Towards Simulation-based Verification of Autonomous Navigation Systems

Tom Arne Pedersen[1], Jon Arne Glomsrud[1], Else-Line Ruud[2], Aleksander Simonsen[2], Jarle Sandrib[2] and Bjørn-Olav Holtung Eriksen[3]

[1] DNV GL, Veritasveien 1, 1363 Høvik

[2] FFI, Instituttveien 20, 2007 Kjeller

[3] Norwegian University of Science and Technology (NTNU), Høyskoleringen 1, 7491 Trondheim

E-mail: tom.arne.pedersen@dnvgl.com, jarle.sandrib@ffi.no, bjorn-olav.holtung.eriksen@ntnu.no

Abstract

Autonomous ships are expected to change water-based transport of both cargo and people, and large investments are being made internationally. There are many reasons for such transformation and interest, including shifting transport of goods from road to sea, reducing ship manning costs, reduced dangerous exposure for crew, and reduced environmental impact.

Situational awareness (SA) systems and Autonomous navigation systems (ANS) are key elements of autonomous ships. Safe deployment of ANS will not be feasible based on real-life testing only, but will require large-scale, systematic simulation-based testing in addition to assurance of the development process.

DNV GL proposes to use a *digital twin*, meaning a digital representation of key elements of the autonomous ship as a key tool for simulation-based testing. The digital twin contains comprehensive mathematical models of the ship and its equipment, including all sensors and actuators. The complete simulation-based test system complementing the digital twin should consist of a *virtual world* to simulate environmental conditions, geographical information and interaction with other maritime traffic. Finally, the test system must include a *test management system* that controls simulation of the digital twin and the virtual world, generates test scenarios as well as evaluates the test scenario results. An automatic scenario generation tool should search for low ANS performance, and ultimately establish sufficient coverage of the possible scenario space. The test scenario evaluation should automatically consider safety, conformance to collision regulations at sea (COLREG), and possibly also the efficiency of the ship navigation.

This paper presents a comprehensive prototype of a test system for ANS. Key topics are simulation-based testing, interfacing the simulator and ANS, cooperation with ANS manufacturers, dynamic test scenario generation, automatic assessment towards COLREG and experiences from the cooperation with The Norwegian Defense Establishment (FFI).

Keywords: Autonomous navigation, digital twin, simulation-based testing, dynamic test scenario, automatic test scenario generation

1. Introduction

Ships have always been operated by seafarers, with a crew size depending on the ship size, type and mission. In recent years, substantial development has been achieved in sensor technology, machine learning, automation and connectivity. This means that, at least in theory, it may be possible to reduce or even remove the crew from the ship. However, this will require either shore-based remotely monitored and operated ship systems, or autonomously operated ship systems based on algorithms.

Remotely or autonomously controlled functions are not necessarily implemented to reduce cost only, but also for safety reasons. 80% to 96% ([1] [2]) of marine accidents are caused by human errors, and 56% of these are related to one or more COLREG rule violation [3]. However, not all errors would have been avoided by introducing unmanned and autonomous ships. Wróbel et al. [4] investigated 100 accidents to analyse if the accidents would have been avoided if one or both ships were fully autonomous and they concluded that navigational related accidents could at least be reduced. The possibility of using autonomously and remotely operated ships are also introducing novel or changed transport systems and business models where e.g. smaller unmanned vessels can be used the last mile bringing cargo from a mother ship to smaller less area-demanding harbours.

In the maritime industry, autonomous ships are on everyone's lips, but what this entails can vary widely. Several definitions of the level of autonomy exists, commonly defined as a system's increasing ability to operate without human control or intervention. The scale ranges from no autonomy where the human operator needs to take all decisions, to fully autonomous without a human operator in the loop. Autonomous does not equal unmanned and many levels of autonomy do not contain this aspect. Figure 1 maps different levels of autonomy (vertical axis) against level

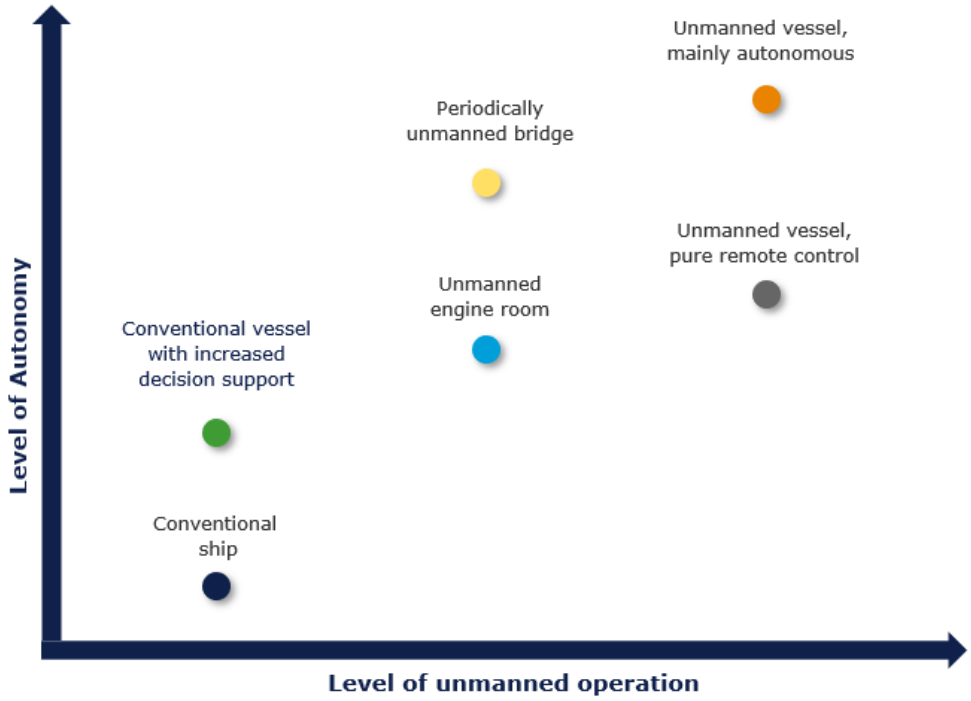


Figure 1: Level of Autonomy vs level of unmanned operation

of unmanned operation (horizontal axis). Conventional ships are placed in the lower left corner, with a low level of autonomy and unmanned operation. Ships with added decision support have a higher level of autonomy and are thus placed higher on the left part of the figure, though still with a low level of unmanned operation. Presently, the engine rooms onboard ships are unmanned certain time periods, and it is required that the engine room can operate at least 24 hours without manual monitoring and control. The engine room operator, however, must be onboard the ship. Unmanned engine rooms, found in the centre of Figure 1, indicates that the engine room can operate without manual control from onboard crew for weeks or months, and that the control and monitoring is done from an on-shore control site. A periodically unmanned bridge is also placed in the middle of Figure 1. At the right part of Figure 1, unmanned ships that are either remotely controlled or autonomous are found. One may notice that a typical ship is not either conventional or autonomous/remote operated, but instead some ship systems may be unmanned, while others are not.

To navigate safely, either the ship crew or the ANS needs to detect any elements that may affect the planned path of the ship. In Figure 2, the ship navigation function is broken down into sub tasks.

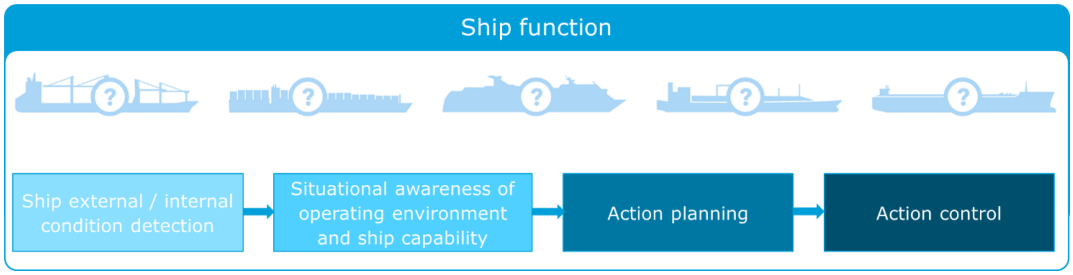


Figure 2: Ship functions broken down into sub tasks. Based on [5]

Initially, the ship navigator needs to know the external and internal operational and ship conditions, such as geography, bathymetry, fixed or floating objects, and weather conditions together with the conditions of the ship and its equipment. A priori information may come from e.g. Electronic Navigation Charts (ENC), Automatic Identification System (AIS), etc. Not all ships transmit AIS data and not all AIS data are reliable, thus additional exteroceptive sensors such as radar, camera, infrared camera and lidar need to be used to detect objects relevant for the navigation

To achieve sufficient SA, the different objects need to be classified and their states determined. Computer vision using camera is a field that has come far in detecting and classifying surrounding objects, but in the maritime industry there is still a long way to go. Computer vision is usually based on machine learning which needs to be trained using pre-existing pictures or video footage that are currently limited in the maritime industry.

Once the system has analysed the situation and SA is achieved, the course of action needs to be planned. The planning is done by the ANS using information from the predefined ship mission and the predefined set of navigation rules such as the COLREG.

The last sub task is action control. The engines, rudders or thrusters are operated to navigate the ship.

Risk is an important factor to consider while navigating. An autonomous navigation system should evaluate its performance, and if outside acceptable limits, or the risk of continuing its ongoing operation is considered too high, the ship should enter a pre-defined minimum risk condition (MRC). The MRC will vary depending on aspects such as location, type of operation and surroundings, and the resulting action may be e.g. to stop and enter DP mode, to go to nearest port, or similar.

When introducing new technology or using existing technology for new purposes, uncertainties are also introduced. The risks of safe operation in unmanned shipping have among other been studied in the MUNIN project [6]. These risks need to be adequately handled, and in [7], the safety qualification process is solved using a goal-based safety case approach. This process is based on the recommended practice for technical qualification, DNVGL RP-A203 [8]. During a qualification process, the safety goals and risks are identified, and qualification activities are then performed to collect evidence for reaching the goals and mitigating the risks.

In the perspectives of assurance and testing, it is of utmost importance to ensure that ANS algorithms are safe and do not cause accidents, meaning the ANS should go through a qualification process where testing of the actual ANS should be an important activity. Testing may be done in real life using the actual ship, in the virtual world using simulators, or in combination. Real-life testing is too time consuming and many required test scenarios will be impossible to test, thus a combination of simulation-based and real-life testing would be the preferred solution. Real-life testing could be used to validate the digital twins, digital models and simulators [9] and spot checking results from simulator tests.

In the next sections we explore simulation-based verification, unpack the components of a test verification system, the role of an open simulation platform as an important enabler, and discuss the evolution of test scenarios.

2. Simulation-based testing

Simulation-based testing will be an important tool when collecting evidence of safe ANS algorithms. A proposal for a test system, named *TestIT*, is shown in Figure 3 consisting of the test management system and a virtual world. The different parts of TestIT are explained in the following.

