Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023)
Edited by Mário P. Brito, Terje Aven, Piero Baraldi, Marko Čepin and Enrico Zio
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-18-8071-1_P320-cd

Considerable Risk Sources and Evaluation Factors for Artificial Intelligence in Maritime Autonomous Systems

Changui Lee

Korea Conformity Laboratories, Changwon, Republic of Korea. E-mail: phdculee@gmail.com

Seojeong Lee

Korea Maritime and Ocean University, Busan, Republic of Korea. E-mail: sjlee@kmou.ac.kr

Alongside of the MASS implementation, Artificial Intelligence (AI) is becoming a prominent issue. As a part of that, it is necessary to prepare possible risk sources and reasonable evaluation mechanism. There already exist international standards regarding AI, machine learning and risk assessment to be able to be considered and interpreted to fit to the maritime sector.
This article is aiming to find out risk sources for AI in maritime sector, based on the document on AI concepts and terminology (ISO/IEC 22989) such as level of automation, lack of transparency and explainability, complexity of environment, system life cycle issues, system hardware issues, and technology readiness. Also, it is to propose evaluation factors applied practically for those risk sources, which are robustness, reliability, resilience, controllability explainability, predictability, transparency, fairness, jurisdictional issues, precision, recall, accuracy and F1 score coming from risk management and AI and Machine Learning (ML) related international standards. The article also reviews two MASS related guidelines from Det Norske Veritas (DNV) and American Bureau of Shipping (ABS) to know current status how and what risk sources are contemplated.
As such, the combination of risk sources and evaluation factors proposed can be applied to evaluate AI practically after adjusting to fit to applied context specifically. Also, all kind of MASS risk stakeholders can be potential users taking into account on these factors and methods, such as risk and test managers, equipment makers, ship owners, classification societies.

Keywords: Artificial intelligence, Risk management, Risk source, Evaluation factor, MASS.

1. Introduction

As the remarkable advancements of machine learning and deep learning have occurred, artificial intelligence (AI) has brought about significant changes in various industries. AI is being used for medical diagnoses, speech recognition, and recommendation systems, among other applications. The maritime sector is also following this trend by introducing systems with electricity, electronics, and software to increase operational efficiency, moving away from traditional mechanical equipment. In the shipbuilding and shipping industry, discussions on the development and operation of Maritime Autonomous Surface Ships (MASS) have been gaining momentum, reflecting the influence of the fourth industrial revolution. At the 98th meeting of the Maritime Safety Committee (MSC) of the International Maritime Organization (IMO) held in June 2017, MASS was defined as a ship that can operate independently of human interaction to a varying degree. At the 99th meeting, an official international agreement was reached on this definition. Alsos et al. (2022).

There have been experienced with various kinds of AI related applications regarding MASS including autonomous navigation, object detection, predictive maintenance, weather routing, cargo optimisation, and autonomous inspection. For example, AI-enabled ship navigation systems can be defined to detect obstacles and make accurate decisions to navigate ships safely in independent and autonomous way, where the systems gather appropriate datum from different kind of sensors, cameras, and so on. These systems are also assumed to be able to identify

any surrounding obstacles like other ships, buoys, and to predict maintenance needs, to plan the most efficient route based on real-time feeding weather conditions, and to optimise cargo loading and unloading. Öztürk et al. (2022)

Despite the potential benefits of AI in the maritime sector, its application is not always positive. This is because AI has unique features and development processes that differ from features and processes for traditional software. Faults and failures of AI could at least influence to or directly cause physical damage, economic damage, and physical injury. In addition, the inherent malfunctions and logical errors further compound the potential dangers of adopting AI-enabled systems. As such, as introducing AI in MASS could pose safety threats, this article emphasises the needs of researches and standardisation efforts to ensure the safety and reliability of AI in the maritime sector. This should be concerned crucially to mitigate potential risks associated with AI in MASS to promote safe and efficient operations. Miyoshi et al. (2022)

This article aims to identify and evaluate the possible risk sources associated with the use of AI in the automation systems of the maritime sector, and to provide practical evaluation factors. To do that, the paper refers existing useful international standards, such as ISO/IEC 22989, ISO/IEC 23894, and ISO/IEC 23053. The reviews on two MASS-related guidelines from DNV and ABS helps to understand the current status on this topic.

