TRUSTWORTHINESS EVALUATION FRAMEWORK FOR DIGITAL SHIP NAVIGATORS IN BRIDGE SIMULATOR ENVIRONMENTS

Hosna Namazi
UiT, The Arctic University of Norway
Tromsø, Norway

Lokukaluge Prasad Perera
UiT, The Arctic University of Norway
Tromsø, Norway

ABSTRACT

The maritime industry is going towards implementing digital navigators, i.e., AI created by machine learning algorithms, on autonomous vessels in the future. Digital navigators can be developed by utilizing machine learning algorithms, e.g., deep learning type neural networks trained by data sets from human navigators. Even though there is significant importance in studying the trustworthiness of these digital navigators, a proper framework to evaluate it has not yet been developed. This study identifies the appropriate key performance indicators (KPIs) in the trustworthiness of digital navigators in autonomous vessels.

The trustworthiness of AI-based applications, including digital navigators, can be studied from two primary levels: Software and hardware levels. Each of these levels must have certain characteristics to be called trustworthy. In other words, software codes and algorithms should be Transparent, i.e., Explainable, Fair, and Accountable/Responsible. Moreover, the trustworthiness at the hardware level can be elaborated under two concepts of Resilience and Availability of the relevant systems and technologies. In addition, some concepts, such as Reliability, Privacy, Security, and Safety, should be studied for both levels since those concepts can overlap in both software and hardware levels.

In this paper, the main focus is on investigating the software's trustworthiness. After an introduction on the importance of the topic and digital navigator's development steps, the existing literature on trustworthy AI is reviewed, and the proper approaches for evaluating trustworthiness in AI-based digital navigators are identified and proposed.

Keywords: Artificial Intelligence, Machine Learning, Shipping, Maritime, Human Navigator, Digital Navigator, Trustworthy AI, Explainable AI, Fairness, Responsible AI.

1. INTRODUCTION

The commercial and economic importance of the shipping industry is indispensable since that accounts for the transportation of around 90% of traded goods globally [1]. In 2019, the international maritime trade volume was equal to 11,076 million tons of loaded goods [2], which is constantly growing under various economic conditions. With the growing development in the field of sensor technology and decision support algorithms, various technology developers are going towards replacing conventional shipping technologies with autonomous applications due to their low operation and maintenance costs, fewer maritime traffic congestions, and consequently better air quality and safety [3].

Although humans are better to cope with complex decision-making situations than machines, human errors in decision-making in ship navigation and operation situations are widely regarded as the primary cause of the respective accidents [4]. This can be caused by tiredness or an overwhelming workload. Since the human ability to process large amounts of data is limited, an essential factor in preventing accidents is a proper human-machine system interface to support collision avoidance type situations in conventional ship navigation [5]. On the other hand, the transition towards autonomous shipping would be a disruptive innovation progression in the evolution of shipping, and there is a growing interest in developing autonomous vessels. Many technology developers worldwide are trying to develop and adopt new technologies to make autonomous shipping a reality.

In this study, the basic assumption is that there is a general understanding of the term "autonomous" as a structure and state of a system, i.e., what it means to be an autonomous system, as opposed to automation as a technology. If a brief definition of the term is needed, one can say that an autonomous system will execute the function, generally without the possibility for a human to intervene on the functional level [3]. Based on this

definition, the term' autonomous vessels' refers to self-navigating vessels. Relling et al. [6] conducted an excellent survey of the term "autonomy"; for further details, one can refer to this research.

One of the essential characteristics of an autonomous system, which distinguishes it from conventional systems, is decision-making without human inference. Autonomous ships can act independently and take proper decisions and actions in different scenarios. For this purpose, powerful data analysis tools should be utilized to extract information from vessel operational and environmental conditions. The difficulty here is that usually, there is a large amount of data and conventional methods based on various empirical ship performances, and navigation models cannot handle this amount of information on a real-time basis. Other problems for conventional mathematical models are system-model uncertainties, sensor noise and fault conditions, and complex parameter interactions that cannot be handled by the same. As a result, such models may not predict actual ship performance and navigation information correctly, jeopardizing the validity of ship navigation strategies and decisions taken afterward [7].

However, autonomous vessels still face operational challenges in complex navigational environments. In this respect, a considerable amount of research studies have been conducted in this area [8]. Perera [9] presented a structured framework to address navigation considerations, including collision avoidance of autonomous ships, and introduced deep learning as a suitable choice for the core algorithm of the decision-making module for navigating autonomous vessels. Murray and Perera [10] formulated a deep learning framework to predict the future 30-minute trajectory of a selected vessel using historical AIS data to aid in proactive collision avoidance by decomposing the historical ship behavior in a given geographical region into clusters using a variational recurrent autoencoder. Bui and Perera [11] developed a big data analytics framework using Gaussian Mixture Model for capturing the clusters in a data set obtained from a bulk carrier. This framework was utilized for localized ship performance monitoring by providing key performance indicators for quantifying trim-draft performance for different engine modes. These are some of the decision support systems that can also be utilized for autonomous shipping. Taghavi and Perera [12] used Gaussian Mixture Models to capture the operating regions of an ocean-going vessel's main engine using one month of engine power, fuel consumption, and engine speed data. They also performed a singular value decomposition in each cluster to find the relationship between the variables, which can serve as a basis for digital twin development. One should note that digital twin applications can play an important role in autonomous future vessels.

