



UiT The Arctic  
University of Norway



University of  
South-Eastern Norway



Western Norway  
University of  
Applied Sciences



NTNU  
Norwegian University of  
Science and Technology

Faculty of Technology, Natural sciences and Maritime Studies

Department of Maritime Operations

University of South-Eastern Norway

## **Agent Transparency and Human Performance in Supervisory Control**

Koen van de Merwe

Dissertation for the degree of Philosophiae Doctor – August 2024



# **Agent Transparency and Human Performance in Supervisory Control**

Koen van de Merwe

© Koen van de Merwe 2024

UiT The Arctic University of Norway  
Faculty of Science and Technology  
Department of Technology and Safety

Norwegian University of Science and Technology  
Faculty of Engineering  
Department of Ocean Operations and Civil Engineering

University of South-Eastern Norway  
Faculty of Technology, Natural Sciences and Maritime Studies  
Department of Maritime Operations

Western Norway University of Applied Sciences  
Faculty of Business Administration and Social Sciences  
Department of Maritime Studies

**Doctoral dissertations at the University of South-Eastern Norway no. 203**

ISSN: 2535-5244 (print)

ISSN: 2535-5252 (online)

ISBN: 978-82-7206-882-9 (print)

ISBN: 978-82-7206-883-6 (online)

This dissertation was supervised by:

Prof. Dr. Salman Nazir	University of South-Eastern Norway Department of Maritime Operations Borre, Norway	Main supervisor
Dr. Steven Mallam	University of South-Eastern Norway Department of Maritime Operations Borre, Norway	Co-supervisor
	Memorial University of Newfoundland Fisheries & Marine Institute St. Johns, Canada	
Mr. Øystein Engelhardtzen	DNV Group Research & Development Høvik, Norway	Co-supervisor

Composition of the doctoral committee:

Dr. Ziaul Haque Munim	University of South-Eastern Norway Department of Maritime Operations Borre, Norway	Chairperson
Dr. Mica Endsley	SA Technologies Gold Canyon, Arizona, USA	1 <sup>st</sup> opponent
Prof. Dr. Ingrid Utne	Norwegian University of Science and Technology Department of Marine Technology Trondheim, Norway	2 <sup>nd</sup> opponent

This work was supported by DNV and the Norwegian Research Council under grant number 311365.



# Dedication

*For my wife, Fenna, and my children, Simon and Robin*





## Preface

Pursuing a PhD has been something I wanted to do since I graduated from Leiden University in 2004 with an MSc. in Cognitive Psychology, I just never got around to doing it. After my first degree, I decided to pursue a second one within Industrial Ecology, with the aim to “do something within sustainability”. But when I obtained a position as a Human Factors researcher at the National Aerospace Laboratory in Amsterdam, I immediately knew this was what I wanted to do – optimizing human performance in safety critical domains. Focussing on human-automation interaction for commercial-, fighter pilots and Air Traffic Controllers, it occurred to me that I could specialise in this even further by doing a PhD. However, my wife and I decided to accept an even bigger challenge. In 2010, we emigrated to Norway.

Without any social network, nor any Norwegian vocabulary, we basically started our adult life anew. However, we both were fortunate enough to be able to find relevant, interesting, and challenging work as consultants within the oil & gas industry. In 2012, our first child was born, and in 2015, the second one. Whilst still learning Norwegian, changing nappies, and a full-time job, we had enough on our plate to keep us occupied for a while. This meant that although the thought of pursuing a PhD degree never left my mind, it began to fade, and at some point, I thought I would never do one. Until at the end of 2019, an opportunity came along that sparked my interest: an industry PhD position within my employer’s Research & Development department on the topic of autonomous shipping and human performance. As this was right up my alley, I applied, got the position, was approved by the University of South-Eastern Norway’s joint PhD program, and started on my new journey. Finally.

The excitement was short-lived as, on my very first day as a PhD candidate, I was in quarantine because of a COVID-19 case at work. Two weeks later, Norway was in lockdown. Here I was, behind my PC at the dining table, together with my wife and two children. We were doing morning-, afternoon-, and evening shifts where one was working, and the other was home-schooling and entertaining the children. In between activities, I was trying to focus on whatever I was going to research. The thought of performing a task as large as a PhD, alone, and in these conditions, was daunting. Needless to say, it was a challenging time.

