Assuring Safe Implementation of Decision Support Functionality based on Data-driven Methods for Ship Navigation

Andreas Brandsæter

Group Technology & Research, DNV GL, Norway. E-mail: andreas.brandsaeter@dnvgl.com

Georg Smefjell

Statutory Support, DNV GL, Norway. E-mail: Georg.Smefjell@dnvgl.com

Koen van de Merwe

Risk Management Advisory, DNV GL, Norway. E-mail: koen.van.de.merwe@dnvgl.com

Vegard Kamsvåg

Group Technology & Research, DNV GL, Norway. E-mail: vegard.kamsvag@dnvgl.com

The rapid technology development related to machine learning and data-driven models for autonomous and unmanned vessels continues. Also manned vessels can make use of this technology, for example to enhance situational awareness of an on board navigator. Potentially, this can contribute to increase safety and to optimize operations by transferring tasks and functions to where they are most effectively handled, ashore and on board. However, the introduction of decision support systems and functionality to enhance situational awareness can have detrimental consequences, due to for example misunderstandings, wrong use of the functionality, malfunctioning user-interface, as well as bad or wrong decision proposals. This can be the case, even when manning levels are kept unchanged. To ensure safety, we argue that the system must be rigorously tested, and the system's limitations, uncertainties and capabilities must be correctly conveyed to its users. Based on current regulations, including the International Maritime Organization (IMO) resolution *Principles of minimum safe manning*, we investigate how minimum safe manning of a vessel should be established considering relevant factors, including the ship's level of automation and shore support. We also discuss challenges related to lack of specification, which is an inherent challenge to decision support systems based on object detection and image classification since these tasks rely on perception of the environment, which can only partially be specified using rules. Furthermore, challenges related to lack of explainability are discussed, and potential benefits of using methods for black-box explanation during operation and during testing are investigated. We emphasize the importance of testing and verification of the dataset used to train the models, ensuring that it sufficiently covers relevant scenarios. We also discuss challenges related to human factors, and emphasize the importance of safety management systems used to identify risks, responsibilities, resources and competencies ensuring compliance with rules and regulations.

Keywords: Situational Awareness, Decision support, Autonomous ships, Machine Learning, Human Machine Interface, Safe Manning, Assuring data-driven methods, Training data analysis, Explainable AI.

1. Introduction

The maritime industry has a long history seeking to optimize operations, from sailing vessels to steam ships, to the technologically sophisticated vessels of today. The rapid technology development related to machine learning and data-driven models for autonomous and unmanned vessels continues. Changes and transformation will be the new normal, and companies who are able to utilize this technology can have major advantages through optimizing operations ashore and on board. Also manned vessels can make use of this technology, for example to enhance situational awareness of an on board navigator. This implementation also facilitates shifting tasks and functions between the ship and shore, having them performed where it is more effective. Potentially, this can contribute to increase maritime safety. For example, in the DNV GL led ROMAS project (**?**), the small, but important, task of keeping track of the number of passengers on a ferry was delegated from the ship to a shore control center, which would, in case of an emergency, communicate with appropriate emergency response and rescue teams. By letting the on-board crew concentrate on other pressing issues, transferring this small but

Proceedings of the 30th European Safety and Reliability Conference and
the 15th Probabilistic Safety Assessment and Management Conference
Edited by Piero Baraldi, Francesco Di Maio and Enrico Zio
Copyright © ESREL2020-PSAM15 Organizers.*Published by* Research Publishing, Singapore.
ISBN: 978-981-14-8593-0; doi:10.3850/978-981-14-8593-0

critical task to shore, can contribute to increase safety. When tasks and functions are moved between shore and ship, this can affect manning levels, and in Section 2 of this paper, we review current rules and regulations related to safe manning.