Several technology providers are expected to introduce autonomous vessels into the real-world ocean environment in the near future. These autonomous vessels should be equipped with reliable digital navigators based on AI to navigate them in the ocean, and such systems should guarantee the required safety levels of vessel navigation. Hence, these digital navigators should be able to act similarly to human navigators in various situations, especially in close vessel encountering scenarios with other vessels, which can lead to possible ship collision situations. In other words, decision-making systems should be provided to digital navigators to make it possible to detect obstacles or other ships and make proper decisions to avoid near-encounter situations, including possible collision situations.

Moreover, it will be times when there are manned vessels, as well as autonomous/remote-controlled vessels in the ocean, also known as a mixed environment, and this will introduce additional complexity to the situation. So, AI-based digital navigators should be able to guide ocean-going vessels in mixed environments safely while avoiding any possible collision situations. Of course, due to vessel under actuation or maneuvering difficulties, it would be difficult for ocean-going vessels to execute some navigation decisions. Such difficulties can influence near-encounter scenarios. Furthermore, it would be challenging to evaluate such close encounter situations due to several reasons, such as the high costs associated with conducting such exercises and the fact that it is not possible to test developed digital navigators in a real ocean environment due to environmental conditions' influence on vessel maneuvers. Since such environmental conditions can introduce additional collision risk, using bridge simulators for simulating these scenarios and testing developed algorithms and systems to support the digital navigator can be a safer and more economical approach, at least for the initial evaluations. More matured algorithms and systems to support autonomous ship navigation can then be implemented under real ocean environmental conditions by considering the results from the bridge simulators.

Although data-driven methods have many benefits, there are some downsides to these techniques, mainly originating from their limitation on a limited number of tasks. On the other hand, human is capable of handling a complex number of tasks, and that may relate to the difference between the human brain and neural networks. Most of the tasks that are going to be automated in future vessels will be formulated previously based on the human mind and behavior. These tasks have evolved to be convenient for the human mind, and it should be guaranteed that they are in coordination and harmony with data-driven machine-learning algorithms that can create autonomy. Therefore, the trustworthiness of AI based systems, such as autonomous vessels, should be evaluated to guarantee reasonable behavior from intelligent systems.

Many of these tasks have the natural ability of humans, and making them a part of an autonomous system requires a great deal of effort and expertise. To find solutions to this problem, one should look at the differences between the human mind and machines. There is a fundamental difference between humans and machines based on their capabilities. These basic differences between humans and autonomous systems or machines arise from humans' superiority in detection, perception, and judgment. Machines will do what they have been programmed to do and if an unprecedented situation comes up, automated systems may fail to have a proper reaction in some situations. This is due to the reason that the development of proper consciousness in

machines is still a challenging problem for the research communities. Hence, an adequate trustworthiness framework for autonomous systems should be developed to avoid any possible failures at the respective decision-making levels. In the best cases, these systems try to find the most similar scenario to the current situation and then utilize the respective actions since they cannot think outside the box in some situations. Therefore, adequate measures to guarantee trustworthiness in such situations should be developed at a system level.

In some systems, the software and hardware units can play an important role in controlling the respective systems, and a light problem or malfunction in the system hardware or software, i.e., just negligence from the software developer, can lead to catastrophic situations. The crash of Boeing 737 MAX of Lion Air, Flight 610, in October 2018, and the crash of Boeing 737 MAX of Ethiopian Airlines, Flight 302, in March 2019 [13] are some examples of software and hardware malfunction in an autopilot system. So, many aspects and precautions about these kinds of systems should be taken into account during their operational phases, especially in autonomous navigation situations.

Despite all the differences between AI algorithms and the human mind, digital navigators should be able to act like human navigators in different situations. That is why these navigators are usually developed based on the data sets from human navigator behavior on real-world encounter scenarios using deep learning based neural network modeling. It is clear that human navigators can make proper decisions in various situations based on both their talent and innate ability, as well as the skills and knowledge they have acquired throughout training and experiences. However, this might raise the question of if a digital navigator is developed by cloning human decisions and actions, how will it eliminate the errors caused by human decisions. To answer this question, one should bear in mind that multiple data sets are going to be used for the development of digital navigator's training, and most of the data used to this end are going to be the right decisions, i.e., remove the data sets that consist of wrong decisions and actions. However, the erroneous part of data sets will be detected as anomalies by AI-based approaches, and it will be isolated, recovered (if possible), and compensated with the right data from the respective application.

2. DIGITAL NAVIGATOR

In the previous section, some of the characteristics of AI-based digital navigators are briefly discussed. However, there are several key pillars in autonomous ship navigation to support the digital navigator. As was discussed, one of the most important yet neglected elements in the development of the digital navigator is the human navigator's role. Similar to the automotive industry, the human navigator's skills should be cloned under a deep neural network to create digital navigators. Therefore, these digital navigators will be able to mimic human capabilities and behaviors to replace them in autonomous ship navigation. As it is shown in FIGURE 1, the next component is the technology, i.e., automation and navigation systems.