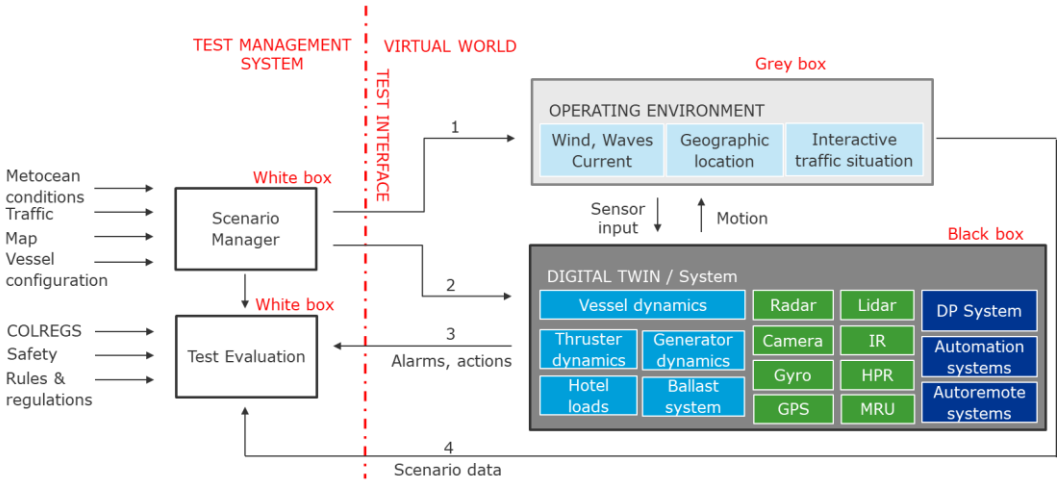


Figure 3: TestIT, a test system for autonomous navigation systems

2.1 Digital twin

The digital twin is a vital part of the test system shown in Figure 3. The digital twin is a virtual representation of a particular ship, called *own ship*, that will be controlled by the ANS under test. It is a comprehensive mathematical model of the real ship including models of the ship-specific hull dynamics, its power, propulsion, and ballast system, and sensors and actuators etc, in addition to emulated control system hardware running actual control system software. Control system software included in the digital twin may be dynamic positioning (DP) system, power management system (PMS), automation control system, see Figure 3. The different models need to be sufficiently accurate to capture relevant dynamics of the ship, and the control system should "believe" it is controlling the actual ship systems.

2.2 Operating environment

The operating environment is another vital part of the virtual world. To play out relevant and realistic test scenarios, it is important to have full control of the environmental conditions such as wind, waves and current, in addition to geographic location and interactive traffic, *target ships*. The word interactive is in this context important. If using e.g. historic AIS data recorded from ships in a specific area as basis for simulating the target ships, these will not interact with own ship, but only replay recorded AIS information. Instead, AIS data can be used as input to construct test scenarios, and then let the target ships interact both with each other and with own ship as in real life. To achieve this, also target ships need to be navigated, either by a human navigator or by other ANS algorithms. Occasionally, other ships may not behave as expected and this also needs to be handled by an ANS, thus the operating environment should include target ships not behaving in full compliance with COLREG.

2.3 Test management system

The test management system shown in Figure 3 consists of two parts. The scenario manager initiates each test episode by setting environmental conditions, traffic, location and ship parameters. The testing should focus on safety, COLREG and failure scenarios., but testing performance is possible given the digital twin is sufficiently accurate.

Test evaluation is the second part of the test management system. By using results from the simulation, ANS algorithms will be evaluated against COLREG, safety and other relevant rules and regulations. Test evaluation is discussed in more detail in Chapter 3.

2.4 Test interface

From Figure 3, one may notice the test interface between the test management system and the virtual world. It is important that the scenario manager has full control of the operating environment, arrow 1, configuring the test scenario exactly as desired. It must be possible to initiate position, course and speed of the target ships in addition to setting path plans or waypoints and decide to which degree they shall follow COLREG. Environmental disturbances and location are other important elements the scenario manager must control.

The scenario manager also needs to interface own ship controlled by the ANS algorithm under test, arrow 2. Initial position, course and speed together with path plan or waypoints need to be transferred. The scenario manager will not interfere with the ANS algorithm after initial parameters are set.

Arrow 3 in Figure 3 indicates that the test evaluation module also needs to communicate with various control systems in own ship. Any alarm, action or ship position, course or speed throughout the scenario is used for evaluation of each episode. To do a full assessment of each episode, also course, speed and position for all target ships will need to be supplied to the test evaluation module, see arrow 4 in Figure 3.

When performing simulation-based testing, the test interface must be capable of communicating all I/O between the control systems and the simulator at a rate sufficient for closed loop operation of the control system software. Normally, when performing Hardware-In-the-Loop (HIL) testing, it has been a requirement that the simulator must run in real time. In a HIL setup, the control system software is running on a Programmable Logic Controller (PLC) or similar with real time operating system. For more information on HIL, the reader can refer to [10].

When testing ANS, it will be necessary to test large numbers of traffic scenarios of relatively long duration, and if this is tested in real time, the time consumption will be high. It is desirable to reduce the total test time as much as possible without sacrificing test scope. This may be achieved either by running several simulations in parallel or run the simulation faster than real time, or a combination of these. For this to be possible, the control system software will have to run on emulated or virtual hardware, most probably in the cloud, where the simulation platform controls the simulation and computer clock cycle time.

Open Simulation Platform (OSP) [11] is a simulation platform which may potentially be used to ease the interfacing between the test management system and the virtual world with all its components. In addition, OSP will be running in the cloud, facilitating the possibility for control of the cycle time of the simulation and the virtual hardware. The platform is currently under development, and a short description is given in the following.

2.5 Open simulation Platform

OSP [11] is under development through a Joint Industry Project (JIP) with in total 24 participants, where Kongsberg, SINTEF Ocean, NTNU and DNV GL are the main partners.

The goal of the JIP is to develop a co-simulation platform to be used among ship designers, equipment and system manufacturers, yards, ship owners, operators, research institutes and academia. The co-simulation platform supports the functional mock-up interface (FMI) which is a tool independent standard to support both model exchange and co-simulation of dynamic models [12]. Supporting the FMI standard will enable the users of the OSP to develop their simulation components in their known modelling environments. These components are then compiled to functional mock-up units (FMUs), before imported to the OSP for simulation-based testing. Each vendor may add their simulation components as FMUs, making it easier to set up large simulations for a complete ship with components and control software delivered by many different suppliers.

3. Test scenario evaluation

For evaluation purposes, COLREG, safety and other rules and regulations should be used. A lot of research has been conducted for path planning algorithms where COLREG is taken into use, and some examples are [13], [14], [15], [16] and [17]. However, for evaluation of COLREG compliant ANS, not so much has been done. One of the most complete COLREG evaluation techniques was developed by Woerner [18], and both [19] and [20] were inspired by this. Woerner et al. have continued the work on COLREG compliance metrics [21] and [22] while Stankiewicz and Mullins [23] investigated both COLREG evaluation and adaptive scenario generation.

The COLREG are by purpose written such that seafarers need to use their judgement and common sense to interpret many of the rules. In order to practice good seamanship, also the autonomous ships need to follow the COLREG, and vague rules may make it difficult to design the collision avoidance systems. Moreover, different maritime environments and different types of vessels necessitate different implementations and parameters for its collision avoidance systems. For instance, what is a reasonable maneuver for a large oil tanker in open waters is very different for that of a small rib in a more confined archipelago area. This makes it challenging to design both the collision avoidance systems themselves as well as the tools for evaluating them. The COLREG contain in total 38 rules divided into 5 parts in addition to four annexes. Not all parts of COLREG is possible nor relevant for testing using simulation-based tools, and it is therefore important to clarify which of the COLREG rules that are handled by the ANS and included in the testing.

In this paper we use the terms *scenario* and *episode* with specific meanings. The term *episode* is borrowed from artificial intelligence (AI) and reinforcement learning (RL) meaning a single simulation or episode in playing a game from start until being terminated when winning/loosing. In this paper an episode is a single simulation from initial conditions until a traffic episode is terminated due to vessels having passed each other or violated COLREG or other evaluation criteria. The term *scenario* is in this paper used for a set of similar episodes, such as *crossing*, *head-on,* or, *overtaking scenario*.

In the following, two different evaluation methods are described. Woerner [18] has proposed a method where a total COLREG score, combined with safety score and penalty scores for each part of the evaluation algorithms are calculated for each encounter. Court decisions have been used for setting evaluation parameters. Another method is suggested by Nakamura and Okada [24]. By using the distance between own ship and target ships and rate of change in bearing, the authors propose a method defining *Danger area*, *Caution area* and *Safety area* for bow and stern crossing and same way situation, to evaluate different encounters. The two methods are briefly described below in Chapter 3.1 and Chapter 3.2, respectively.

3.1 COLREG score and penalties

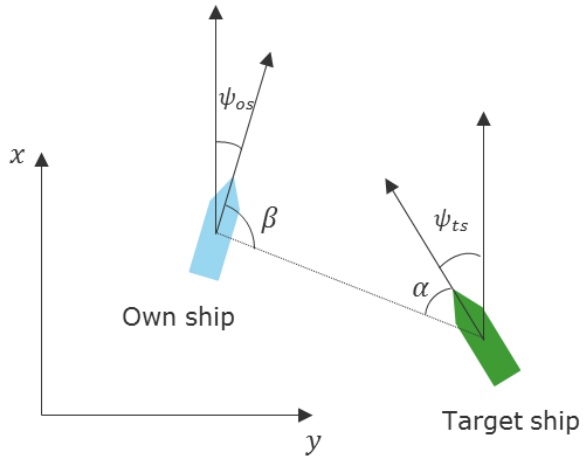


Figure 4: Pose between own ship and target ship

Pose is given by the relative bearing and contact angle and may be used for evaluating different traffic situations. The contact angle α is the angle between the heading of target ship and the straight line between own ship and the target ship seen from the target ship. The relative bearing β is the angle between the heading of own ship and the straight line between target ship and own ship seen from own ship, see Figure 4.

COLREG rule 14 is used as an example to describe the score and penalties method proposed by [18]. The rule shall prevent two ships on nearly reciprocal courses from colliding, and the rule requires a port to port passing which may be evaluated using a combination of contact angle and relative bearing at closest point of approach (CPA). CPA is defined as the point on own ship's future track where the range between own ship and target ship is at its minimum.