2. Standard documents related to AI

ISO and JTC 1 collaborate to set standards for information technology since 1987. In 2017, JTC1 established the JTC1 subcommittee (SC) 42 for AI standardization, consisting of five working groups and one joint working group. ISO/IEC JTC1/SC 42 covers various topics related to AI, such as defining AI, big data, and trustworthiness, and publishes standard documents on these topics. Issa et al. (2022)

2.1. *ISO/IEC 22989 (Artificial Intelligence Concept and Terminology)*

ISO/IEC 22989 is a standard that provides a common language for understanding AI in the IT industry. It covers concepts of categorization, from strong and weak AI to modern machine learning methods. It also defines common concepts and definitions, such as AI-enabled systems and learning data. ISO/IEC 22989 presents nine characteristics that enable trustworthiness validation and defines it as the verification of whether the system satisfies stakeholder expectations. Table 1 shows these characteristics.

Table 1. Characteristics for trustworthiness.

Characteristic	Description
Robustness	The ability to maintain performance level even under external interference or harsh environmental conditions
Reliability	The ability to perform the necessary functions without breakdown during the intended
Resilience	The ability to quickly recover the operation status after the accident
Controllability	The ability to take over control rights by an external agent involved
Explainability	The ability to explain why the AI-enabled system made this decision
Predictability	The ability to enable a person who can trust in the results of the AI-enabled system
Transparency	The degree of disclosing what data is needed and how it collected and educated the data
Fairness	The degree of discrimination against different groups
Jurisdictional issues	The difference in regulations applied when the operating area of the AI-enabled system (jurisdiction) is changed

Source: ISO/IEC 22989:2022

2.2. *ISO/IEC 23053 (Artificial Intelligence System Framework Using Machine Learning)*

ISO/IEC 23053 is a framework for machine learning technology, a crucial aspect of modern AI. It defines machine learning pipelines, offering examples of machine learning development processes and introducing concepts such as tasks, algorithms, and performance

metrics. The document also outlines commonly used classification systems, including supervised, unsupervised, and reinforcement learning, along with their typical methods and approaches. Table 2 presents the performance metrics for evaluating machine learning, using the confusion matrix with TP, FN, FP, and TN values. This matrix visualizes the performance of a supervised classification algorithm, where "True" means predicted and actual values match and "False" means a discrepancy. A positive predicted value is P (Positive), and a negative predicted value is N (Negative).

Table 2. Performance metrics for validating machine learning.

Metric	Description
Precision	The ratio of positive on prediction value by positive $Precision = \frac{TP}{TP + FP}$ (1)
Recall	The ratio of correct positive on the actual positive value $Recall = \frac{TP}{TP + FN}$ (2)
Accuracy	The ratio of correctly predicted on the total prediction value $Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$ (3)
F1 Score	Harmonic mean with precision and recall $F1\ Score = 2 \times \frac{(Precision) \times (Recall)}{(Precision) + (Recall)}$ (4)

Source: ISO/IEC 23053:2022

TP: True Positive (correctly predicted positive instances)

TN: True Negative (correctly predicted negative instances)

FP: False Positive (incorrectly predicted as positive when actually negative)

FN: False Negative (incorrectly predicted as negative when actually positive)

2.3. *ISO/IEC 23894 (Artificial Intelligence Risk Management)*

ISO/IEC 23894 is a standard for managing potential risks that may arise when developing and introducing AI or intelligent services. The standard provides general information and frameworks on risk management for AI, as well as risk management procedures, purposes, causes of risk, and risk management methods according to the AI lifecycle. ISO/IEC 23894 categorizes potential causes of risk into seven types, as shown in Table 3.

Table 3. Risk sources of AI.

Risk source	Description
Level of automation	AI is often used in automation of the system and can affect the automation stage. In particular, if you need collaboration with people, handovers with people can be a risk.
Lack of transparency and explainability	If the information related to the development and learning of the AI-enabled system is not transparently disclosed, or if it is not explained so that people can understand the basis for the judgment of AI, the AI will not be trusted.
Complexity of environment	AI is mainly used to handle complex and diverse surrounding environments, and complex environments can cause additional risks compared to simplicity.
System life cycle issues	AI has a different characteristic from the existing system development life cycle, which can cause danger. For example, it can be an inappropriate verification method or process.
System hardware issues	The defect in the hardware or sensor can be interrupted or incorrectly measured by the service. In addition, the lack of system performance or communication bandwidth for AI can cause risk.
Technology readiness	There may be a risk if you still use less technically less mature AI algorithms or models for real work.
Risk sources related to machine learning	The development of AI is associated with machine learning or deep learning. Risks may occur if the quality or learning process of data required for learning is inappropriate.