Human navigators should interact with this technology, i.e., navigation and automation systems, to achieve navigation objectives, such as controlling speed, steering, etc. Meanwhile, when human navigators are interacting with the technology, the information from these systems should be illustrated to them using onboard visual aids. In addition, this visualized information should not be in a way to overwhelm human navigators in some situations, and it should be properly presented to support human capabilities. The other key pillar is the rules and regulations, i.e., regarding ship operations, and all ship navigators should follow the respective rules and regulations to ensure safe voyages. Last but not least, in order to make autonomous shipping possible, AI or the digital navigator should be developed as the final element. Future autonomous vessels will be navigated by these digital navigators, which is created using the data obtained from human navigators' knowledge and skills during ship operations to mimic their behavior. AI-based navigators will also have similar interactions to human navigators' behavior with the required technologies to make autonomous ship navigation a success. However, the information from the respective navigation and automation systems should be visualized and conveyed to digital navigators in a way that it can be understandable for them. Also, like human navigators, digital navigators should comply with the rules and regulations of the ocean.

So, there should be a single regulatory framework to be followed by both humans in manned vessels and digital navigators in autonomous vessels in mixed environments. Eventually, in the future real ocean environment, where both manned and autonomous vessels co-exist, these digital and human navigators should be able to interact with each other. Finally, the trustworthiness of digital navigators must be studied from both the software and hardware level to facilitate the required safety conditions. For evaluating a digital navigator's trustworthiness, it is better to have a first look at the actions that a digital navigator can be taken in different navigation situations. The digital navigator's decisions can be divided into two categories: changing the speed and/or changing the vessel course. In other words, the digital navigator can decide to increase, decrease, or keep the vessel speed and to change the course to port, starboard, or keep it straight to navigate the respective vessel in different scenarios and avoid possible collisions or near-miss situations. By identifying possible ship maneuvers, it is much easier to quantify and validate the trustworthiness of digital navigators, especially in simulator environments.

3. AI TRUSTWORTHINESS

Although the introduction of AI into different aspects of life is beneficiary and influences individual, societal, and economic developments, it also comes with ethical and legal challenges. Therefore, these AI-based applications should follow ethical and legal values to be trusted and so called trustworthy. Then, such technologies can be adopted into various industrial sectors.

To realize the importance of a trustworthy AI, we should first understand the definition of trust and its necessity. Based on

the Merriam-Webster Dictionary [14], trust is defined as an assured reliance of a party on the ability, strength, and character of another party since the first party does not have control over the actions of the other one in doing a responsibility. This action is usually accompanied by taking a risk for the first party.

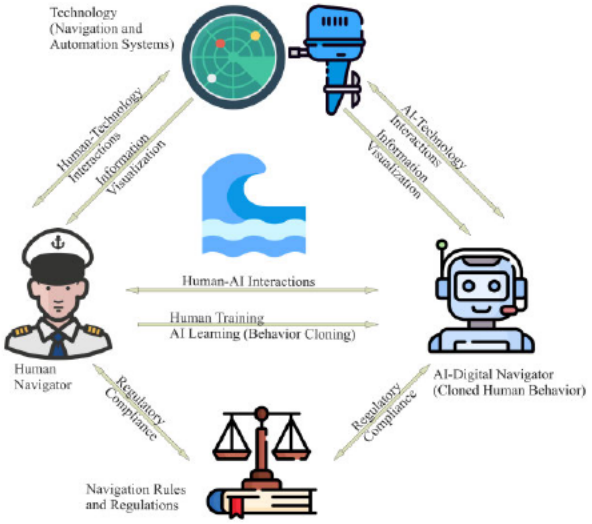

FIGURE 1: KEY PILLARS IN AUTONOMOUS SHIPPING

Trustworthy AI has been an open debate for a while. Several researches have been carried out in this field to identify the right factors of trustworthy AI and its characteristics despite different terminologies used in various research studies. For instance, Vadlamudi [15] mentioned five principles to make trustworthy AI which is usefulness, non-maleficence, autonomy, justice, and logic. In another study [16], being lawful, ethical, and robust are mentioned as three criteria for making reliable AI. Shneiderman [17] highlighted Human-Centered Artificial Intelligence (HCAI) methods to produce systems that are safe, reliable, and trustworthy. These methods indicate the extent of human control and computer automation applications to enhance human performance and avoid the dangers associated with excessive human or system controls. Researchers also took a look into trustworthy AI in various industries. For instance, Haspiel et al. [18] studied the trust between humans vs. automated vehicles (AVs) and the effect of explainability in the adoption of AVs. They focused on the timing of giving explanations on AVs' behavior to human drivers, and the results showed that the provided explanations before the respective actions could increase user acceptance and trust. Dikmen and Burns explored the level of drivers' trust in the autonomous features of Tesla cars. The results showed that different indicators, such as the personal experience and usage time of these systems, can affect the user's trust. Apart from these, the transparency in decision-making processes and having awareness about the related mechanisms that these systems use to handle different situations can influence the reliability in autonomous systems [19].

Studying trustworthiness is not limited to the transportation industry. Several research studies have been carried out in other fields, such as robotics, healthcare, and finance, to understand trustworthy AI. For instance, Langley et al. investigated the explainability of intelligent autonomous systems in automotive and robotic industries. They stated that transparency in these systems' decision-making, like every other system, will increase user trust and satisfaction [20].