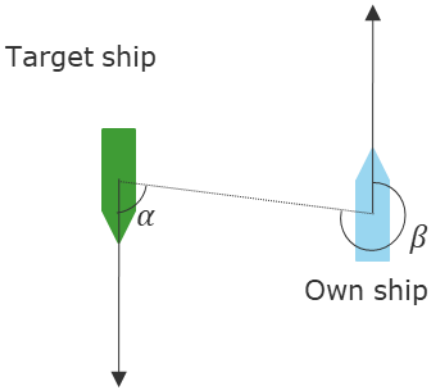


Figure 5: True port to port passing at CPA in head-on encounter

Figure 5 shows a true port to port passing at CPA, which is the preferred way of passing when in a head-on situation. A true port to port passing is achieved if $\alpha_{cpa} = -90°$ and $\beta_{cpa} = 270°$. Looking at α_{cpa}, one possible score function $S^{14}_{\alpha_{\text{cpa}}}$ is

$$S^{14}_{\alpha_{\text{cpa}}} = \left(\frac{\sin(\alpha_{cpa})-1}{2}\right)^2, \tag{1}$$

see Figure 6. The proposed score function gives maximum score at $\alpha_{cpa} = -90°$, while a starboard passing will result in 0 score.

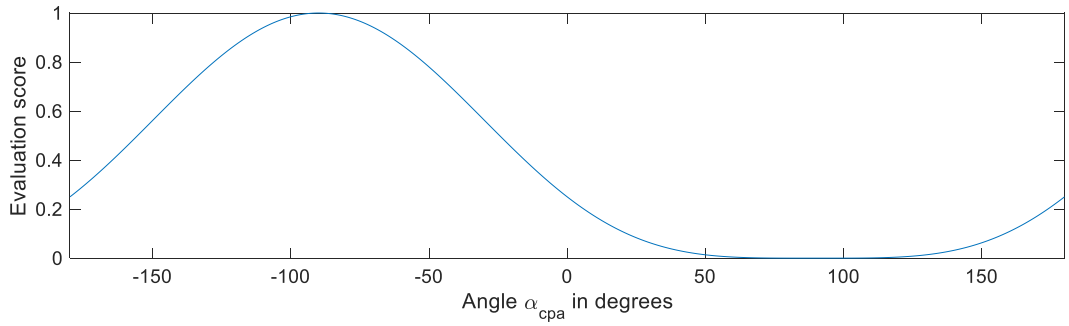


Figure 6: Plot of $S^{14}_{\alpha_{\text{cpa}}}$

Similar score function may be used for β_{cpa}, and combining them gives the following score function for a true port to port passing

$$S^{14}_{\Theta_{\text{cpa}}} = S^{14}_{\alpha_{\text{cpa}}} S^{14}_{\beta_{\text{cpa}}} = \left(\frac{\sin(\alpha_{cpa})-1}{2}\right)^2 \left(\frac{\sin(\beta_{cpa})-1}{2}\right)^2 \tag{2}$$

The penalty for evaluating the passing may then be given as

$$P^{14}_{\Theta_{\text{cpa}}} = 1 - S^{14}_{\Theta_{\text{cpa}}} = 1 - \left(\frac{\sin(\alpha_{cpa})-1}{2}\right)^2 \left(\frac{\sin(\beta_{cpa})-1}{2}\right)^2 \tag{3}$$

The penalty value goes from 0 to 1, where P = 1 indicates the highest penalty which again gives the worst score.

In a head-on encounter, the rule requires a starboard manoeuvre to be commanded. [18] did not propose a function for evaluating a non-starboard course change, therefore a new penalty function is proposed. By using the position of own ship at the time the target ship is detected, t_{0cpa}, as initial position, $\boldsymbol{p}_0$, and calculating a second position, $\boldsymbol{p}_2$ at t_2 assuming constant speed and heading, such that

$$t_2 = 100\, t_{0cpa}$$

$$\mathbf{a} = \mathbf{p_0} - \boldsymbol{p}_2 \tag{4}$$

$$\mathbf{b} = \mathbf{p} - \boldsymbol{p}_2$$

where $\boldsymbol{p}$ is the position of the own ship at any given time after the target ship has been detected, see Figure 7. If own ship for some reason is deviating from initial heading, the cross product between $\boldsymbol{a}$ and $\boldsymbol{b}$ may be used to decide if own ship has deviated to port side or starboard side of the initial course. Using this together with

$$d = \frac{\|\boldsymbol{a} \times \boldsymbol{b}\|}{\|\boldsymbol{a}\|} \tag{5}$$

where d is the distance between the position of the own ship perpendicular to the line between the points $\boldsymbol{p}_0$ and $\boldsymbol{p}_2$, the penalty function P_{nsb}^{14} may be given as

$$P_{nsb}^{14} = \begin{cases} 1 & \mathrm{d} \geq d_{threshold}\ and\ \mathrm{c} > 0 \\ 1 - \left(\dfrac{2(d_{\mathrm{threshold}} - \mathrm{d})}{d_{\mathrm{threshold}}}\right)^4 & \dfrac{d_{threshold}}{2} < \mathrm{d} < d_{threshold}\ and\ c > 0 \\ 0 & \mathrm{c} \leq\ 0 \text{ or } \mathrm{d} \leq \dfrac{d_{threshold}}{2} \end{cases}, \tag{6}$$

where c is the third element of the cross product $\boldsymbol{a} \times \boldsymbol{b}$

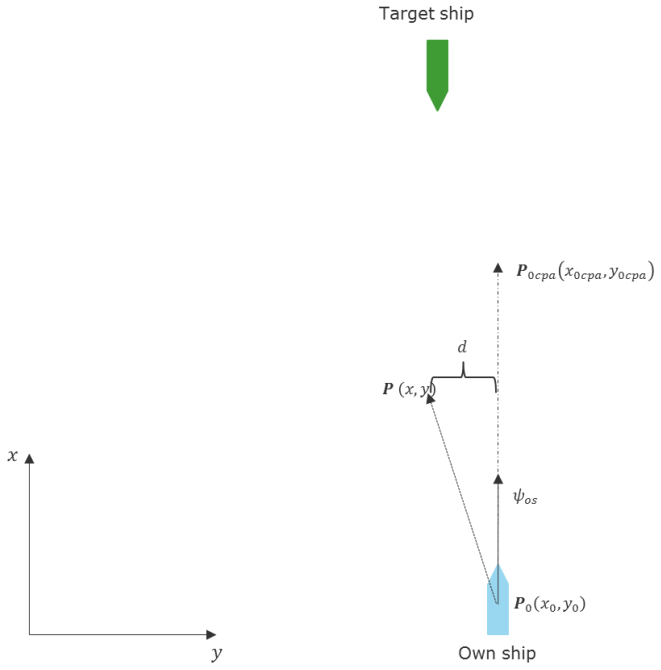


Figure 7: Head-on encounter

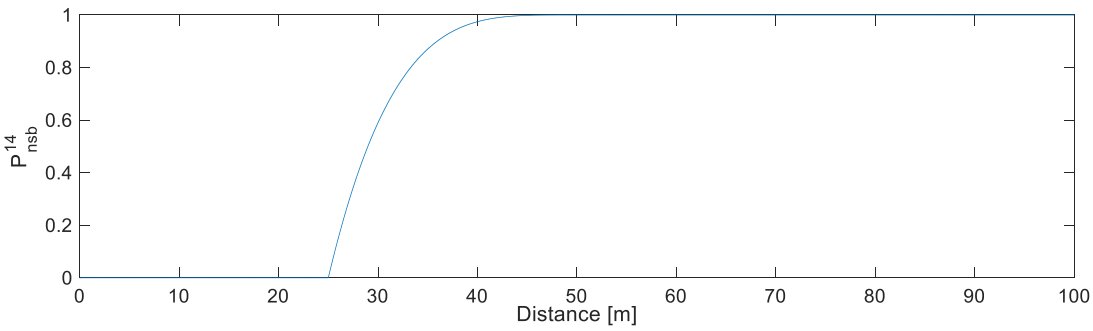


Figure 8: Penalty score for non-starboard course change

The penalty function is shown in Figure 8 for d ≥ 0 using $d_{threshold} = 50[m]$ for $c > 0$.

Woerner [18] suggested to also add penalty for delayed action to the evaluation of a head-on encounter. By using the range of the maneuver relative to the detection range and CPA range, the following penalty function may be used to evaluate delayed action:

$$P^{8}_{delay} = \left(\frac{r_{detect} - r_{maneuver}}{r_{detect} - r_{cpa}}\right) \tag{7}$$

where r_{detect} is the range of detection , $r_{maneuver}$ is the range when own ships starts the manoeuvre and r_{cpa} is the range at closest point of approach.

The total score for rule 14 is now given as

$$S^{14} = sat^{1}_{0}\left\{\left(1 - \gamma_{nsb}P^{14}_{nsb} - \gamma_{\psi_{app}}P^{8}_{\Delta\psi_{\text{app}}} - \gamma_{delay}P^{8}_{delay}\right)\left(1 - P^{14}_{\Theta_{cpa}}\right)\right\}, \tag{8}$$

where γ_{nsb}, $\gamma_{\Delta_{app}}$ and γ_{delay} are penalty coefficients which may be tuned.

3.2 Evaluation using anxiety estimation

Nakamura & Okada, [24] proposes a method using anxiety estimation for evaluating an ANS towards COLREG. They have been collecting experience data where 12 captains and pilots were participating in navigational experiments. In total 135 encounters where simulated and 30 000 data points where collected.

According to the authors, the navigators use the distance between ships, rate of change in bearing and crossing direction to recognize the risk of collisions with other ships. Due to these factors, they propose a set of evaluation diagrams, shown in Figure 9, where the diagram is divided into *Danger area*, *Caution area* and *Safety area* using distance and bearing change rate as input variables.

The evaluation is done by summing the time used in the different phases in the evaluation area diagram. The time spent in the Safety area gives no penalty, while the time spent in Caution area and Danger area is multiplied with -1 and -2 respectively for penalty calculation. The authors propose to use the following equation to calculate the evaluation score for each scenario:

$$\text{score} = \frac{\sum_{t=0}^{t_{end}} - (2 \cdot Dangerous_t + 1 \cdot Cautionary_t)}{t_{end}} \tag{9}$$

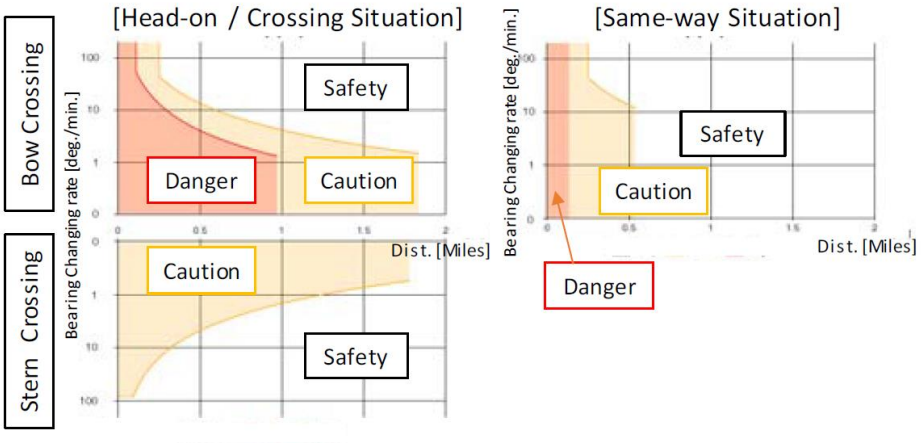


Figure 9: Evaluation area diagram [24]

The variable $Dangerous_t$ is the period/time own ship was in the Danger area during the scenario, while $Cautionary_t$ is the period/time own ship was is in the Caution area. t_{end} is the period/time of ship manoeuvring.