Fortunately, with the guidance and unwavering support of my supervisors, employer, family, friends, and colleagues, I was able to take it step-by-step, persevere, and ultimately made this a successful and extremely rewarding chapter of my life. I look back on it with fulfilment, joy and pride.

Høvik, Norway

Koen van de Merwe



# Acknowledgements

First and foremost, I would like to thank my supervisors Prof. Dr. Salman Nazir, Dr. Steven Mallam, and Øystein Engelhardtson for their unwavering support and supervision. Throughout the four-year period, in our online bi-weekly meetings, I could always count on good discussions, guidance, earnest opinions, and support. Your reflections have raised the level of the dissertation and have helped me in critically reflecting on science and my work. Thank you.

When I had the opportunity obtain assistance from a summer student to help me in some practical issues, I was very fortunate to hire Koen Houweling. As a navy-certified navigator and student at USN at the time, Koen was pivotal in creating realistic traffic situations and the transparency symbology. What started as a two-month hire-in, quickly became a full year in which Koen showed his dedication and hard work in assisting me and others within DNV. Your contributions, experience, and reflections have been invaluable to this dissertation.

I was fortunate to meet several colleagues within the university community that have supported me with their knowledge and experience. In particular, I would like to mention Dr. Ole Andreas Alsos and Dr. Erik Veitch, for inviting me to the NTNU Shore Control Lab and join their research. Furthermore, I would like to mention Douglas Owen from Aalto University for sharing valuable reflections and insights.

A sincere gratitude goes to the professionals that have made this PhD a possibility: Bastø Fosen, Massterly, the various nautical training institutes in Norway, colleagues at USN, and other participants in the various workshops, interviews, and experiment. Your contribution has been invaluable in producing the scientific results in this dissertation. The support of Andreas Madsen, Georg Steintveit, and Tom Eystø has especially been helpful in acquiring participants for the experiment.

I would like to thank my colleagues at DNV, USN, and fellow PhD-students for the discussions and support during the years. Even though a PhD can be a lonely task, your social, technical, and scientific support was essential in persevering and keeping up the pace. A special mention goes to Kristian Gould who has helped me in reflecting on my work during critical moments. Furthermore, I would like to thank Prof. Dr. Partick Hudson from Leiden University, and Prof. Dr. David O'Hare from University of Otago, for inspiring me to pursue a career in Human Factors.

My family and friends have supported me in numerous ways, forms, and stages throughout the PhD process. As we live in Norway, the weekly conversations with my parents are necessarily online. Nevertheless, they have helped in keeping the focus, motivation, and stamina to continue.

Finally, I would like to mention my wife and children. Fenna, thank you for your encouragement throughout the years, your wisdom, stamina, and support in making it work whilst running a busy family. Now this chapter can be closed. Simon and Robin, thank you for bringing immense joy to my life. Your enthusiasm, engagement, patience, kindness, reflections, and love have been the greatest inspiration.



# Summary

The maritime industry is investing in advanced technologies to reduce its environmental footprint, become more attractive to personnel, improve its safety record, and enhance its resilience against adverse conditions, whilst maintaining profitability. To contribute to achieving these goals, Artificial Intelligence (AI) is anticipated to play a central role in supporting navigators in critical decision making and possibly even allowing ships to sail without direct human involvement. Considering the safety-critical nature of maritime navigation, this means that AI-enabled systems need to demonstrate a high degree of reliability and robustness across a wide range of situations. However, given the limitations of such systems to operate in novel and complex situations, careful design, implementation, management and operation is required when deploying these in real-world environments. Therefore, proposed autonomous ship concepts typically employ human operators to monitor, supervise, and potentially intervene in the system to ensure the required performance and safety levels are achieved.