As it was discussed before, deploying AI as digital navigators can also have similar challenges. In FIGURE 2, a two-level trustworthiness study framework of an AI-based digital navigator is presented. That consists of both software and hardware level components that relate to trustworthy AI. The software level focusing on the developed AI algorithms should be transparent, fair, and responsible. Furthermore, the systems that will be used in the implementation and development of these navigators should be trustworthy as well. In other words, they should be resilient and available. Apart from these, both levels should pass the reliability, privacy, security, and safety considerations to be able to trust the digital navigators.

4. TRUSTWORTHINESS AT SOFTWARE LEVEL

In this section, the approaches for studying trustworthy AI at a software level, considering characteristics such as Explainability/Transparency, Fairness, and Responsibility/Accountability in autonomous vessels, i.e., digital navigators, are presented. Since there are several comprehensive reviews on different characteristics and categories of AI trustworthiness, it is reasonable to use the existing approaches to evaluate the trustworthiness of AI-based navigators as well. Therefore, the relevant research and comprehensive reviews in each category are studied, and the important current applications are selected and presented in the following sections.

4.1 Explainability of the Digital Navigator

Nowadays, AI applications are used to find solutions for various industrial challenges. With an unprecedented development speed in some areas, many AI-based applications are successfully functioning without human interference. However, these applications introduce new challenges in some situations, e.g., due to data quality issues or abnormal system events. It is important that these AI approaches should be transparent and explainable, especially in such situations where human lives are affected, such as healthcare, judicial, and transportation sectors. Transparency and the overall model understanding of these systems are highlighted in debugging and understanding of whether the models make decisions based on the right information or not. Also, bias detection, assessing the suitability of the model for deployment, and understanding whether the model predictions in decision-making situations are trustworthy or not can be other reasons for developing transparent systems. Adadi and Berrada [21] conducted a review paper that discusses the methods that can be used to help the AI-based models be transparent and explainable so that they are not stopped from functioning due to lack of transparency. Guidotti et al. focused on black-boxed decision support systems. They classified the

different explanation methods of black boxes based on their type and the applications they are used for [22].

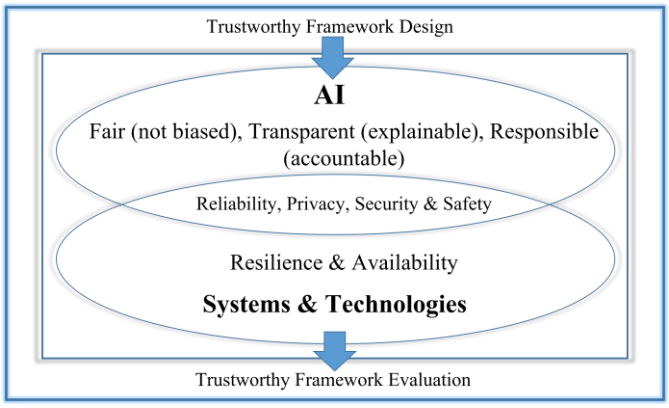

FIGURE 2: TRUSTWORTHINESS EVALUATION FRAMEWORK OF AI

The proposed digital navigator is developed based on the data collected from the behavior of human navigators using Deep Neural Networks (DNN). In this approach, a series of scenarios, along with the correct response of the human navigator, is fed into the neural network, and the network learns and simulates the correct behavior in different scenarios. In the future, based on the similarity of the situation to one or a combination of the scenarios, the digital navigator can decide its action. As a result, an explainability evaluation framework suitable for Deep Neural Networks should be selected for the digital navigator. In [23], it is suggested that post-hoc local explanations and feature relevance techniques are the most suitable and adopted methods for explaining Deep Neural Networks. The explainability evaluation framework suitable for Deep Neural Networks has been categorized based on three different groups of networks, Multi-Layer Neural Networks or Multi-Layer Perceptron (MLP), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) [24]. By considering all the highlighted criteria from the reviewed articles in [24] for each neural network, the frameworks for mentioned categories are as follows.

a- Multi-Layer Neural Networks (MLP)

Many researchers have proposed different explainability criteria for MLP, including model simplification, feature relevance estimator, text explanation, local explanation, and model visualization approaches.

Model simplification can be performed by decomposing the neural network and rule extraction by creating rule-based graphs or decision trees. This approach can be applied to simple networks with one hidden layer, but since the MLP can be used for the digital navigator that has a complex architecture, model simplification may not be an adequate method for explaining this network. Simple MLPs usually can perform simple tasks, and as the task gets more complicated, the number of neurons and hidden layers can increase, which makes it very difficult to utilize the model simplification approaches. In an MLP with several layers and neurons, simplifying itself can result in a complex set of steps that is not easy to explain. As a result, explaining neural networks by feature relevance methods has become more popular approaches.

One of the approaches [25] for this purpose is to decompose the network classification decisions into contributions of its inputs. In this method, the mathematical logic and algebraic equations of each neuron are considered, and each neuron is looked at as an object that can be decomposed and expanded. Then, these decompositions are back-propagated through the network from the output to the input. In this way, the relationship between the output and each of the network inputs is derived. Apart from these methods, text explanations, local explanations, and model visualizations can be used to clarify the network performance. A local explanation identifies what dimensions of a single input were the most responsible unit for a DNN's output.

b- Convolutional Neural Networks (CNN)

CNNs are widely used for image processing applications, which have crucial importance in digital navigation development. These networks consist of a sequence of convolutional layers which can automatically learn a high level of features. Explainability for CNNs is more straightforward than other types of networks because of human cognitive skills in visual data. These networks try to find different features or objects in each image, which is a natural ability of the human mind. As a result, it is easier to explain the performance of each layer, i.e., object detection, to a human mind.