3.3 Final scenario assessment

A final assessment of the scenario evaluation needs to be taken by a human, but most probably it will not be feasible for a human operator to evaluate every one of the scenarios used for testing ANS, especially not when testing is done in parallel and the test setup is running faster than real-time. Instead, the idea is that the test evaluation should trigger a human assessment. If, for example, a test result is below some threshold, the human operator should check the result and approve if acceptable. However, it is important to secure that the test evaluation algorithm does not let an actual ANS failure pass without signalling the need for a manual check.

4. Automatic test scenario generation

One of the main challenges when it comes to implementing ANS, is to make the systems sufficiently safe, but what does that mean? An acceptance goal for autonomous systems to be as safe as conventional systems is challenging to prove. One important method can be to test the algorithms in traffic scenarios that best represent the potential traffic scenarios a ship might meet within e.g. 50, 100 or even 200 years of operation. This approach would in theory aim at performing exhaustive testing of all representative traffic scenarios, even though that might not be completely attainable due to the huge amount of simulations necessary. The testing therefore needs to be as time and resource efficient as possible, thus the generation (selection), simulation and evaluation of scenarios must be highly automated. Automatic test scenario generation is addressed next.

4.1 Scenario characteristics

Li et al. [25] categorized existing testing approaches for autonomous vehicles in the automotive industry into scenario-based and functionality-based testing, where scenario-based testing entails testing complete systems in environment-relevant operational situations, while functionality-based testing focus around each tested function itself and its role in the complete system. The authors argued that using either one of these methods is not enough and instead a combination of them should be used to design a more complete set of simulation-based tests for autonomous vehicles. The proposed method may be used to design tests for autonomous ships, but additional robustness testing should be included. Robustness testing should demonstrate the ability of the ANS to handle uncertainty as well as varying ship capabilities. Uncertainties arise from e.g. inaccuracies, failures, environmental conditions or noise through e.g. signals and sensors, while reduced capabilities can be caused by failures or limited resource in any critical system or component of the ship. We envisage that based on COLREG rule 6 of "Safe speed", it is a prerequisite that an ANS can safely handle uncertainty of the situational awareness as well as limitation in the manoeuvring capability. An ANS is a high level system function that rely on a large part of the underlying systems to work properly and in relation to [25], a scenario-based approach is most relevant. To conclude, the scenario characteristics appropriate for testing of ANS should focus around relevant traffic scenarios and the safe and conforming ANS behaviour according to COLREG.

4.2 Initial scenario selection and coverage validation

Critical to the confidence in an ANS, gained from the testing, is how representative and relevant the scenarios are. To our knowledge, there are no mature methods to at the highest level, automatically select scenarios that answer the representativeness and relevance needs. Therefore, scenarios need to be predefined based on operative experience and system and technical knowledge, but this does not mean we should not lean towards established test processes and test designing techniques [26]. Hazard identification and system analysis, such as STPA [27], can also be important contributors in identifying a complete scenario space and [28] used STPA to derive verification objectives and scenarios for maritime system testing.

Test coverage is a central goal for any testing and also for testing ANS. The defined scenarios and their theoretical coverage must be evaluated and the complete coverage must be validated. This depends very much upon the test problem at hand and for safe navigation we can rely heavily on the COLREG. The set of COLREG rules have evolved into what they are due to massive operative marine experience and we can confidently say COLREG at a high level are validated, and as such base our scenarios on them. Also, historic AIS-data could be used as backing evidence that the chosen scenarios are relevant and gives good coverage. In [29], a method for identifying collision risk between vessels and grounding situations using historical AIS data, has been developed.

A possible test scope for ANS could therefore initially be predefined, generic and stylistic single COLREG scenarios, with own ship and only one target ship, but where complexity

increases in the next level by introducing several target ships approaching own ship from different positions and with different headings. A third level could be introducing location and operation specific test scenarios.

4.3 Automated scenario selection

Within a predefined scenario, such as e.g. crossing scenarios with varying target ship speed and heading, automated scenario or more precisely episode search and selection can be applied. This can be considered a combinatorial problem using methods from the field of exhaustive testing [26] where various test coverage techniques might help in efficiently obtaining good coverage. Also, search or optimization techniques can be applied using the evaluation score from already performed episodes in adaptively searching for new episodes revealing inadequate behaviour. The search algorithm will gradually build an increasingly complete understanding of the evaluation response from the scenario simulation, i.e. building a model of the *response surface* [30]. This response surface model is unique to the predefined scenario and the actual ANS being under test and can only be inferred during testing. The inference of the evaluation score response surface model will gradually happen during testing, but for efficiency there are different ways of exploring the unknown parts depending on the characteristics of the actual evaluation response surface. Elements of a response surface might be described being continuous, discontinuous, asymptotic and even random and when dealing with software-based algorithms in ANS, all aspects might be the case. Therefore, efficient exploration of the unknown parts of the response surface need to handle such elements to work robustly and be able to reach critical areas without getting stuck at local points or excluding important areas. A simple method of gradient search might fail just because of discontinuous characteristics, but a more robust method might be found from Bayesian Optimization [31] and the use of Gaussian Processes (GP) regression [32]. The GP can be used to model the current scenario response surface and the current estimated mean value and variance can be computed for any point on this surface, i.e. any episode. Already simulated episodes or points on the surface will have an actual value and a zero variance since it is known, but the GP for non-simulated episodes will have a variance according to the "distance" or uncertainty to the known episodes. This knowledge can then be used to plan the next episodes to maximize some high-level test goal. For the successful testing of ANS, the main goals might be twofold: (1) efficiently find discrepancies (fail fast) and (2) establish confidence in the coverage, based on the actual coverage of the response surface. This process is at the core of Bayesian Optimization that use the strength of the currently known (i.e. the inverse of the variance) to plan the next episodes or experiments. This is only one proven method that might give efficient results for this test problem, but we envisage this field of research will continue to be developed in the future, including other optimization methods or even AI techniques.

Another efficiency aspect is related to simulation time which can be reduced by faster than real-time simulation as well as parallelization of episodes. By conducting testing of the pure software parts of the ANS and virtualisation of the ANS computing HW one can achieve both methods which can be integrated with the efficient scenario search and coverage methods earlier described.

With these testing approaches, immature ANS systems would fail fast while mature systems would fail late or eventually not fail, a strategy that can be considered agile and cost efficient. Re-testing of updated mature systems could be done using the same strategy. One could also envisage that even self-learning or adaptive ANS could be frequently or continuously tested in a similar way.

5. Simulation-based testing process

Simulation-based testing can be utilized in different phases of an ANS lifecycle process, like during development, internal testing at manufacturer, or during formal testing. In this paper, the focus is on collecting evidence in a more formal assurance or certification process. Typically, formal processes involve different parties, such as system manufacturers, ship building companies, ship owners and verification organizations. The described simulation-based test system is specifically intended as part of the formal assurance process where the key parties are the ANS manufacturer, the end user of the ship and the verification organization, a role DNV GL or other class bodies could take.

Three critical aspects of the testing process are covered in the following, namely the cooperation between the ANS manufacturer and the verification organization, the aspect of independence and finally the validation of the test results.

5.1 Manufacturer cooperation

Manufacturer cooperation is key when performing testing of an ANS. The verification organisation depends on the manufacturer to be able to:

- secure correct software used for testing,
- interface their control system,
- understand how the ANS is working,
- commission the test setup and interface, and
- validate the test results

The last bullet in the above list is of high importance. As may be seen in Chapter 3.3, the scenario evaluation should trigger a manual check of the test results by a human operator/tester in case of deviations. The human operator will then do an assessment of the scenario and flag this for follow-up if necessary. All items flagged for follow-up will then be discussed with involved parties, such as manufacturer, ship owner, class etc. If necessary, the scenario may be replayed while the manufacturer is checking their software.

Simulation-based testing does not require access to the source code of any part of the manufacturer ANS, since it is a black-box testing method, only considering the inputs and outputs of the SW-based parts. This can make it easier to cooperate with different manufacturers in a competitive business environment. Securing the IP of manufacturer SW is also an important aspect of the OSP platform due to the same reasons.

5.2 Independence

Objectivity is important when testing software and the closer the relation between the developer and the tester, the more difficult it is to be objective. The level of independence, and therefore the objectivity, increases with the 'distance' between the developer and the tester. The IEEE 1012 Standard for System and Software Verification and Validation [33] defines three types of independence: technical independence, managerial independence, and financial independence. Technical independence means that the verification personnel or tools should not be involved or used in the development of the system. Managerial independence means that the verification organization should be independent from the system vendor organization, while financial independence means that the budget of the verification effort should be independent of the budget for the system development and delivery. The IEEE 1012 also defines five forms of independence: classical, modified, integrated, internal and embedded. Classical independence is when the verification organization is an external organization (different company), and embodies all three types of independence (technical, managerial, financial). This is the level of independence adequate when testing safety critical systems.

Manufacturers often have their own simulators which they use in development or internal testing of control system software. This also applies to the ANS manufacturers. It is possible to maintain classical independence even though the manufacturer simulator is used in the test setup. In such a setup, the verification organisation should provide a test interface between the control system subject to test and the simulator controlled by this control system. In this way, the verification organisation will have full control of all the signals interchanged between the simulator and the control system. In addition, the simulator should be validated by the verification organisation to be fit for purpose:

- simulators shall not set restrictions on the test scope and test scenarios,
- it shall be possible to get access to all relevant signals through the test interface,
- it shall be possible to validate the correctness of the simulator and all its components, and
- it shall be possible to validate the correctness of the interface between simulator, test interface and control system software.

5.3 Test result validation

It is crucial to validate the results from the use of the simulation-based testing to achieve the needed confidence in the test activity and finally in the correctness of the ANS under test. Apart from the fact that the ANS successfully should handle all the simulated scenarios according to the evaluation criteria, confidence arise from especially two aspects:

- correctness of the simulation-based test results and
- the sufficiency or completeness of the tested scenarios, i.e. the level of coverage.

Correctness of the simulation-based test results depend on the validation of the digital twin, meaning the digital models, emulated systems, co-simulation of models and test interfaces. Validation is done in several ways and at different places in the testing process:

- interface and validation testing prior to starting the testing
- comparison of the digital twin simulation-based test results to results and data from real testing
- cooperation with the manufacturer during testing where the manufacturer gives input whether the simulations and results are valid or trustworthy, as discussed in 5.1
- test results review activities performed by the manufacturer, ship owner and the verification organization, aiming at concluding and validating the end results of the testing activity

The final challenge of any testing is how sufficient, complete or representative the test scope is in addressing all the critical behaviour, functionality, robustness or performance of the system under test, i.e. the level of coverage. Confidence is often initially perceived by the absence of failed test results, but in the end, it is the level of coverage that finally creates the needed confidence. ANS need to handle a very large number of different scenarios, and methods for assessing which scenarios that are representative or important to test and which are not, are a future research question relevant for many complex algorithms e.g. in autonomous or AI technologies. Initial methods for addressing this problem is briefly discussed in chapter 4.