Source: ISO/IEC 23894:2023

The AI is often utilised for system automation, and depending on the level of automation, collaboration with humans may be necessary. Therefore, collaboration between humans and systems can become a risk. If humans cannot confirm the basis of AI learning or decision-making processes during collaboration, it may lead to a lack of trust in the AI and potential risks. Moreover, the performance of AI can be affected by the complex and diverse surrounding environment, leading to potential risks. Handling various situations may require many computations, necessitating high-performance hardware and various sensors. However, performance deficiencies in systems or sensors can also result in risks. The development of AI differs from that of traditional systems. Applying traditional development methods as is may make it impossible to sufficiently verify the AI software, which can result in potential risks. Additionally, using immature or unverified AI algorithms prematurely can lead to risks. Furthermore, the performance of AI can vary greatly depending on the learning process, and potential risks can arise if the quality of data required for learning is poor or if the learning process is inadequate. Razmjooei et al. (2023)

3. Proposed evaluation factors for risk sources

AI-enabled systems are increasingly utilised across various industries, offering numerous benefits and opportunities. However, they also pose potential risks and challenges that must be managed. To effectively manage these risks, the ISO/IEC 23894 provides a standardised framework that covers the various risk sources that may arise when developing and operating AI.

A method is necessary to evaluate the degree of risk stemming from these risk sources. This article defines characteristics for trustworthiness and performance metrics as evaluation factors and endeavours to categorize the evaluation factors that may arise from each specific risk source. Consequently, this article can propose an approach that suggests combination of evaluating each source, by deriving evaluation factors associated with the respective risk sources.

Table 4 presents a combination of risk sources and evaluation factors for AI, including the seven types of risk sources presented in ISO/IEC 23894, the nine types of characteristics for trustworthiness presented in ISO/IEC 22989, and the performance metrics in ISO/IEC 23053. Each risk source can correspond to one or more evaluation factors, and each evaluation factor can correspond to one or more sources of risk. This means that the relationship between risk sources and evaluation factors is a many-to-many (N:N) relationship, highlighting the complexity of managing risks associated with AI.

Table 4. The combination of risk sources and evaluation factors for AI.

Risk source	Evaluation factor
Level of automation	Controllability, Explainability, Jurisdictional issues
Lack of transparency and explainability	Explainability, Predictability, Transparency
Complexity of environment	Robustness, Reliability, Resilience
System life cycle issues	Explainability, Predictability, Transparency, Reliability
System hardware issues	Robustness, Reliability, Resilience
Technology readiness	Precision, Recall, Accuracy, F1 score
Risk sources related to machine learning	Predictability, Fairness

The proposed method in this article

The handovers involving people, depending on the level of automation, can pose a risk, and the regulations governing the handover process between humans and AI can vary by country, potentially limiting the transfer of certain functions. To manage this risk source and ensure successful collaboration between humans and AI, it is important to consider evaluation factors such as controllability, explainability, and jurisdictional issues.

Transparency and explainability are essential for building trust in AI. Lack of transparency and explainability is another risk source that can lead to a lack of trust in AI. Explainability, predictability, and transparency are evaluation factors that can help manage this risk source and ensure that people understand the basis for AI's judgments.

The complexity of the environment can also be a risk source for AI. A complex and diverse environment can affect the performance of AI, leading to potential risks. To manage this risk source, evaluation factors such as robustness, reliability, and resilience can be implemented.

The different characteristics of AI-enabled system development from traditional systems can create a risk source related to system life cycle issues. Applying traditional development methods as-is may not be sufficient to verify AI software, which can result in potential risks. To manage this risk source, evaluation factors such as explainability, predictability, transparency, and reliability can be implemented.

Hardware or sensor defects can pose a risk source for AI, leading to incorrect measurements or interruptions in service. In addition, a lack of system performance or communication bandwidth for AI can cause risk. Evaluation factors such as robustness, reliability, and resilience can help manage this risk source.

Using less technically mature AI algorithms or models for real work can pose a risk source related to technology readiness. Evaluation factors such as precision, recall, accuracy, and F1 score can help manage this risk source.

Machine learning or deep learning is a common approach used in AI, and risks may arise during its development. Poor quality or inadequate learning processes can result in potential risks. To manage this risk source, evaluation factors such as predictability and fairness can be implemented.

By understanding and managing these risk sources and evaluation factors, developers and organizations can ensure that their AI-enabled systems are trustworthy, reliable, and beneficial for MASS.

4. Review of related guidelines

According to the IMO's autonomous shipping implementation plan, various studies are being conducted in the industry, and classification societies are managing autonomous shipping technology by providing guidelines for autonomous ships and awarding certification for systems that comply with them. Det Norske Veritas (DNV) published guidelines for autonomous and remote operation ships, American Bureau of Shipping (ABS) also released guidelines for autonomous and remote-control functions. Both DNV and ABS's guidelines describe the procedures and functional requirements necessary to develop and operate ships that are operated autonomously or remotely. In this section, we aim to apply the proposed combination of risk sources and evaluation factors to the DNV and ABS MASS-related guidelines.