There are many innovative approaches proposed in the literature [24] that can be used for this purpose. For instance, one of the ways to identify each layer process is to use the occlusion sensitivity method, which is iteratively feeding the network with the same image and blocking a different region at a time [26]. Generally, these approaches aim to understand the CNNs performance, and that can be categorized as:

1- Understanding the process by mapping back the output into the input to see which parts of the input were contributing to the output.

2- Interpreting how the layers work in general and not necessarily related to any specific input.

c- Recurrent Neural Networks (RNN)

RNNs have been extensively used for predictive applications of sequential data and time series analysis. In a digital navigator, a vessel route can consist of latitudes and longitudes in the form of a series of data points. A series of these coordinates define the path a selected vessel has taken, and the future path of the ship can be predicted using this historical position data. Such information can also be useful for close ship encounters with collision avoidance type maneuvers. Time series datasets exhibit long-term interdependencies that are too complex for a machine learning (ML) algorithm to analyze. On the other hand, RNNs can handle these time dependencies by preserving knowledge in the respective neurons.

The research studies pertaining to explain RNN models can be divided into two main groups [24]:

1- Understanding what an RNN model has learned from the input time series mainly by feature relevance methods.
2- Modifying RNN architectures one at a time to get insights about their decision-making process as an approximately local explanation approach.

In developing the digital navigator, all these three kinds of Neural Networks will be used for different purposes, such as trajectory estimation [27], image processing application, and decision making, and considering the complex structure of these navigators' nature. As a result, it is not possible to explain the resulting model and check its transparency using simple approaches. Therefore, all three categories presented here must be considered in evaluating the explainability of a digital navigator.

4.2 Fairness of the Digital Navigator

As it was discussed in the previous studies, using AI-based approaches in decision-making and decision-support applications is getting popular, which has a great impact on human lives both in personal and societal levels. These algorithms have been incorporated into many industrial applications, as discussed before. Moreover, these applications are developed by humans using data from real-world situations, which are susceptible to bias.

Therefore, it is crucial to make sure that suggestions and decisions made by such applications or systems may not be biased in favor of a certain group, and the feature selection process for decision-making may not discriminate against specific groups or characteristics. Many researchers have presented different approaches recently to address such challenges and try to propose guidelines and frameworks to mitigate their adverse consequences. Mehrabi et al. [28] have presented a comprehensive review of different sources of biases that can affect AI applications and proposed a taxonomy for fairness definitions for machine learning applications. Richardson and Gilbert [29] have also performed a systematic review of the algorithmic bias issues that have been defined, and fairness solutions have been proposed. Alikhademi et al. [30] performed a review of the fairness of machine learning algorithms in the criminal justice system with predictive policing. There are also other articles in the literature that performed a review in this context, such as Fairness Testing [31] and Software Fairness [32]. Madaio et al. [33] conducted an iterative co-design process with 48 AI practitioners from 12 technology companies, working on 37 separate products, services, or consulting engagements, to develop a checklist for AI fairness. The study claimed that this approach guarantees that the proposed checklist meets the practitioners' expectations and the presented checklist can provide organizational infrastructure for formalizing ad-hoc processes.

From all the comprehensive studies in the recent literature, Agarwal and Agarwal [34] presented a step-by-step model for formulizing AI fairness evaluations inspired by an open system interconnection (OSI) model to standardize AI fairness handling. The study claimed that despite the previous works that have been done in this area, they had developed a seven-layer model that studies the system fairness in all its lifecycle stages. In this paper, the seven layers of an AI lifecycle are introduced as the purpose layer, data selection layer, pre-processing layer, algorithm layer, training layer, independent audit layer, and usage layer. Since this study proposes a comprehensive model compared to other papers, it is suggested to investigate the digital navigator's fairness by the same approach, which is discussed in the following sections.

The first layer aims to inspect the probable biases that may exist before data handling or the coding stages. It is stated that the developers should study the relevant documentations and existing models thoroughly to understand the features prone to biases, as well as the acceptable limits of the respective bias. Also, in this layer, the fairness study approach that will be used in the next stages should be identified.

Generally, the second layer aims to find and remove the biases regarding the data collection sources and label them. Since one of the leading reasons for biased AI is proven to be due to biased training data, it is important to recognize them and eliminate them in the first stages before the training process begins. The collecting process can also be the reason for the unfair AI, from the biases considering the people or sources that the data is collected from, to the devices that gather data. Apart from those, some biases might be introduced during the labeling process. It is recommended that the data get labeled by the respective researchers or organizations themselves rather than third parties since the personal insight of individuals can cause additional biases. In conclusion, it is important to make sure that the data is accurate, complete, well-distributed, and verified.

As it is mentioned, the third stage is allocated to data pre-processing. During this process, the raw datasets are prepared for analyzing through some stages that can easily result in introducing bias in the AI application. For example, aggregating, scaling, removing outliers, and introducing a lot of estimated data can cause unfair AI models. To avoid unfairness in the algorithm development step is the goal of this stage. Whether the decision-making algorithms are developed by humans or automated tools, they can be affected by biases in the respective decision-making processes. It is important to develop explainable and transparent algorithms to understand if it is working based on the right and fair information or not.