6. Use cases and results

A prototype of the test system, TestIT, has been developed and is used in cooperation with Norwegian Defence Research Establishment (FFI), who is developing an ANS for the unmanned surface vessel (USV) Odin [34], see Figure 10. The control system has been interfaced to TestIT and several scenarios have been used to gather experience for further development of TestIT.

In the following, the test setup and the test cases are described. The chapter ends with a discussion of the results and experience obtained during the initial testing.

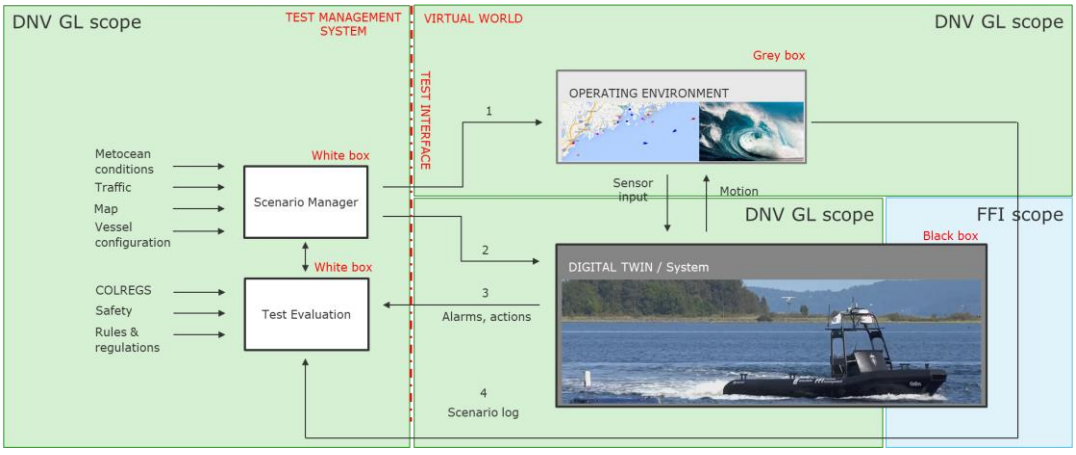


Figure 10: Illustration of the test setup

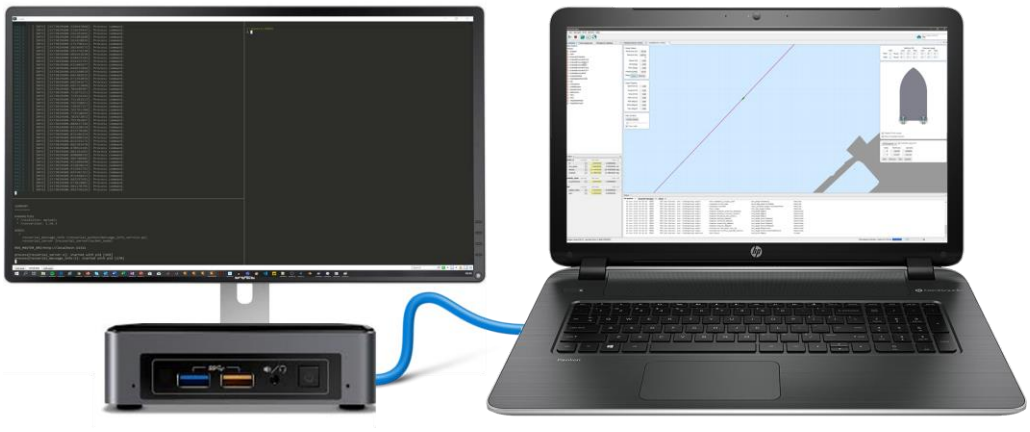


Figure 11: Test setup

6.1 Test setup

An illustration of the test setup is shown in Figure 10. As may be seen from the figure, DNV GL is responsible for the test management system (TestIT), the operating environment and the digital twin of Odin apart from the FFI-developed ANS controlling Odin.

Odin is a 10.5m long and 3.5m wide high-speed USV operating in the displacement, semi-displacement and planning regions. A simplified model of the USV, able to operate in all three regions, has been developed using a model and parameter identification method proposed by Eriksen and Breivik [35] based on log data received from FFI.

The ANS is running on a mini-pc, running Linux and Robotic Operating System (ROS) [36], see Figure 11. The simulator, which in this setup includes vessel, water jet modules, target ships and encounter evaluation, is interfaced to the control system using an external ROS node. Environmental disturbances have been excluded for simplification. The target ships may be operated manually or guided by waypoints, and their collision avoidance system may be enabled or disabled.

The ANS running on the mini-pc is identical to the ANS running on Odin. The simulator is interfaced to all necessary input and output signals of the ANS, making the ANS “believing” it is controlling the real vessel. The Odin ANS contains an experimental collision avoidance algorithm under development at FFI. Currently it is compliant with parts of the COLREG, and the system can distinguish between head-on, crossing and overtaking situations as well as last resort emergency maneuvers in order to avoid running ashore or colliding with obstacles. When calculating possible evasive maneuvers, the algorithm is parametrized in order to mimic behaviours that are judged to be reasonable for a vessel of Odin's size when manoeuvring in a coastal area with traffic dominated by other vessels of a similar size.

It is only the navigation part of the ANS that is included in the test setup, while the situational awareness is emulated by the simulator by outputting a list of target ships with necessary information to the ANS. This list may however be manipulated to emulate the effect of measurement noise, failed sensors, inadequate target ship tracking, etc.

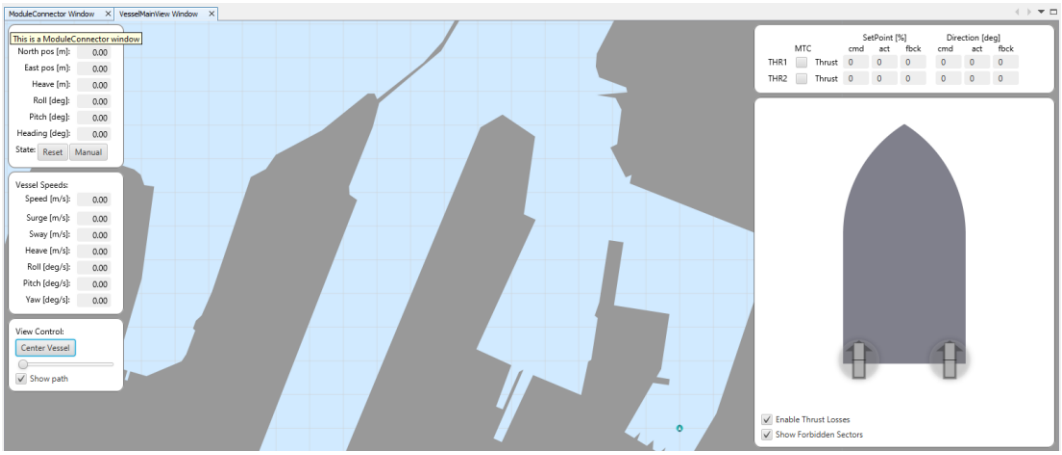


Figure 12: Initial position in Dora basin, Trondheim

6.2 Test cases

The main reason for testing the ANS under development by FFI, is to gather initial experience of the test process for such systems. Questions like which systems need to be included in the test setup, how to evaluate encounters, what information need to be published by the ANS etc. may hopefully be answered during this cooperation.

Several scenarios have been generated including head-on, crossing and overtaking situations where Odin has been both stand-on and give-way vessel. Each scenario has numerous episodes where small adjustments to initial position, speed and heading. So far, for all the episodes, collision avoidance of the target ships has been deactivated. Some of the episodes will be described in the following.

All scenarios and episodes are simulated in the area just outside Trondheim harbour and the initial position, (x=0) and (y=0), is put in the Dora basin (latitude = 63.44054 [deg] and longitude 10.42294 [deg]), see green dot in Figure 12.

6.2.1 Head-on encounter, 1

The first episode is a head-on encounter using one target ship sailing straight towards Odin, see Figure 13. Odin (white) is sailing at 25 knots, while the target ship (green) is sailing at approximately 20 knots, see Figure 13. The results are shown in Figure 14 to Figure 16.

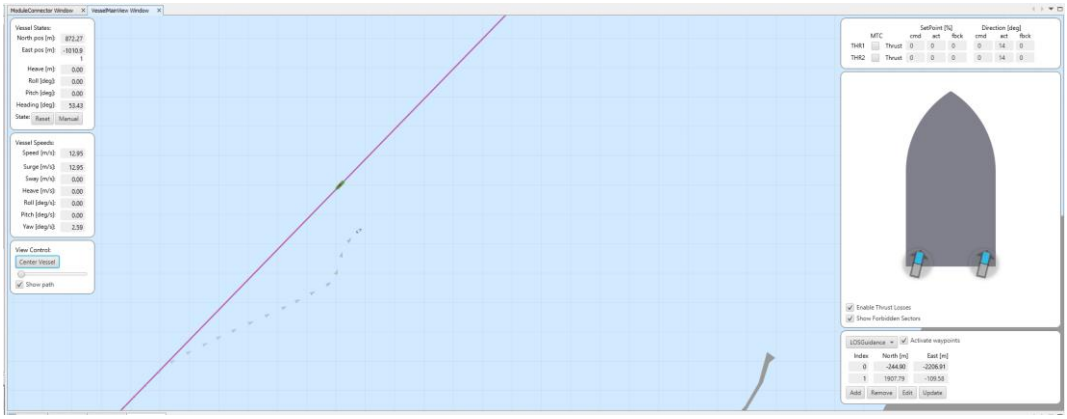


Figure 13: Head-on encounter, screen shot during simulation

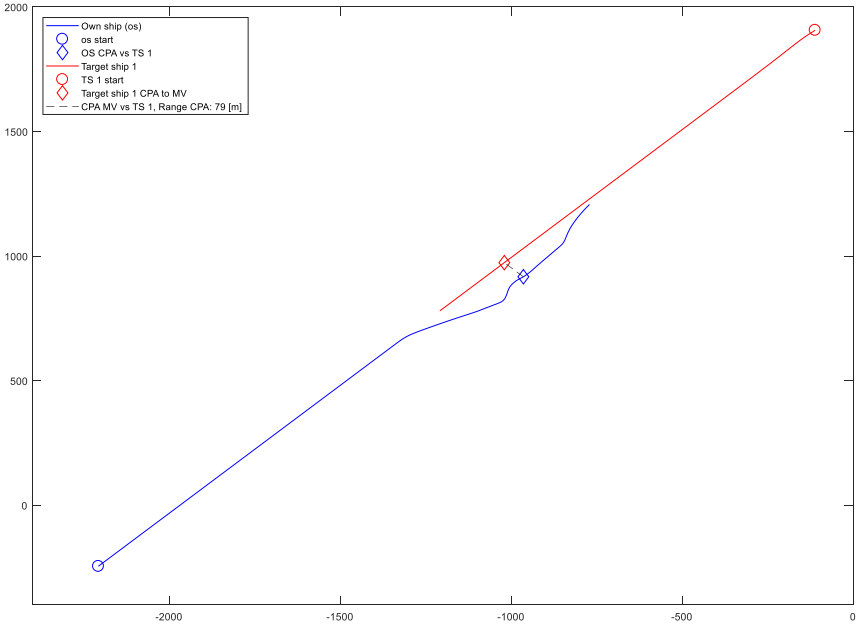


Figure 14: North - east position of own ship and target ship

Figure 14 shows the north and east position of both Odin and the target ship relative to the initial position, which is set in the Dora basin, Trondheim. The initial position of Odin is in lower left corner in the figure, and Odin is sailing north east (blue line), while the target ship starts in the upper right corner sailing south west (red line).