4.1. *MASS-related guideline of DNV*

DNV guidelines cover the management of new operational concepts and human functions that are not accommodated by existing regulations. These guidelines define the requirements for navigation, ship engineering, remote control centre, and communications, with a particular emphasis on crucial cybersecurity and software testing in autonomous and remote operations. Applying the level of automation, one of the risk sources, to the DNV guidelines, the results are presented in Table 5. The table presents requirements for automation and support from personnel on board, local/manual actions, and auto remote vessels. The evaluation factors for these requirements are controllability, explainability, and jurisdictional issues.

4.2. *MASS-related guidelines of ABS*

ABS guidelines establish a goal-based framework for implementing autonomous and remote-control functions in ships and marine structures. The guidelines cover the functions that should be performed based on a four-stage decision loop (monitoring, analysis, decision, action), as well as interactions with stakeholders. In particular, the guidelines use a risk-based approach to describe the requirements for evaluating and implementing autonomous and remote-control functions. The results of applying the level of automation, one of the risk sources, to the ABS guidelines are presented in Table 6. The requirements cover the extent of automation and support from personnel, local/manual actions, operator and operations supervision level, possibility of retaking control, and final integration and onboard test. The evaluation factors for these requirements are controllability, explainability, and jurisdictional issues.

Requirements	Reference	Evaluation factors
1.2 Extent of automation and support from personnel on board automatic support (AS) Operation of the vessel function by automation systems and personnel in combination. Automation system(s) may partly or fully perform data acquisition, interpretation and decision. This mode is a collective term for all variants of decision support where the automatic support function may need complementary human sensing, interpretation or decision-making and where the action is not automatically effectuated.	Section 5 chapter 1	Controllability, Explainability
2.3 Local/manual actions 2.3.2 Auto remote vessels For auto remote vessels, it is generally not considered feasible to mitigate effects of failures and incidents by manual actions performed on board. 2.3.3 Automatic Operation (AO) Even if conventional manual operations on board will be replaced by purely automation systems, capabilities for remote supervision and emergency control should be arranged in the RCC. 2.3.4 Automatic Support (AS) If conventional manual operation on board will be performed by the remote engineering watch in RCC, decision support functions should be arranged which provide a firm basis for making decisions and executing control actions.	Section 5 chapter 2	Controllability, Jurisdictional issues

Source: Autonomous and remotely operated ships (DNV-CG-0264), DNV, September 2021

Table 6. Evaluation factors related to the level of automation in ABS guidelines.

Requirements	Reference	Evaluation factors
2.1.2 Operator and Operations Supervision Level An operator is to be designated and will have responsibility over the Autonomous Function. The operator may be physically located onboard the vessel or in a remote location. The operator station is to be constantly manned. (i) The operator is to supervise the function executions either continuously, periodically or as needed (ii) The operator is to be able to intervene, override, and take over the operation when deemed necessary by the operator	Section 5 chapter 2	Controllability, Explainability
2.5.2 Possibility of Retaking Control It is to be possible for the Operator to intervene and regain control of the action from the autonomous function at all times.	Section 5 chapter 2	Controllability
3.4 Final Integration and Onboard Test Manual Control (for autonomous function) The operation of manual control takeover using human interface systems and controls onboard is to be confirmed to be functioning satisfactorily. Manual Control (for remote control systems) The operation of manual controls at the remote controlling station using human interface systems and controls is to be confirmed to be functioning satisfactorily.	Section 7 chapter 3	Controllability, Jurisdictional issues

Source: Requirements for autonomous and remote control functions, ABS, August 2022

4.3. *Discussion*

To provide a comprehensive understanding of potential risks associated with MASS, the DNV and ABS guidelines were thoroughly examined for seven risk sources, encompassing those identified in Sections 4.1 and 4.2. This in-depth analysis allowed for a more detailed evaluation factors associated with each risk source. The results of this review are presented in Table 7, which offers a clear overview of the potential risks and the degree to which they are addressed in the guidelines. This comprehensive review helps to identify potential gaps and areas where further improvement is necessary to effectively manage risks associated with MASS.