The fifth layer studies the model training fairness. It is important to make sure that the data used for training, testing, and validation steps are well divided and balanced. Also, training and performance study parameters, as well as benchmarks to evaluate the model performance, should be chosen correctly to avoid any unfairness. The fairness of the developed model should be reviewed by an independent audit in the sixth stage, and the standard fairness process should be checked for all the protected attributes defined by the developers in the previous

step. This step is especially important for the AI applications used at a societal level to make sure that the platform is just for all members of society and is not biased towards a specific group.

After introducing an AI-based system to the end user, it is still important to check its fairness and retrain the respective model based on the new data periodically. Also, the performance indicator of the developed system should be checked regularly to understand when to retrain the system. This approach can be more beneficial compared to most studies, which only focus on bias in datasets and algorithms, as it suggests a comprehensive method that includes all stages of AI development, from conceptualization to implementation. Given that the stages involved in constructing AI models are often similar, that is why this approach can also be applied to digital navigators. Ensuring fairness in AI-based digital navigators at each stage of development can effectively mitigate any potential biases to a good extent.

4.3 Accountability and Responsibility of the Digital Navigator

A meaningful accountability process is composed of three phases: information, explanation or justification, and (the possibility for) consequences [35]. Although extensive research studies have been focused on explaining the performance of AI applications, the respective research studies on the accountability of AI applications are not as extensive. Since many advanced AI applications used in industrial or scientific domains are developed by a group or an organization, accountability is not usually the main interest of individual researchers, who may focus on the technical aspects of AI. It is essential to promote responsible humans, research institutes, nonprofit organizations, or even governmental institutes in times of adverse consequences of AI.

In [36], Dignum proposes a definition of accountability and responsibility of AI. Accountability of AI refers to the system's ability to justify its decisions based on the algorithms used by itself to the end users. On the other hand, responsibility is defined as the role of people and developers in the AI's actions and decisions, and it is not only governing the AI applications based on the rules but also being responsible for the whole socio-technical environment in which the systems and people are interacting.

Vassilakopoulou [37] proposed seven principles considering a socio-technical perspective for the design of accountable AI through regulations and operational coordination. Variance control, boundary location, and power authority are three principles that aim the accountability through the respective regulations. Accountability through operational coordination consists of four principles: core process integration, task allocation among humans and machines, congruence, and design ownership by the managers and users.

Wang et al. claimed data governance, ethical design solutions, risk control, and training and education as four practices of a responsible AI [38]. By studying 10 cases of responsible AI in different industries, the study has identified five strategies, including the emergence of chief responsible AI officers, balancing economic and social sustainability of AI use, transparent and customer-centric data policy, creating socially responsible initiatives with AI, and reward and punishment mechanism to regulate AI usage to help other organizations in developing responsible AI.

Based on the research work of Cheng et al. [39], a socially responsible AI is designed and developed based on values such as fairness, transparency, reliability, privacy, and security. The main goal is to create algorithms that are not only functional and useful for society's needs but also align with its ethical values. Besides this definition, a pyramid that outlines the different responsibilities of AI in a hierarchical order is also suggested. According to this pyramid, a socially responsible AI must meet the functional, legal, ethical, and philanthropic responsibilities, which address its effective functionality, compliance with laws, compliance with societal and ethical expectations, and finally, the social challenges, respectively. The authors suggest that to create socially responsible AI, five steps should be taken.

Firstly, a foundation for the development of AI should be defined thoroughly. In other words, it should elaborate on which values are important in designing AI algorithms. In the second step, the social issues that the designed AI should address must be identified. Third, the means and approaches for developing a socially responsible AI should be selected. After developing the AI, it should be validated as the fourth step, meaning that it must be assured that it meets the defined goals. As the final step, after deploying the AI application in the real world, it should be monitored to make sure that it is socially responsible.

AI-based digital navigators should have similar characteristics to every other AI-based application discussed in this study, which means that it should pass all the criteria, such as fairness, transparency, as well as privacy, and security issues, to be called responsible. By implementing the suggested steps, this study will make sure that the digital navigators are functionally and legally responsible and can navigate future vessels based on the extant rules and regulations efficiently and effectively. Also, these digital navigators must meet the ethical requirements to address societal challenges as well.

5. DISCUSSION

Based on the previous discussions, it can be understood the characteristics of trustworthy AI are not completely isolated concepts. The features like transparency, fairness, and responsibility of an AI-based digital navigator are not mutually exclusive but are somehow related and dependent on each other. In other words, in some of the fairness validation steps, the developed machine learning algorithms must be transparent and explainable to understand how the AI is acting and based on what information it is operating. Likewise, in the responsibility validation, the algorithms should act fairly, and they should be explainable besides other features.

The trustworthiness studies may not be yet extensively relevant to the maritime domain since most of the studies for autonomous vessels are still in the research phase and the technical readiness for it is not high enough in the maritime field as opposed to other industries such as the automotive industry in

which there are existing examples of driverless cars. Hence, bringing relevant knowledge from other industries into the maritime industry is an excellent way to evaluate the developed machine learning algorithms, i.e., digital navigators, in the future.

