoing vessel concept model-according to selected criteria. In: Research Methods and Solutions to Current Transport Problems: Proceedings of the International Scientific Conference Transport of the 21st Century, 9–12th of June 2019, Ryn, Poland 15. Springer International Publishing, pp. 274–283.

Lundberg, H., Mowla, N.I., Abedin, S.F., Thar, K., Mahmood, A., Gidlund, M., Raza, S., 2022. Experimental analysis of trustworthy in-vehicle intrusion detection system using eXplainable artificial intelligence (XAI). IEEE Access 10, 102831–102841.

Faas, S.M., Kraus, J., Schoenhals, A., Baumann, M., 2021. Calibrating pedestrians' trust in automated vehicles: does an intent display in an external HMI support trust calibration and safe crossing behavior? Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems 1–17.

Mallam, S.C., Nazir, S., Sharma, A., 2020. The human element in future Maritime Operations–perceived impact of autonomous shipping. Ergonomics 63 (3), 334–345.

Muir, B.M., 1987. Trust between humans and machines, and the design of decision aids. Int. J. Man Mach. Stud. 27 (5–6), 527–539.

Nass, C., Jonsson, I.-M., Harris, H., Reaves, B., Endo, J., Brave, S., Takayama, L., 2005. Improving automotive safety by pairing driver emotion and car voice emotion. CHI'05 Extended Abstracts on Human Factors in Computing Systems, pp. 1973–1976.

Omeiza, D., Anjomshoae, S., Webb, H., Jirotka, M., Kunze, L., 2022. From spoken thoughts to automated driving commentary: predicting and explaining intelligent vehicles' actions. IEEE Intelligent Vehicles Symposium (IV), 1040–1047, 2022.

Omeiza, D., Kollnig, K., Web, H., Jirotka, M., Kunze, L., 2021a. Why not explain? Effects of explanations on human perceptions of autonomous driving. 2021 IEEE International Conference on Advanced Robotics and its Social Impacts (ARSO), pp. 194–199.

Omeiza, D., Web, H., Jirotka, M., Kunze, L., 2021b. Towards accountability: providing intelligible explanations in autonomous driving. IEEE Intelligent Vehicles Symposium (IV), 231–237, 2021.

Omeiza, D., Webb, H., Jirotka, M., Kunze, L., 2021c. Explanations in autonomous driving: a survey. IEEE Trans. Intell. Transport. Syst. 23 (8), 10142–10162.

Petersen, L., Robert, L., Yang, X.J., Tilbury, D.M., 2019. Situational awareness, drivers trust in automated driving systems and secondary task performance. arXiv Preprint arXiv:1903.05251.

SAE International, 2018. *Taxonomy and Definitions for terms Related to driving automation Systems for on-road motor vehicles* (J3016_201806). SAE International.

Salay, R., Queiroz, R., Czarnecki, K., 2017. An analysis of ISO 26262: using machine learning safely in automotive software. arXiv Preprint arXiv:1709.02435.

Schneider, T., Hois, J., Rosenstein, A., Ghellal, S., Theofanou-Fülbier, D., Gerlicher, A.R., 2021. Explain yourself! Transparency for positive ux in autonomous driving. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–12.

Schneider, T., Hois, J., Rosenstein, A., Metzl, S., Gerlicher, A.R., Ghellal, S., Love, S., 2023. Don't fail me! The level 5 autonomous driving information dilemma regarding transparency and user experience. Proceedings of the 28th International Conference on Intelligent User Interfaces, pp. 540–552.

Shafi, K., 2017. A machine competence based analytical model to study trust calibration in supervised autonomous systems. In: 2017 Ninth International Conference on Advanced Computational Intelligence (ICACI). IEEE, pp. 245–252.
Shen, Y., Jiang, S., Chen, Y., Campbell, K.D., 2020. To explain or not to explain: a study on the necessity of explanations for autonomous vehicles. arXiv Preprint arXiv: 2006.11684.
Singh, V., Osen, O.L., Bye, R.T., 2023. Explainable Artificial Intelligence for Autonomous Surface Vessels by Fuzzy-Based Collision Avoidance System. International Conference on Smart Computing and Communication, pp. 145–163.
Taylor, R.M., 2017. Situational awareness rating technique (SART): the development of a tool for aircrew systems design. In: In Situational Awareness. Routledge, pp. 111–128.
van de Merwe, K., Mallam, S., Nazir, S., 2024. Agent transparency, situation awareness, mental workload, and operator performance: a systematic literature review. Hum. Factors 66 (1), 180–208.
Van Der Laan, J.D., Heino, A., De Waard, D., 1997. A simple procedure for the assessment of acceptance of advanced transport telematics. Transport. Res. C Emerg. Technol. 5 (1), 1–10.
Veitch, E., Alsos, O.A., 2021. Human-centered explainable artificial intelligence for marine autonomous surface vehicles. J. Mar. Sci. Eng. 9 (11), 1227.
Wiegand, G., Eiband, M., Haubelt, M., Hussmann, H., 2020. "I'd like an explanation for that!" Exploring reactions to unexpected autonomous driving. 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 1–11.
Wiegand, G., Schmidmaier, M., Weber, T., Liu, Y., Hussmann, H., 2019. I drive-you trust: explaining driving behavior of autonomous cars. Extended Abstracts of the 2019 Chi Conference on Human Factors in Computing Systems 1–6.
Wintersberger, P., Janotta, F., Peintner, J., Löcken, A., Riener, A., 2021. Evaluating feedback requirements for trust calibration in automated vehicles. IT Inf. Technol. 63 (2), 111–122.
Zhang, Q., Yang, X.J., Robert Jr, L.P., 2021. Drivers' age and automated vehicle explanations. Sustainability 13 (4), 1948.
Zhang, Y., Wang, W., Zhou, X., Wang, Q., Sun, X., 2023. Tactical-level explanation is not enough: effect of explaining AV's lane-changing decisions on drivers' decision-making, trust, and emotional experience. Int. J. Hum. Comput. Interact. 39 (7), 1438–1454.