Several accidents are caused, or partly caused, by inadequate situational awareness. More extensive use of decision support systems and functionality to enhance the navigator's situational awareness can therefore prove useful and contribute to increased safety. In their investigation of the collision between the two manned vessels, the Norwegian frigate HNnMS Helge Ingstad and the oil tanker Sola TS, in 2018, the **?** explains that once people have established a mental model of the situation, they tend to seek cues from the environment that confirm rather than reject the model, without being aware of this (often referred to as confirmation bias). Hence, if the system conveys a wrong or incomplete representation of the situation (e.g. misclassifies a ship) or suggests dangerous maneuvering actions, implementation and use can increase risk and lead to serious accidents, even on manned ships. We discuss this further in section 3, where we also describe challenges related to the interface between humans and machine (HMI), particularly regarding transfer of functions to shore control centers.

Because the introduction of decision support systems and functionality to enhance situational awareness can have detrimental consequences, even on manned ships, we argue that thorough testing and verification of decision support tools and situation displays are needed. In Section 4, we discuss how the use and implementation of data driven models introduce new challenges to assurance and verification. Important challenges including lack of specification and lack of interpretation are discussed. In Section 5, we propose an approach to assurance based on the pursuit of a capability assessment based on claims and supporting evidence. Furthermore, we emphasize the importance of reproducibility and appropriate cross validation, we discuss how training data from different sources such as early development, test track and operation can be utilized, and illustrate how apparently insignificant small changes in the input data can result in missed detection. Finally, in Section 6, we provide some concluding remarks.

2. Existing Regulation

The **?** places the responsibility for safe operation of a ship and pollution prevention on the owner, or an organization or person such as the manager, or the bareboat charterer, who has assumed the responsibility for operation of the ship. The legal entity holding the responsibilities is named company and DOC-holder (coming from the required Document of Compliance which all companies operating vessels of size 500 GT and above, in international trade, must hold).

It is important to remember that autonomous and remote supported vessels will, for the foreseeable future, be operated together with vessels operating in more traditional modes. We have noted some interest in developing a new quality management code or revising the ISM code specifically targeting autonomous vessels and shore control centers. This is being discussed as a part of the ongoing Scoping Exercise in the International Maritime Organization (IMO). As new operational modes will be used together with conventional ships, and the ISM code is goal based, we advocate keeping the existing ISM code and its solutions and requirements.

2.1. *Safety management objectives*

In accordance with ISM code 1.2.2, the Safety management objectives of the company should, inter alia:

(1) provide for safe practices in ship operation and a safe working environment;
(2) assess all identified risks to its ships, personnel and the environment and establish appropriate safeguards; and
(3) continuously improve safety management skills of personnel ashore and aboard ships, including preparing for emergencies related both to safety and environmental protection.

Furthermore, according to paragraph 1.2.3, the safety and management system should ensure:

(1) compliance with mandatory rules and regulations; and
(2) that applicable codes, guidelines and standards recommended by the Organization, Administrations, classification societies and maritime industry organizations are taken into account.

2.2. *Duties and responsibilities of the Company*

We expect that the duties and responsibility of the DOC-holder will be maintained. Accordingly, the safety management systems of DOC-holders must be revised to ensure compliance also when introducing and utilizing new technologies and new operational methodologies and forms of support. In resolution A.1118(30) *Revised Guidelines on the Implementation of the International Safety Management (ISM) code by Administrations*, the IMO is asking its member states to enable a company to ”develop solutions which best suit that particular company”. We expect this goal to be maintained for vessels and companies operated in novel ways. Furthermore, we expect that companies will continue to be required to have Documents of Compliance, and that ships (500 GT and above) will have to have Safety Management Certificates. Additionally, any manned vessel will continue to have to comply with the Maritime Labour Convention (as per its scope). It is further expected that flag states and/or their recognized organizations will, as requested in A.1118(30), "ensure that assessments are based on determining the effectiveness of the safety management system in meeting the objectives" of the code.