The mentioned approaches for studying each of the characteristics in section 4 have been used and tested in other industries such as health care, automotive, and finance sectors since they are in a mature technology development phase compared to the maritime industry. The results of the evaluations show that the discussed methods can have a positive impact on the explainability, accuracy, and overall trustworthiness of AI-based applications. As an example, [34] presents a case study where a seven-layer model for evaluating AI fairness is used for analyzing an AI system to predict loan repayment risk. The system is trained and validated using the German credit dataset, but the target user is intended to be an Asian country. The respective findings demonstrate that although the AI application appears to be fair based on the fairness metrics and the overall fairness score, checks made at layers 1, 2, and 7 imply that it might not be relatively fair for the intended users.

A deep learning-based explainable AI approach has been implemented in driverless cars through the creation of explainable models that reveal the decision-making capabilities of these vehicles. This approach aims to enhance human understanding and trust in autonomous navigation systems. Karmakar et al. proposed two deep learning-based models to study the trustworthiness of autonomous cars. Based on the simulation results and real-world traffic data, the study concluded that the proposed methods could be utilized to evaluate the trust in driverless cars compared to non-learning-based methods [40].

6. CONCLUSION

The application of digital navigators in autonomous vessels will bring various challenges, such as AI trustworthiness. A trustworthy digital navigator must be explainable, and the decision-making process should be transparent. In addition, the digital navigator must act fairly and responsibly during its ship navigation.

In this paper, previous research studies done in various fields are reviewed, and the state-of-the-art approaches for confirming AI-based navigators' trustworthiness at the software level are gathered and presented. Based on the complex nature of machine learning algorithms developed for the digital navigator, some post-hoc explaining approaches, such as MLP, CNN, and RNN must be considered. To make sure that the digital navigator is fair, a seven-layered model is suggested to study the fairness in each step of its development from the concept layer to the implementation layer to detect and eliminate all possible biases. Finally, a framework is considered for the accountability and responsibility study of these navigators. It can be concluded that all these characteristics are somehow related to each other and are not mutually exclusive. To be fair and responsible, the machine learning algorithms for AI-based navigators must be transparent and explainable.

ACKNOWLEDGMENT

The work is a part of the Center of Excellence in Maritime Simulator Training and Assessment (COAST) originates from MARKOM and was established as the Center for Outstanding Education (SFU) in 2020.

This work is also supported by the MARKOM II project under the project title "Onshore Operation Center for Remotely Controlled Vessels (OOC 2023)" under contract number PMK-2022-10014.

REFERENCES

[1] "Organisation for Economic Co-operation and Development, Ocean shipping and shipbuilding." https://www.oecd.org/ocean/topics/ocean-shipping/

[2] E. S. Han and A. goleman, daniel; boyatzis, Richard; Mckee, *Review of Maritime Transport 2020*, vol. 53, no. 9. 2020. [Online]. Available: https://unctad.org/system/files/official-document/rmt2020_en.pdf

[3] "Autonomous and remotely-operated ships." https://www.dnv.com/maritime/autonomous-remotely-operated-ships/index.html

[4] L. Zhang, H. Wang, Q. Meng, and H. Xie, "Ship accident consequences and contributing factors analyses using ship accident investigation reports," *Proc. Inst. Mech. Eng. Part O J. Risk Reliab.*, vol. 233, no. 1, pp. 35–47, Feb. 2019, doi: 10.1177/1748006X18768917.

[5] E. Pedersen, K. Inoue, and M. Tsugane, "Simulator Studies on a Collision Avoidance Display that Facilitates Efficient and Precise Assessment of Evasive Manoeuvres in Congested Waterways," *J. Navig.*, vol. 56, no. 3, pp. 411–427, Sep. 2003, doi: 10.1017/S0373463303002388.

[6] T. Relling, M. Lützhöft, R. Ostnes, and H. P. Hildre, "A Human Perspective on Maritime Autonomy," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10916 LNAI, no. 2, 2018, pp. 350–362. doi: 10.1007/978-3-319-91467-1_27.

[7] T. Villegas, M. J. Fuente, and M. Rodríguez, "Principal component analysis for fault detection and diagnosis. Experience with a pilot plant," *Int. Conf. Comput. Intell. Man-Machine Syst. Cybern. - Proc.*, pp. 147–152, 2010.

[8] T. Kim and J.-U. Schröder-Hinrichs, "Research Developments and Debates Regarding Maritime Autonomous Surface Ship: Status, Challenges and Perspectives," 2021, pp. 175–197. doi: 10.1007/978-3-030-78957-2_10.

[9] L. P. Perera, "Deep Learning Toward Autonomous Ship Navigation and Possible COLREGs Failures," *J. Offshore Mech. Arct. Eng.*, vol. 142, no. 3, pp. 1–39, Jun. 2020, doi: 10.1115/1.4045372.

[10] B. Murray and L. P. Perera, "An AIS-based deep learning framework for regional ship behavior prediction,"

Reliab. Eng. Syst. Saf., vol. 215, no. May, p. 107819, Nov. 2021, doi: 10.1016/j.ress.2021.107819.

[11] K. Q. Bui and L. P. Perera, "Advanced data analytics for ship performance monitoring under localized operational conditions," *Ocean Eng.*, vol. 235, no. June, p. 109392, Sep. 2021, doi: 10.1016/j.oceaneng.2021.109392.