Decades of research has demonstrated that there are significant human performance challenges associated with assigning humans a supervisory role of highly automated systems. Since operators are tasked with supervising systems that make their own decisions and actions, they are typically removed from much of the information- and decision-making loop. Consequently, they may find it challenging to evaluate, understand, and predict the behaviour of such systems. In addition, passive information processing may lead to complacent behaviour, biases in decision making, a reduced ability to detect critical information, an over- or underreliance on the system, and high workload when switching from supervised- to manual control. As a result, their capability to intervene is affected. Nevertheless, recent research has suggested that by disclosing the system’s decisions, planned actions, and internal reasoning to the operator, i.e., by making the system “transparent”, some of these challenges may be alleviated. However, considering the novelty of the application of AI-enabled systems in safety-critical domains, there is limited experience with the effect of transparency in these settings. As such, there is an urgent need to generate new knowledge, methods, and tools with regards to how humans may successfully interact with these types of technologies. Therefore, this dissertation aims to explore the following overarching research question (RQ):

*How does agent transparency support human performance in supervisory control?*

The main RQ is decomposed in the following sub-questions (see Table 1).

*Table 1. The research questions addressed in this dissertation.*

No.	Research question
1	What is the relationship between agent transparency and Situation Awareness, mental workload, and task performance?
2	How is human performance achieved in conventional- and supervised maritime collision avoidance?
3	How does a model for human information processing form the basis for agent transparency in the ship autonomy context?
4	How should a maritime collision avoidance system be made transparent to a human supervisor?
5	What is the relationship between agent transparency and Situation Awareness, mental workload, and task performance in maritime autonomous collision avoidance manoeuvring?

To answer the RQs, a mix of quantitative and qualitative methods were deployed. The Preferred Reporting items for Systematic review and Meta-Analysis method (PRISMA) was used to systematically map and assess the scientific literature for empirical research addressing the effect of transparency on central Human Factors variables: Situation Awareness (SA), mental workload, and task performance (RQ1). A Goal-Directed Task Analysis (GDTA) was used to identify information requirements for conventional- and supervised collision and grounding avoidance (CAGA) manoeuvring (RQ2). A model of human information processing was adapted to the autonomous shipping context and was used to structure the data from the GDTA and generate layers of transparency for a hypothetical CAGA system (RQ3). An iterative design process was used to develop traffic situations and Human Machine Interface (HMI) concepts for displaying various levels of transparency (RQ4). A controlled experiment was used to assess the relationship between transparency and agent performance (RQ5). In addition, a range of secondary methods were used to gather, structure, validate, and quality-assure the data. Finally, professional navigators played a key role as subject matter experts (SME) throughout the dissertation to ensure external validity.

As depicted in Table 2, the Systematic Literature Review (SLR) found a promising effect of transparency on SA and task performance, without affecting mental workload, for studies where participants were responding to proposals or supervising automation. As documented in Article 1, it was suggested that strategies to improve human performance, when interacting with intelligent agents, should focus on allowing humans to see into its information processing stages, considering the integration of information in existing HMI solutions. By using the PRISMA method, the results contributed by systematically mapping the scientific knowledge regarding transparency as a design principle for effective human-automation interaction. In addition, the results provide an incentive for designers to apply transparency principles when developing systems in which operators are tasked with responding to proposals or perform supervisory control, e.g., as proposed in autonomous shipping concepts.

The GDTA mapped and analysed the goals, decisions, and cognitive tasks associated with conventional- and supervised collision and grounding avoidance. As reported in Article 2, data was obtained from in situ observations and interviews with nine navigators onboard passenger ferries, an appraisal of the collision regulations, and from relevant company documentation. The results provide a detailed analysis of the change in information requirements from conventional- to supervised collision avoidance. The study explored the shift in cognitive activities when the navigator's task changes from performing collision avoidance to supervising a system performing collision avoidance. To support operators in this change, explicit information requirements were identified that should allow for insight into the agent's decisions, planned actions, and underlying reasoning.

As addressed in Article 3, a model for human information processing was adapted and repurposed to function as a model for transparency. The model by Parasuraman, Sheridan, and Wickens (PSW) was contextualized to the maritime collision avoidance setting such that the information from the GDTA could be structured into unique and distinct layers. It was suggested that this model may serve as a framework for transparent design as the steps in the model could represent the agent's input parameters, analysis, decisions, and planned actions. Using the model in this way, a minimum set of information requirements was made, organised in layers per processing step, resulting in a model for transparency.