2.3. *Proposal for minimum safe manning*

A proposal for the minimum safe manning shall be prepared by the Company (shipowner, or the management company) and submitted to the Flag State Administration for their approval/acceptance and consequently issuance of the safe manning document. Internationally, manning levels on board vessels are regulated by IMO resolution A.1047(27) *Principles of minimum safe manning*. Additionally, there may be some national requirements as the manning is set by the competent authority in the flag states. The objective of the resolution is to ensure that a vessel is sufficiently, effectively and efficiently manned to provide safety and security of the vessel. This includes safe navigation and operation at sea, safe operations in port, prevention of human injury or loss of life, the avoidance of damage to the marine environment and to property, and the welfare and health of seafarers through the avoidance of fatigue. These are goals which are overlapping with the ISM code and accordingly the DOC-holders must have measures in place in the safety management system to ensure ongoing compliance and continuous improvement.

According to resolution A.1047(27), minimum safe manning of a vessel should be established taking into account all relevant factors, including:

(1) size and type of vessel;
(2) number, size and type of main propulsion units and auxiliaries;
(3) level of automation;
(4) construction and equipment of the vessel;
(5) method of maintenance used;
(6) cargo to be carried;
(7) frequency of port calls, length and nature of voyages to be undertaken;
(8) trading areas, waters and operations in which the vessel is involved;
(9) extent to which training activities are conducted on board;
(10) degree of shoreside support provided to the vessel by the company;
(11) applicable work hour limits and/or rest requirements; and
(12) the provision of the approved Ship's Security Plan

In most countries manning is set through an application where the DOC-holder documents how they are operating their vessels, i.e. answering the underpinning questions of how the 12 elements mentioned above are impacting operations and how their measures are providing the necessary safety level. It is expected that a combination of effective management of new technologies, keeping the ISM code as is but revising the measures in the safety management systems and a systematic handling and documentation in relation to IMO Resolution A.1047(27) will be necessary to get support and acceptance of optimizing solutions. Utilizing the safety management system as a mechanism for reaching goals, ensuring compliance, understanding and mitigating risks and ensuring resources needed for operations is necessary.

3. Human Factors Engineering

Automation has proven useful in many applications. However, due to the fact that automation is inherently complex and difficult to understand, many challenges arise in the interface between humans and machines, leading to catastrophic failures (**??**). Hence, the human-machine interface (HMI) should be carefully designed and tested. Human operators play an important role in complex technological systems due to their flexibility and ability to learn and adapt to unexpected situations (**?**). However, monitoring the output of automated functions have proven to be challenging for humans because of a lack of situation awareness resulting in the out-of-the-loop performance problem (**??**). In addition, the operator may not adequately understand the inner workings of the automation leading to degraded performance (**???**).

In line with IMO A.1047(27) *Principles of safe manning*, safe operation can be demonstrated by ship owners utilizing on shore control centers (SCC) which are supporting vessels through advanced automation. It is important to note that although automation is frequently implemented with the goal of reducing manual workload, **?** argue that automation can sometimes increase workload of a pilot during already high workload periods, such as when a route change is needed or when a problem occurs. Also, **?** argue that automation of industrial processes may expand rather than eliminate problems with the human operator.

4. Data-Driven Models Introduce New Challenges to Assurance

As new technology is taken into consideration when establishing levels for minimum safe manning, assurance and verification of the new technology and its reliability and robustness is required. But data-driven models introduce new challenges to assurance. **?** argue that the safety standards available within the automotive industry and any other industry, including ISO 26262 *Road Vehicles - Product safety* and ISO/PAS 21448 *Road vehicles - Safety of the intended functionality*, have been defined without explicitly considering the specifics of machine learning algorithms and data-driven models. **?** summarise *Part 6 - Product development at the software level* of ISO 26262 as a specification of the process requirements for the level of rigour needed in developing the software for a function. Algorithms for machine perception and situational awareness are usually partly or fully based on machine learning algorithms whose functional reasoning are challenging or even impossible to understand and predict (**?**). Hence, the verification of such a system needs to be fundamentally different from a traditional verification process based on physical understanding. The machine learning algorithms are data driven, and completely dependent on the quality of the training data. Effective verification will therefore likely need to be carried out by a combination of testing, simulations and benchmarking against real and synthetic datasets.

Both traditional programmed software and machine learning software exhibits some error rate. Our focus is on documenting error, and reducing the number of errors and their consequences to an acceptable level. **?** list lack of specification and non-interpretability as the two main obstacles to safety assurance of machine learning algorithms and data-driven models, in addition to dataset collection and its requirements, and handling uncertainty.