Odin continues straight towards the target ship until the range between the two vessels is approximately 830 m. The ANS is then commanding Odin to change course to starboard, as seen in Figure 15.

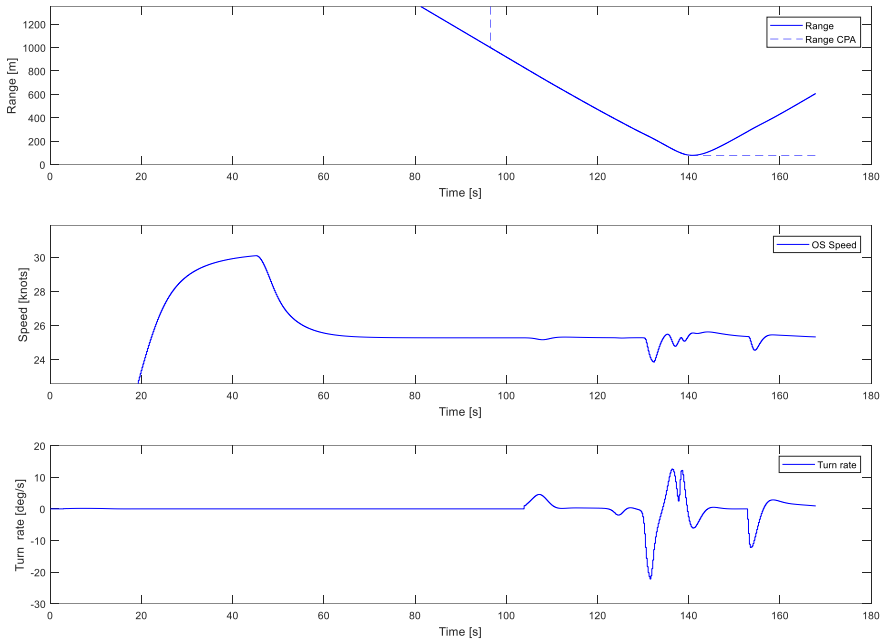


Figure 15: Range, speed and turn rate of Odin

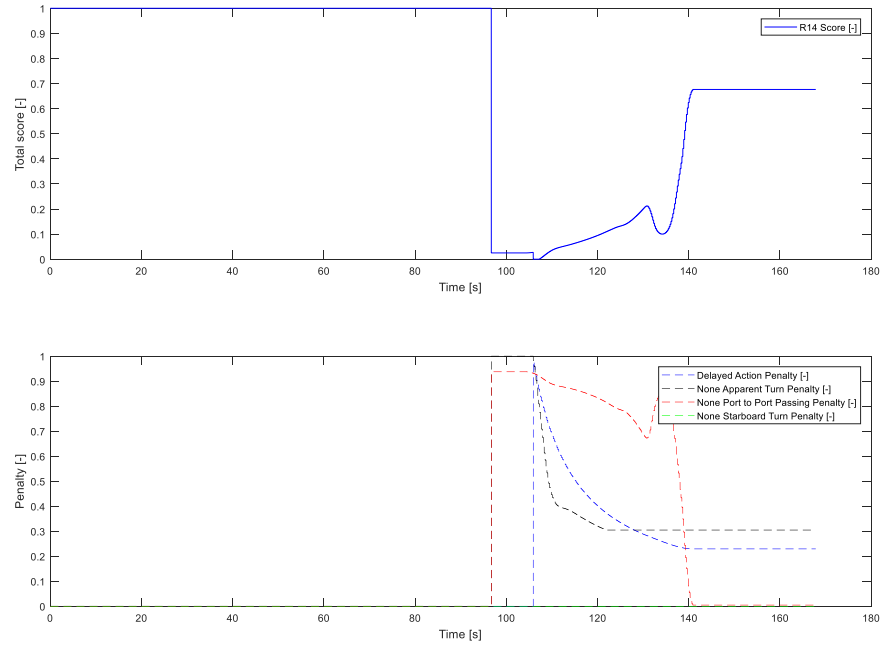


Figure 16: Total score and penalties for head on encounter

As seen in Figure 14, Odin continues straight forward for some time after first being commanded to starboard. After approximately 25 seconds, Odin is commanded into a port turn, making Odin sailing towards the target ship. After only a few seconds, Odin is again commanded starboard and passes the target ship with a CPA distance of approximately 80 m.

The evaluation is set to start when the range between the vessels are less than 1000 m. The upper part of Figure 16 displays the total score for the head on scenario, while the lower part of the figure displays the penalties which reduces the total score, see equation (8). The score value ranges from 0 to 1, where 1 is the best possible score.

CPA is reached after approximately 140 seconds. As may be seen from Figure 16, the estimated penalties are dynamic until the target ship is passed and the actual range and pose at CPA are known. The ANS is penalized for delayed action and none-apparent turn.

6.2.2 Head-on encounter, 2

The first head-on encounter was slightly modified to give the episode which is shown next. In the first head-on encounter, the initial range between the Odin and the target ship was 3000 m. In the next encounter, the initial range was approximately 1450 m. The speed of the target ship was also increased from 20 to 23 knots.

As seen in Figure 17, showing the north-east position of own ship and target ship, the manoeuvre that Odin performs is slightly different compared to the first episode. Odin is commanded to take a starboard turn, then goes straight before a slight port turn is initiated just before Odin is at its closest point of approach. The range at CPA is 121 m.

Figure **18** displays range between Odin and the target ship, together with speed and turn rate of Odin for the second head-on encounter. If comparing this figure to Figure 15, one may notice that Odin in both figures has an overshoot in speed. The speed setpoint is set to 25 knots, but due to the overshoot, the speed is approximately 29 knots when the maneuver is initiated. The difference in speed may explain the slightly difference in navigational behaviour.

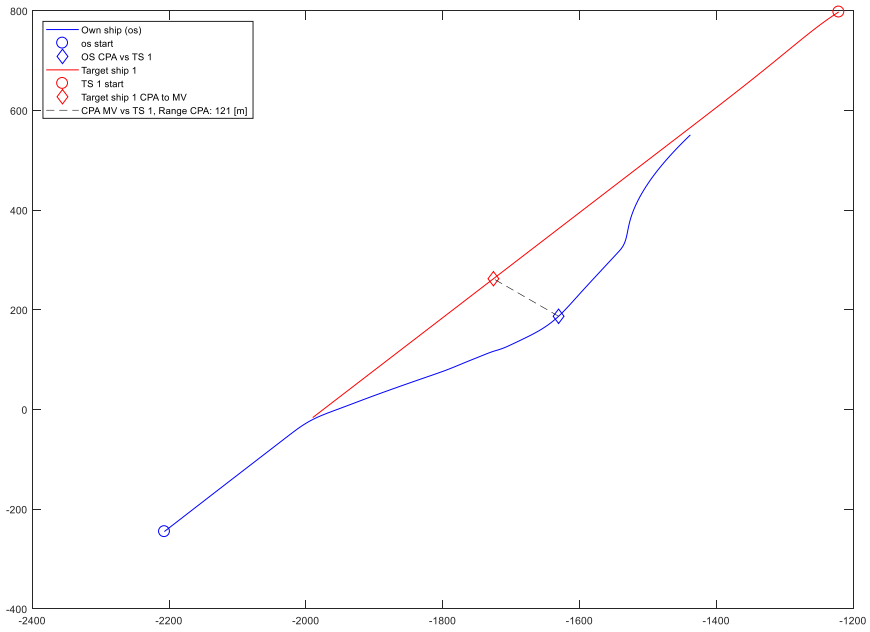


Figure 17:North - east position of own ship and target ship

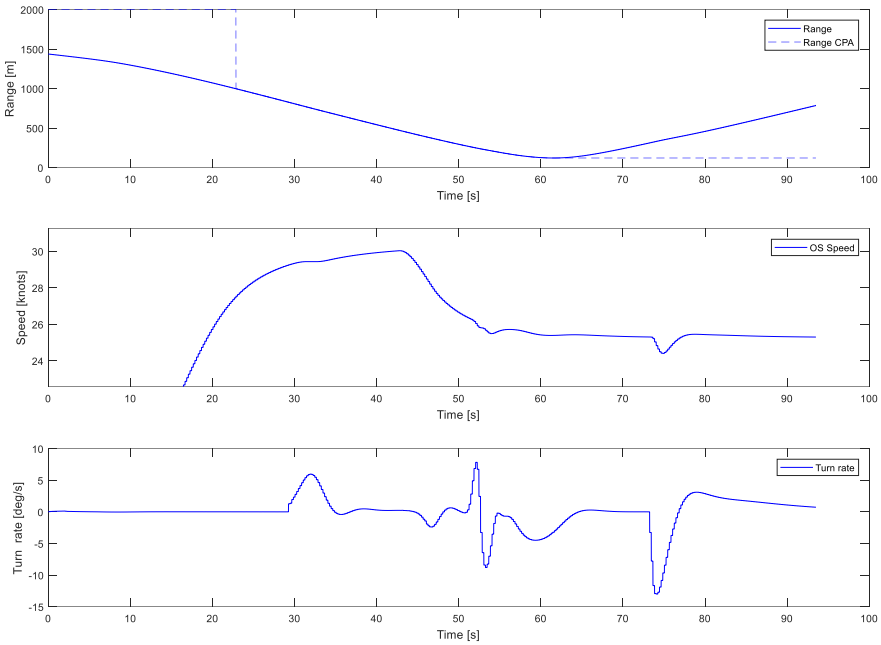


Figure 18**: Range, speed and turn rate of Odin**

6.2.3 Multiple target ships encounter

The last episode shown is a multiple target ship encounter, see Figure 19. Odin is sailing head-on to one target ship. At the same time, Odin is stand on for a crossing vessel coming from port side of Odin. This gives a situation where Odin is both a give way and a stand on vessel at the same time.