Table 7. Comparison requirements between ABS and DNV MASS-related guideline

Risk source	DNV document	ABS document
Level of automation	★★★	★★★
Lack of transparency and explainability	★☆☆	★☆☆
Complexity of environment	★★★	★★★
System life cycle issues	★★☆	★★☆
System hardware issues	★★★	★★★
Technology readiness	★★☆	★☆☆
Risk sources related to machine learning	★☆☆	★★☆

The symbol ★★★ indicates that the classification societies' requirements regarding that specific risk source are adequately described, ★★☆ signifies that only certain aspects of the risk source are described, and ★☆☆ suggests that additional consideration is required for the requirements.

The guidelines from both DNV and ABS specify the roles and responsibilities of people and systems based on the level of automation in autonomous shipping. However, it might need to be noted that there is room for improvement in terms of transparency and explainability in both guidelines, as they have some deficit part of specific requirements. Considering this point of view could help ship's safe navigation as introducing AI. Higher complexity of the environment means more risk sources with more complicated situational awareness and more difficult decision-making, as well as more kinds of regulatory situations.

While ABS uses the adaptable V-model, DNV requires adherence to traditional development processes such as ISO/IEC 12207. If these guidelines are to be utilized in AI development as well, it is imperative to take into account the distinctive nature of AI development processes. Both DNV and ABS acknowledge the importance of addressing hardware failures, such as power loss and fires, as part of their system hardware requirements. DNV also places significant emphasis on requirements related to communication errors. Both consider software quality measurement and imply the needs of metrics. Also, they mention data quality which is one of the key parts of safety and security of AI

5. Conclusion

This article aims to provide a comprehensive understanding of the potential risks associated with the use of AI in automation systems in the maritime sector. Despite the significant benefits and opportunities that AI offers, it also poses potential risks and challenges that must be managed properly. To address these concerns, this article proposes combination of risk sources and evaluation factors for AI that combines international standards such as ISO/IEC 22989, ISO/IEC 23894, and ISO/IEC 23053. The combination provides a valuable tool for understanding the various sources of risk and their corresponding factors that need to be considered when developing and operating AI-enabled systems in the maritime sector.

In addition to proposing the combination, this article also reviewed DNV and ABS guidelines for seven risk sources to provide a more detailed evaluation factors associated with each risk source. The findings of this article provide a practical approach to managing the risks associated with the use of AI in the maritime sector, and highlight the need for industry-specific risk management frameworks and guidelines that consider the unique challenges and risks associated with the use of AI in the maritime sector.

Moving forward, the plan is to select target systems for MASS (such as collision avoidance systems using cameras) and examine the risk sources and evaluation factors in detail, creating metrics that can quantitatively evaluate them.

In conclusion, this article contributes to the growing body of knowledge on managing the risks associated with AI in the maritime sector and can help ensure the safe and efficient operation of MASS. By understanding and managing the risk sources and evaluation factors proposed method, developers and organizations can ensure that their AI-enabled systems are trustworthy, reliable, and beneficial for MASS.

Acknowledgment

This research was supported by Korea Institute of Marine Science & Technology Promotion (KIMST) funded by the Ministry of Oceans and Fisheries, Korea (RS-2023-00238653).

References

Alsos, Ole Andreas, et al. "Maritime autonomous surface ships: Automation transparency for nearby vessels." Journal of Physics: Conference Series, vol. 2311, no. 1, IOP Publishing, 2022.

Öztürk, Ülkü, Melih Akdağ, and Tarık Ayabakan. "A review of path planning algorithms in maritime autonomous surface ships: Navigation safety perspective." Ocean Engineering 251, 2022, 111010.

Miyoshi, Toshiyuki, et al. "Rules required for operating maritime autonomous surface ships from the viewpoint of seafarers." The Journal of Navigation 75, no. 2, 2022, 384-399.

Issa, Mohamad, et al. "Maritime Autonomous Surface Ships: Problems and Challenges Facing the Regulatory Process." Sustainability 14.23, 2022, 15630.

Razmjooei, Damoon, et al. "Investigating Industry 4.0 Concept And Its Technologies in the Maritime Industry." Road 31.114, 2023, 235-244.

Lee, Seojeong, et al. "A study of S-100 based product specifications from a software implementation point of view: focusing on data model representation, similar features and symbols, and ECDIS and VTS software." The Journal of Navigation 75, no. 5, 2022, 1226-1242.

ISO/IEC 22989: Information technology -- Artificial intelligence -- Artificial intelligence concepts and terminology. International Organization for Standardization, 2022.

ISO/IEC 23053: Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML). International Organization for Standardization, 2022.

ISO/IEC 23984: Information technology -- Artificial intelligence --Guidance on risk management. International Organization for Standardization, 2023.

DNV. Autonomous and remotely operated ships (DNV-CG-0264), 2021.

American Bureau of Shipping (ABS). Requirements for autonomous and remote control functions, 2022.