4.1. *Lack of specification*

The lack of specification is an important challenge when testing and verifying a model, especially for use in safety critical domains. A training set is necessarily incomplete, and it is not possible to guarantee that it is even representative of the space of possible inputs (**?**). For example, machine perception is a functionality which is not completely specified. What is for example the specification for recognizing a kayak? Problems which involve advanced functionality that are not completely specifiable has motivated the implementation of machine learning based software which learns from examples rather than being programmed from a specification (**??**). Based on experimental data reviewed, **?** argue that human categorization is also dependent on stored exemplars, in addition to abstracted rules.

4.2. *Lack of interpretations*

? claims that although interpretability is often suggested as a remedy, few articulate precisely what interpretability means or why it is important. The paper discusses the interpretability of human decision-makers, and what notion of interpretability these explanations satisfy, and argues that human explanations seem unlikely to clarify the mechanisms or the precise algorithms by which brains work. Nevertheless, the information conferred by an interpretation may be useful. In the context of machine learning, **?** defines interpretability as ”the ability to explain or to present in understandable terms to a human”.

? distinguishes between *interpretability*, that is how well a human can understand the decisions in a given context, and *explanations* of specific decisions. Similiarly, **?** distinguish between *trusting a model*, that is whether the user trusts a model to behave in reasonable ways if deployed, and *trusting a prediction*, that is whether a user trusts an individual prediction. However, **?** points out that by explaining multiple (individual) predictions, the global model is also interpreted and trust in the model can be achieved.

Several methods are proposed and developed to interpret black-box models and explain their predictions. Some of these methods are model-specific, that is, they can only be used on a subset of machine learning models, while other methods are model-agnostic. If a task should be solved with machine learning methods, typically, several types of machine learning models are evaluated, and when comparing models in terms of interpretability, it is easier to work with model-agnostic explanations (**?**).

A popular and frequently used model-agnostic approach to interpret and explain the decisions and predictions is feature importance. For a linear regression model, the importance of different features is readily available (if independence between the features can be assumed), and various methods aim to provide a similar interpretation of more complex models. Several methods are available, including perturbation methods (**??**), local surrogate models such as LIME (**?**), Shapley values (**???**), case-based explanations (**?**) and counterfactual explanations (**?**). Since the predictions made by the data-driven methods rely heavily on the training data used, we also advocate explanations which convey how the training data affects the predictions. This includes case-based explanation methods which select particular points of the dataset to explain the behaviour of machine learning models (**?**), and influence functions which tell us how the model parameters change when a point in the training dataset is up-weighted by an infinitesimal amount (**?**). **?** propose a method based on Shapley values. The Shapley value concept originates from coalitional game theory, developed to fairly distribute the payout among a set of cooperating players. This is extended to subset importance, such that a prediction is explained by treating the subsets of the training data as players in a game where the predictions are the payouts.

Due to their subjective nature, it is challenging to quantify and rate the quality of different interpretations and explanations (**?**). A possible approach to test the quality of an explanation, is to use human subject evaluation, assuming that good model explanations are consistent with explanations from humans who understand the model (**?**). One can sometimes also test if explanations can guide users to select the best predictor or classifier, or to improve it (**?**).

5. Capability Assessment - Claims and Evidence

According to the International Maritime Organization's guidelines for the approval of alternatives and equivalents (**?**), the approval of an alternative and/or equivalent design can be performed by comparing the alternative design to existing designs to demonstrate that the design has an equivalent level of safety. Hence, the approval of autonomous systems used in shipping, including decision support systems and functionality to enhance situational awareness, can be based on the equivalence principle: the use of the novel functionality must make the operation safer or at least as safe as the conventional operation. However, measuring and testing equivalence is challenging, and deciding concrete acceptance criteria is difficult.