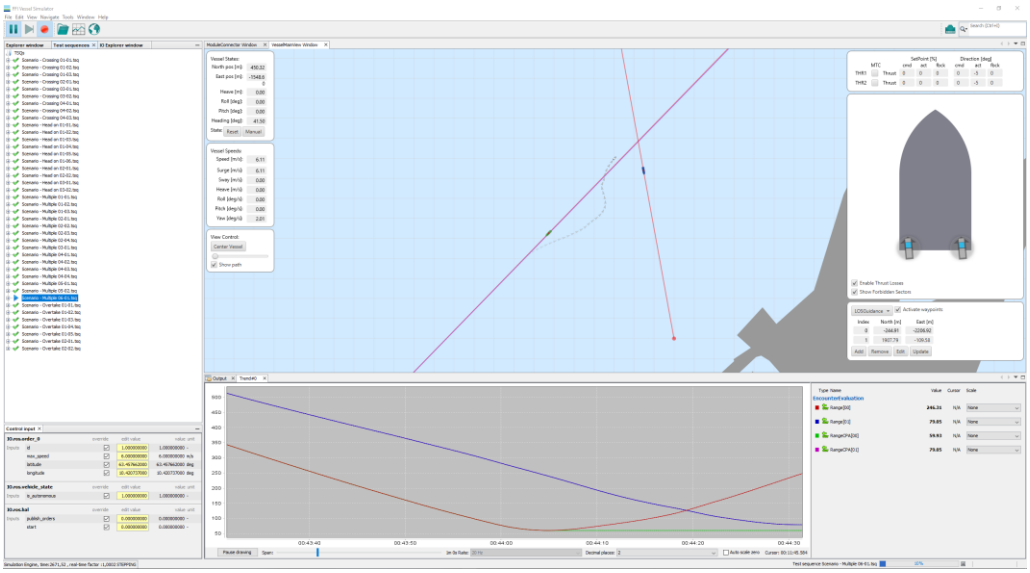
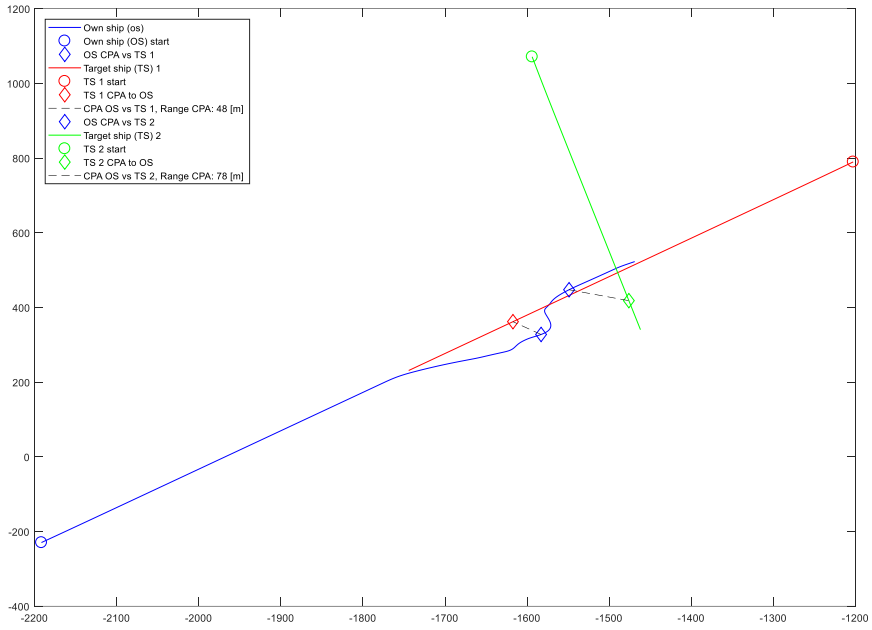


Figure 19: Multiple target ship encounter, screen shot during simulation



580

Figure 20: North - east position of own ship and target ship

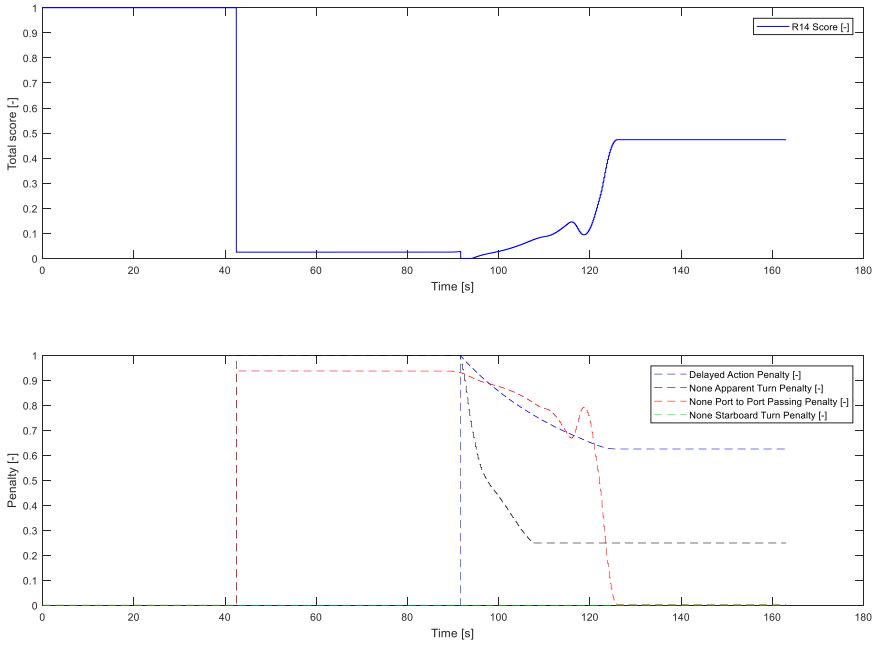


Figure 21: Total score and penalties for multiple target ship encounter: head-on encounter

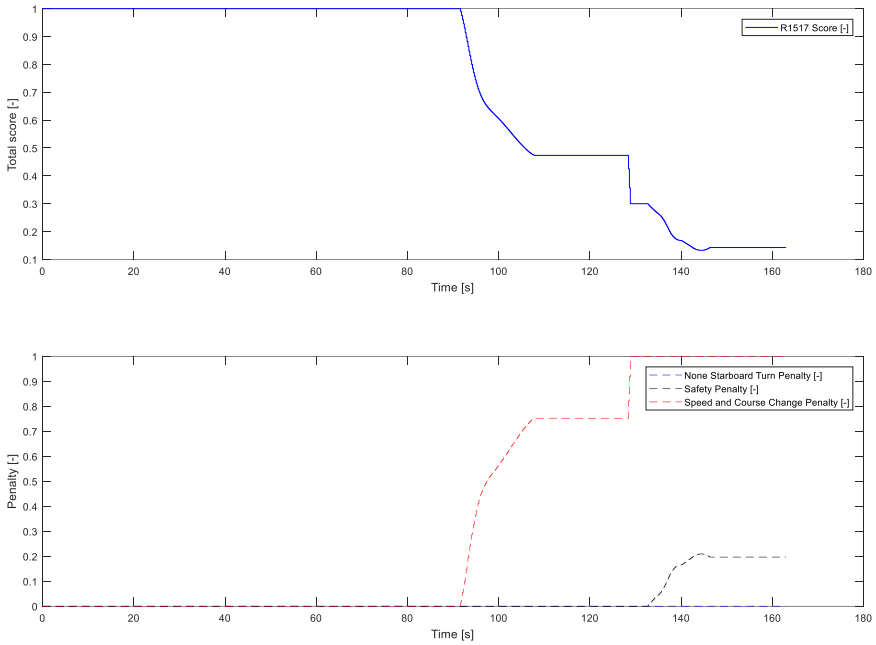


Figure 22: Total score and penalties for multiple target ship encounter: crossing encounter

Figure 20 shows the north-east position of Odin and the target ships during the scenario. Odin is navigating from south-west to north-east (blue line), while one target ship is navigating straight towards Odin (red line). The second target ship is navigating from north to south-east (green line). Initial positions of Odin and the target ships are marked as circles while CPA between Odin and the different target ships are marked with diamond shapes.

The range at closest point of approach between Odin and the first target ship (red) is 48 m, while the range at closest point of approach between Odin and the second target ship (green) is 78 m.

As seen in Figure 20, Odin takes a starboard turn to safely pass the first target ship. Then, to avoid the second target ship, the autonomous navigation system is commanding Odin into a port turn. According to COLREG, the target ship is a give way vessel, and should therefore alter either speed or course or both to avoid Odin, but this is not done since the collision avoidance system on the target ships are deactivated.

Figure 21 and Figure 22 present the total score and penalties for the head-on encounter and for the crossing encounter, respectively. The score and penalty functions for the crossing encounter may be found in [18].

As may be seen from Figure 21, the total score for the head-on scenario is mostly reduced due the delayed action penalty. Delayed action penalty is calculated using range between Odin and the target ship at the time when the target ship is detected, the range at the time when the maneuver starts and the range at CPA. In this situation, the range of detection is 1000 m, while Odin starts the navigation to starboard at a range of approximately 420 m. The range at CPA was 48 m.

Inspecting Figure 22, one may notice that the total score is very low, and the main reason for this is the penalty given for change in speed and/or course. In addition, a small penalty is given for

safety which looks at range between the vessels and the pose between them. However, no penalty is given for Odin taking a port turn to go behind target ship 2, which according to COLREG rule 17 (c) should have been penalised.

6.3 Discussions

Simulation-based testing of ANS has been demonstrated in the sub chapters 6.1 and 6.2. Even though there are still a way to go before the test tool is available for use in the maritime industry, much experience has been gained during this work.

The interface between the test system, the virtual world and the ANS will be important. For the test setup in this demonstration, the test system and the virtual world were communicating with the ANS using an external ROS node, while the test system and the rest of the virtual world were running in the same simulator. One may foresee that other constellations will be used in the future. For this to scale, an easy way of interfacing the different systems needed for testing ANS will play an important role. The Open Simulation Platform may be one piece in solving this.

The COLREG as it is written today may be a challenge both for collision avoidance algorithms and for evaluation algorithms. As pointed out in chapter 3, the COLREG is not written for software implementation, instead COLREG needs to be interpreted, but then how to ensure that this interpretation is correct, and also how the to evaluate COLREG performance?

For the scenarios presented in the above sub chapters, the evaluation is inspired by the evaluation method developed by Woerner [18]. One of the main challenges with this evaluation method is the use of fixed parameters. The parameters may manually be altered based on the specific situation using feedback from e.g. experienced navigators, but this is not feasible when running a lot of different scenarios. In Chapter 4, automatic scenario generation is discussed which will make it even harder to manually adjust the evaluation parameters. Another challenge with this evaluation method is that each encounter is evaluated independently of the others. This may clearly be seen from the third episode, where Odin is give-way for the head-on encounter, while at the same being stand-on for the crossing encounter. According to COLREG rule 17, own ship is only allowed to alter course and/or speed if she *finds herself so close that collision cannot be avoided by the action of the give-way vessel alone, she shall take such action as will best aid to avoid collision* [37]. In this scenario, Odin needs to change course due to the head-on target ship. The range between Odin and the second target ship when Odin starts changing course due to the head-on encounter, is too large for COLREG rule 17(b) to apply, and Odin is therefore penalized for this manoeuvre. This is seen when looking at Figure 22. According to COLREG rule 17 (c) the *vessel shall, if circumstances of the case admit, not alter course to port for the vessel on her own port side*. As may be seen from Figure 22, Odin is not penalized for the port turn initialized after the head-on encounter is passed. For the test tool to be used as a verification tool for autonomous navigation, multiple encounters need to be handled in a better way.