[12] M. Taghavi and L. P. Perera, "Data Driven Digital Twin Applications Towards Green Ship Operations," in *Volume 5A: Ocean Engineering*, Jun. 2022. doi: 10.1115/OMAE2022-78775.

[13] P. Johnston and R. Harris, "The Boeing 737 MAX Saga: Lessons for Software Organizations," *Softw. Qual. Prof.*, vol. 21, no. 3, pp. 4–12, 2019, [Online]. Available: www.asq.org

[14] "Trust," "Meriam-Webster.com," 2023. https://www.merriam-webster.com

[15] S. Vadlamudi, "Enabling Trustworthiness in Artificial Intelligence - A Detailed Discussion," *Eng. Int.*, vol. 3, no. 2, pp. 105–114, 2015, doi: 10.18034/ei.v3i2.519.

[16] S. Jain, M. Luthra, S. Sharma, and M. Fatima, "Trustworthiness of Artificial Intelligence," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Mar. 2020, pp. 907–912. doi: 10.1109/ICACCS48705.2020.9074237.

[17] B. Shneiderman, "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy," *Int. J. Human–Computer Interact.*, vol. 36, no. 6, pp. 495–504, Apr. 2020, doi: 10.1080/10447318.2020.1741118.

[18] J. Haspiel *et al.*, "Explanations and Expectations," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, Mar. 2018, pp. 119–120. doi: 10.1145/3173386.3177057.

[19] M. Dikmen and C. Burns, "Trust in autonomous vehicles: The case of Tesla Autopilot and Summon," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct. 2017, pp. 1093–1098. doi: 10.1109/SMC.2017.8122757.

[20] P. Langley, B. Meadows, M. Sridharan, and D. Choi, "Explainable Agency for Intelligent Autonomous Systems," *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 2, pp. 4762–4763, Feb. 2017, doi: 10.1609/aaai.v31i2.19108.

[21] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[22] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, Sep. 2019, doi: 10.1145/3236009.

[23] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[24] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[25] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognit.*, vol. 65, no. May 2016, pp. 211–222, May 2017, doi: 10.1016/j.patcog.2016.11.008.

[26] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," 2014, pp. 818–833. doi: 10.1007/978-3-319-10590-1_53.

[27] B. Murray and L. P. Perera, "A Data-Driven Approach to Vessel Trajectory Prediction for Safe Autonomous Ship Operations," in *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*, Sep. 2018, pp. 240–247. doi: 10.1109/ICDIM.2018.8847003.

[28] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, Jul. 2021, doi: 10.1145/3457607.

[29] B. Richardson and J. E. Gilbert, "A Framework for Fairness: A Systematic Review of Existing Fair AI Solutions," vol. 1, pp. 1–28, Dec. 2021, [Online]. Available: http://arxiv.org/abs/2112.05700

[30] K. Alikhademi, E. Drobina, D. Prioleau, B. Richardson, D. Purves, and J. E. Gilbert, "A review of predictive policing from the perspective of fairness," *Artif. Intell. Law*, vol. 30, no. 1, pp. 1–17, Mar. 2022, doi: 10.1007/s10506-021-09286-4.

[31] Z. Chen, J. M. Zhang, M. Hort, F. Sarro, and M. Harman, "Fairness Testing: A Comprehensive Survey and Analysis of Trends," pp. 1–27, Jul. 2022, [Online]. Available: http://arxiv.org/abs/2207.10223

[32] E. Soremekun, M. Papadakis, M. Cordy, and Y. Le Traon, "Software Fairness: An Analysis and Survey," May 2022, [Online]. Available: http://arxiv.org/abs/2205.08809

[33] M. A. Madaio, L. Stark, J. Wortman Vaughan, and H. Wallach, "Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Apr. 2020, pp. 1–14. doi: 10.1145/3313831.3376445.

[34] A. Agarwal and H. Agarwal, "A Seven-Layer Model for Standardising AI Fairness Assessment," *arXiv Prepr. arXiv2212.11207*, Dec. 2022, doi: https://doi.org/10.48550/arXiv.2212.11207.

[35] M. Busuioc, "Accountable Artificial Intelligence: Holding Algorithms to Account," *Public Adm. Rev.*, vol. 81, no. 5, pp. 825–836, Sep. 2021, doi: 10.1111/puar.13293.

[36] V. Dignum, *Responsible Artificial Intelligence.*, vol. 51. 2019.

[37] P. Vassilakopoulou, "Sociotechnical Approach for Accountability by Design in AI Systems," *ECIS 2020*, pp. 6–15, 2020, [Online]. Available: https://aisel.aisnet.org/ecis2020_rip/12

[38] Y. Wang, M. Xiong, and H. Olya, "Toward an Understanding of Responsible Artificial Intelligence Practices," in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2020, vol. 2020-Janua, pp. 4962–4971. doi: 10.24251/HICSS.2020.610.

[39] L. Cheng, K. R. Varshney, and H. Liu, "Socially Responsible AI Algorithms: Issues, Purposes, and Challenges," *J. Artif. Intell. Res.*, vol. 71, pp. 1137–1181, Aug. 2021, doi: 10.1613/jair.1.12814.

[40] G. Karmakar, A. Chowdhury, R. Das, J. Kamruzzaman, and S. Islam, "Assessing Trust Level of a Driverless Car Using Deep Learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4457–4466, Jul. 2021, doi: 10.1109/TITS.2021.3059261.