One approach is to pursue a capability assessment based on claims and supporting evidence. Each claim needs a stringent specification. For example, a claim regarding detection of kayaks should include detection distance capabilies, detection rates, the kayaks size, color and shape, external conditions such as weather, waves and lighting, etc. If all tests are performed on red kayaks, it is not necessarily true that the results are valid for blue kayaks. On the other hand, the more general a claim is, the better, assuming that the claim is confidently supported by evidence. Evidence can be based on real-world data from testing, but data from simulations can also contribute.

5.1. *The importance of reproducibility*

? argue that ”reproducibility is a minimum necessary condition for a finding to be believable and informative”. Here reproducibility is defined as ”the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator.” However, although a movement to examine and enhance the reliability of research expands (**?**), the basic terms, including reproducibility, are not standardized . **?** propose to use three different terms to clarify the meaning of reproducibility:

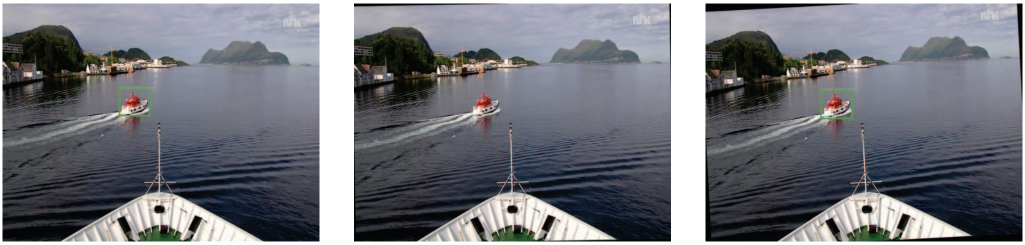
(1) *methods reproducibility*, the ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results;
(2) *results reproducibility*, the ability to produce corroborating results in a new study, having followed the same experimental methods;
(3) *inferential reproducibility*, the ability to make knowledge claims of similar strength from a study replication or reanalysis.

We believe that the capability assessment of a decision support system (or sub-system) should include claims which are reproducible in all three definitions proposed by **?**. This ensures that qualitatively similar claims can be supported by evidence either from a reanalysis of the original test results or from an independent replication of the tests and experiments executed to gather the supporting evidence. When a claim is adequately supported by evidence, a different team of approval engineers, should, on multiple trials, come to the same conclusions regarding the claim even when they use different datasets collected at a different time and place.

5.2. *Analyzing the dataset*

Since the outputs of data-driven models rely heavily on the data used for training, careful and rigorous analysis of the training data can be an essential part of assurance. Statistical distribution should be considered for the classes and the data attributes within each class, defining the class itself as well as environmental attributes that may be encountered within the operational design domain. The dataset should capture all aspects of the future operation, such as types of objects and environmental conditions. **?** state that the dataset should be ”highly representative and complete, particularly regarding corner case inputs”. This is desirable, however, it is difficult to ensure and assure that this is fulfilled.

For testing, one option is to use data gathered by the system developer/owner for compatibility and validity purposes. Using data gathered by the actual system under test ensures that the data is in the correct format, and of the same quality as it would be in operation. Furthermore, using data gathered by the actual sensors used in the system enables testing of the entire processing chain, including its physical elements related to the actual sensors and their installation on the given ship. This also enables implicit testing of the accuracy of sensor calibration, noise properties and low-level signal processing. Additionally, the sensor fusion module depends on accurate external calibration between sensors which is unique for each particular configuration, and might not be able to function properly using separately maintained generic validation datasets. In the following, we investigate utilizing three different sources of information.



(a) Rotation: *0 degrees* (b) Rotation: *- 1 degree* (c) Rotation: *- 3 degrees*

Fig. 1.: A classifier successfully detects and draws a bounding box around the vessel in Fig. (a) and (c). In Fig. (b), the classifier fails to detect the vessel.

5.2.1. *Data used for development and training*

The system developer gathers data during their development and training phase. The performance of data-driven algorithms, such as convolutional neural networks (CNNs), for object detection are reliant of the properties and distribution of this dataset. The system developer typically perform cross-validation (e.g., k-fold cross-validation), and reports the final results. We emphasize the importance of ensuring independence between the validation sets.