While running the scenarios and observing how Odin was navigated, it was evident that more information should be provided by the ANS. For the operator to better trust such systems, the reasoning for why the ANS is making certain manoeuvres, should be given to the operator/tester

which is in line with recommendations for explainability of autonomous vessel [38]. Information which should be provided by the ANS are (list is not exhaustive):

- Indication of target ships observed and assessed
- Indication of ANS calculated CPA information
- Indication of the COLREG rules which own ship is using for the different target ships
- Indication of the existing and possible new route for avoiding target ships

2Automatic scenario generation will be an important part of the test system. As could be seen from the test cases in previous sub chapter, similar episodes with only small adjustments in initial parameters, changed how the ANS solved the encounter. Using a brute force method where all combination of initial parameters are tested will not be efficient. Instead, the automatic scenario generation should be able to use knowledge about already performed tests to generate new scenarios and episodes, and to increase test efficiency.

7. Conclusion

For autonomous ships to be accepted by the society, it is said that autonomous ships need to be as safe or safer than conventional ships. Proving this may be a challenge, especially if only real-life testing is performed.

The ANS should go through a qualification scheme where safety goals and risks are identified, and qualification activities are performed to collect evidence for mitigating the risks and reaching the safety goals. DNV GL proposes to use a combination of real-life and simulation-based testing to assess the ANS. A scenario manager setting up test scenarios using a combination of scenario-based and functional-based testing combined with robustness testing and automatic search for critical scenarios, will be a vital part of the test system. Two different methods for evaluating the results from the testing are described. The test evaluation algorithm will need to trigger human assessment of possible ANS failures, and it is important that the evaluation algorithm does not fail to flag an actual ANS failure without signalling the need for a manual check.

Several scenarios and episodes have been simulated to test the test system, and results show that there are still work to be done both when it comes to generating new scenarios and episodes for efficient testing and when it comes to the evaluation criteria. Developing automatic scenario generation will be an important contribution for testing ANS which needs to be in place for safe deployment of ANS. The evaluation algorithms will also need further work and it is of utmost importance that they can handle multiple simultaneous encounters and at the same time handle situations both in open seas and narrow channels without the need of experts adjusting evaluation parameters in each situation. This will be especially important when introducing automatic testing and evaluation.

685 # References

[1] Safety4sea, “Human error the cause for most coastal vessels accidents in harbours,” https://safety4sea.com/human-error-the-cause-for-most-coastal-vessels-accidents-in-harbours/, 2018.

[2] M. Insight, October 2019. [Online]. Available: https://www.marineinsight.com/marine-safety/the-relation-between-human-error-and-marine-industry/. [Accessed March 2020].

[3] A. M. Rothblum, D. Wheal, S. Withington, S. A. Shappell, D. A. Wiegmann, W. Boehm and M. Chaderjian, “Key to successful incident inquiry,” in *2nd Int. Workshop Human Factor Offshore Oper.*, 2002.

[4] K. Wróbel, J. Montewka and P. Kujala, “Towards the assessment of potential impact of unmanned vessels on maritime transportation safety,” *Reliability ENgineering and System Safety,* vol. 165, pp. 155-169, 2017.

[5] B.-J. Vartdal, R. Skjong and A. L. St. Clair, “Remote-Controlled and Autonomous Ships,” DNV GL Group Technology & Research, Position Paper, 2018.

[6] Ø. J. Rødseth and H. C. Burmeister, “Risk Assessment for an Unmanned Merchant Ship,” in *TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation 9(3)*, 2015.

[7] E. Heikkilä, R. Tuominen, R. Tiusanen, J. Montewka and P. Kujala, “Safety Qualification Process for an Autonomous Ship Protoype - a Goal-based Safety case Approach,” in *TransNav 2017 - 12th International Conference on Marine Navigation and Safety of Sea Transportation*, 2017.

[8] D. GL, “DNVGL-RP-A203: Technology Qualification,” 2017.

[9] M. Wood, P. Robbel, M. Maass, R. D. Tebbens, M. Meijs, M. Harb, J. Reach, D. Wittmann, T. Srivastava, M. E. Bouzouraa, C. Knobel, D. Boymanns, M. Löhning, B. Dehlink, D. Kaule, R. Krüger, J. Frtunikj, F. Raisch, M. Gruber, J. Steck, J. Mejia-Hernandez, S. Syguda, P. Blüher, K. Klonecki, P. Schnarz, T. Wiltschko, S. Pukallus, K. Sedlaczek, N. Garbacik, D. Smerza, D. Li, A. Timmons, M. Bellotti, M. O'Brien, M. Schöllhorn, U. Dannebaum, J. Weast, A. Tatourian, B. Dornieden, P. Schnetter, P. Themann, T. Weidner and P. Schlicht, “Safety First for Automated Driving,” 2019.

[10] T. A. Johansen, T. I. Fossen and B. Vik, “Hardware-in-the-loop testing of DP systems,” in *Dynamic Positioning Conference, Control Systems II*, 2005.

[11] DNV GL, Kongsberg, SINTEF Ocean, NTNU, “Open Simulation Platform,” September 2018. [Online]. Available: www.opensimulationplatform.com. [Accessed 29 Augsut 2019].

[12] FMI, Functional Mock-up Interface, 2019.

[13] J. Zhang, X. Yan, X. Chen, L. Sang and D. Zhang, “A novel approach for assistance with

] anti-collision decision making based on the international regulations for preventing collisions at sea," 2012.

[14] W. Naeem, G. W. Irwin and A. Yang, "Colregs-based collision avoidance strategies for unmanned surface vehicles,"," in *Mechatronics, vol. 22, no. 6,*, 2012.

[15] S. Campbell, W. Naeem and G. W. Irwin, "A review on improving the autonomy of unmanned surface vehicles through intelligent collision avoidance manoeuvres," in *Annual Reviews in Control,*, 2012.

[16] L. P. Perera, "Autonomous ship navigation under deep learning and the challenges in COLREGS," in *Proceedings of the 37th International COnference on Ocean, Offchore and Arctic Engineering*, Madrid, Spain, 2018.

[17] L. P. Perera and B. Murray, "Situation Awareness of Autonmous Ship Navigation in a Mixed Environment Under Advanced Ship Predictor," in *Proceedings of the 38th International Conference on Ocean, Offshore and Arctic Engineering*, Glasgow, Scotland, UK, 2019.

[18] K. L. Woerner, Multi-Contact Protocol-Constrained Collision Avoidance for Autonomous Marine Vehicles, PhD thesis, Massachusetts Institute of Technology, 2016.

[19] P. K. E. Minne, "Automatic testing of maritime collision avoidance algorithms," Master thesis, NTNU, 2017.

[20] E. Henriksen, "Automatic Testing of Maritime COllision Avoidance Methods with Sensor Fusion," Master thesis, NTNU, 2018.

[21] K. Woerner, M. R. Benjamin, M. Novttzky and J. Leonard, "Quantifying protocol evaluation for autonomous collision avoidance," in *Autonomous Robots 43(5)*, 2018.

[22] K. L. Woerner and M. R. Benjamin, "Real-time Automated Evaluation of COLREGS-Constrained INteractions Between Autonomous Surface Vessels and Human Operated Vessels in Collaborative Human-Machine Partnering Missions," in *Oceans-MTS/IEEE Kobe Techno-Oceans (OTO)*, 2018.

[23] P. G. Stankiewicz and G. E. Mullins, "Improving Evaluation Methodology for Autonomus Surface Vessel COLREGS Compliance," MTS/IEEE OCEAN'19 Marseille Conference, 2019.

[24] S. Nakamura and N. Okada, "Development of AUtomatic Collision Avoidance System and Quantitative Evaluation of the Maneuvering Results," *International Journal on Marine Navigation and Safety of Sea Trasportation,* vol. 13, no. 1, pp. 133-141, March 2019.

[25] L. Li, W. L. Huang, Y. Liu, H. H. Zheng and F. Wang, "Intelligence Testing for AUtonomus Vehicles: A New Approach," in *IEEE TRANSACTIONS ON INTELLIGENT VEHICLES*, 2016.

[26] ISO, "Software and systems engineering - Software Testing," ISO, 2013.

[27] N. G. Leveson and J. P. Thomas, "STPA Handbook," March 2018. [Online]. Available: http://psas.scripts.mit.edu/home/get_file.php?name=STPA_handbook.pdf.

[28] B. Rokseth, I. B. Utne and J. E. Vinnem, “Deriving Verification Objectives and Scenarios for Maritime Systems Using the Systems-Theoretic Process Analysis,” *Reliability Engineering & System Safety,* vol. 169, pp. 18-31, January 2018.

[29] A. Bakdi, I. K. Glad, E. Vanem and Ø. Engelhardtsen, “AIS-Based Multiple Vessel Collision and Grounding Risk Indication based on Adaptive Safety Doman,” *Journal of Marine Science and Engineering,* vol. 8, no. 1, 2019.

[30] Wikipedia contributors, “Response surface methodology,” Wikipedia, The Free Encyclopedia, 23 August 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Response_surface_methodology&oldid=912154443. [Accessed 16 December 2019].

[31] Wikipedia contributors, “Bayesian optimization,” Wikipedia, The Free Encyclopedia, 12 December 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Bayesian_optimization&oldid=930478164. [Accessed 16 December 2019].

[32] Wikipedia contributors, “Kriging,” Wikipedia, The Free Encyclopedia, 19 October 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Kriging&oldid=922057035. [Accessed 16 December 2019].

[33] IEEE, “1012 Standard for System and Software Verification and Validation,” 2012.

[34] FFI, “USV Odin,” 2019. [Online]. Available: https://www.ffi.no/aktuelt/nyheter/norske-forskere-laerer-forerlose-bater-sjovett.

[35] B. O. H. Eriksen and M. Breivik, “Modeling, Identification and Control of High-Speed ASVs: Theory and Experiments,” in *Sensing and Control for Autonomous Vehicles*, Springer, 2017, pp. 407-431.

[36] ROS.org. [Online]. Available: www.ros.org.

[37] COLREGS, “International Regulations for Preventing Collisions at Sea,” Lloyd's Register Rulefinder, 2005 - Version 9.4.

[38] J. A. Glomsrud, A. Ødegårdstuen, A. L. St. Clair and Ø. Smogeli, “Trustworthy versus Explainable AI in Autonomous Vessels,” in *Paper presented at the International Seminar on Safety and Security of Autonomous Vessels (ISSAV) 2019. No proceedings available*, Espoo, 2019.