5.2.2. *Data from operation*

Data is also gathered by the vendor within the specified operational design domain over a period of time, ensuring that a significant number of different scenarios and object types are captured. This dataset is not used for training machine learning algorithms and may be captured after deployment in a trial phase. This dataset does not usually contain labels or ground truth information, used to validate the correct output from the detectors. However, it is assumed that the data can still be used for investigating robustness and consistency in varying conditions and situations.

5.2.3. *Data from Unrehearsed test track*

We also investigate how tests can be executed on a test track, designed by the approving body, and unknown to the system developer. This is analogous to the sea trial, traditionally performed when approving a new-build. A number of objects will be present during the test track, and these should be equipped with sensors and positioning equipment such that their position and speed is known to a high degree of certainty. The design of the test track will depend on the operational design domain for the particular system under test, as well as the properties of the dataset used for training any machine learning algorithms.

5.3. *Cross-validation*

As described in **?**, it is well known that when we evaluate predictions from a statistical model on the dataset used to train the model, our accuracy estimates tend to be overoptimistic (**?**). To maximize the utilization of the data, and at the same time avoid overfitting to the test data, cross-validation techniques can be applied. Cross-validation introduces various methods of repetitively splitting the data into exclusive parts, where one part is used to train the model, and the other is reserved for testing. Dependency between the training and test dataset can result in overly optimistic estimates of model performance (**?**). **?** argue that a similar situation can occur when there are dependence structures in the data. If the test data are drawn nearby in the dependency structure, the independence between the training and test data can be compromised. Hence, ensuring independence between the two datasets is essential. A range of different splitting techniques can be used, providing different cross-validation estimates, see for example (**??**).

5.4. *Invariants*

To furhter maximize the utilization of the available data, it is often useful to define a function's invariants, that is the ways the input can change without affecting the output (**?**). For example, classification of a vessel in an image should be invariant to translations, meaning that the classification does not change if the vessel is moved to another location in the image. Similarly, small rotations of an image should not affect the classification. Fig. 1 (a) shows an example where a Region-based Fully Convolutional Network (R-FCN) (**?**) successfully detects and classifies a vessel. In Fig. 1 (b), the image is rotated slightly (-1 degree) and the classifier fails to detect the vessel. Interestingly, when the image is further rotated (-3 degrees), as shown in Fig. (c), the vessel is again successfully detected. The classifiers ability to succeed on further rotations clearly illustrates that the classifier is highly unpredictable. This demonstrates that rigorous testing is needed, and that it is not sufficient to test extremes.

6. Conclusion

In this paper, we discuss how the introduction of new technology and functionality for situational awareness and decision support affect the requirements for safe manning. We argue that errors or inadequate robustness in the decision support functionality can have severe consequences, even on manned ships. Hence, meticulous and thorough testing and verification should be required.

Once the capabilities and limitations of the decision support functionality is sufficiently tested and documented, the functionality should be taken into consideration when establishing the minimum safe manning of a vessel, according to the International Safety Management code (ISM code) for setting manning and the International Maritime Organization's (IMO's) resolution focusing on effective safety management systems solutions. This allows tasks and functions to be executed where it is most effective, and the ISM code and the safety management systems place necessary responsibilities, duties and measures.

We investigate challenges related to assurance and quality assessment of functionality based on data driven methods, and argue that both the lack of specifications and the lack of interpretations associated with data-driven methods makes the verification of such a system challenging and fundamentally different from traditional verification processes. Since the machine learning algorithms are completely dependent on the quality of the training data, rigorous analysis of the training data is needed. Moreover, we discuss how different sources of data should be utilized in development and testing. We also illustrate how we can optimize our utilization of the available data by considering invariants, such as rotated images, showing that rigorous testing is needed, and testing of extremes is not always sufficient.

Acknowledgement

The authors thank Kjetil Kåsamoen in DNV GL Statutory Support, and Øystein Engelhartsen and Knut Erik Knutsen in DNV GL Group Technology & Research for valuable comments and interesting discussions regarding the topics of this paper.