Based on the structured information requirements, HMI concepts were developed for making an autonomous collision avoidance system transparent for its supervisor (see Article 4). Here, traffic situations and symbology were developed to operationalise transparency for a hypothetical CAGA system. The symbology was integrated into the primary task display for collision avoidance, i.e., the radar display, and the PSW model was used to create distinct levels of transparency. For the purpose of this dissertation, this enabled each transparency level's individual contribution to human performance to be evaluated. As such, this activity provided the groundwork for the empirical evaluation of agent transparency in a maritime collision avoidance context. In addition, the results demonstrated the value of the PSW model as a design framework for creating levels of transparency for autonomous agents.

Finally, the effect of transparency on SA, mental workload, and task performance was evaluated using a controlled experiment (see Article 5). Based on the PSW model and the traffic situations, four levels of transparency and two levels of traffic complexity were varied in an experiment with 34 navigators. The results demonstrated a positive effect of transparency on SA without affecting mental workload. However, the time to comprehend the provided information increased with increased levels of transparency. These results indicate the benefits of applying transparency principles to autonomous collision avoidance systems in terms of SA, but care should be taken in time-critical conditions where the added transparency information may affect timely decision making. Furthermore, considering the absence of the effect of transparency on mental workload, these results also indicate the value of proper HMI design through applying a structured and systematic human-centred design process, as applied in this dissertation.

*Table 2. Summary of key point and contributions of this dissertation.*

<b>RQ / Article</b>	<b>Key points</b>	<b>Theoretical contributions</b>	<b>Practical contributions</b>
1	Systematically gathered and assessed empirical evidence for the relationship between agent transparency and key human factors variables	The results contribute to the knowledge regarding transparency as a design principle for effective human-automation interaction	The results provide incentives to designers for applying transparency principles, especially for when humans respond to proposals and perform supervisory control
2	Goals, decisions, cognitive tasks and SA requirements were identified for conventional- and supervised collision avoidance	The results provide a detailed analysis of the change in information requirements from conventional- to supervised collision avoidance	The results provide concrete insights into the SA requirements for supervised collision avoidance
3	Adapted the PSW model to the maritime collision avoidance domain and used the model to organise the SA requirements into layers of transparency	The results expand the applicability of the PSW model to represent a model for agent internal information processing, i.e., transparency	The results provide a set of minimum SA requirements, organised per layer of transparency

<b>RQ / Article</b>	<b>Key points</b>	<b>Theoretical contributions</b>	<b>Practical contributions</b>
4	Developed realistic traffic situations and applied the transparency model and SA requirements to develop realistic HMIs	The results provide the groundwork for empirical evaluations of transparency levels	The results provide insight into the practical value of the model as a design framework for transparent agents
5	Experimentally evaluated the transparency model in autonomous collision avoidance context and found effects of transparency on SA and task performance, but not on mental workload	The results add to the knowledge of the effects of transparency on key human factors variables and empirically evaluate the proposed transparency model	The results provide insight into the anticipated human performance effects of transparency when applied to autonomous agents

To conclude, this dissertation investigated the role of agent transparency in supervisory control and contributed with knowledge, methods, and tools regarding transparency in general and its application to the maritime domain specifically. As humans are foreseen to play a critical role in overseeing the functioning of AI-enabled systems, the operator’s ability to understand, predict, and evaluate agent behaviour becomes a critical aspect of the human’s supervisory task repertoire. Consequently, it is essential that humans are informed and supported in making accurate decisions to enable timely and appropriate control when needed. Therefore, the aim of this dissertation was to generate and advance the knowledge on how supervisory control can be supported through agent transparency. This dissertation has contributed to this aim by recognising the importance of transparency in safety critical domains in terms of human performance, exploring the impact of autonomy on the operator’s cognitive tasks, constructing a model for transparency, operationalising transparency for the maritime navigational context, and assessing its effects in a controlled experimental setting. The results have implications for scientific research and for the application of transparency as a design principle for autonomous agents. In addition, this dissertation has made explicit the role-change that may be anticipated when introducing autonomous systems. With these new insights, meaningful human work may be created where the combined capabilities of human-agent teams can be optimised. Ultimately, this dissertation advocates the relevance of affording human operators with insight into the reasoning of autonomous systems and established transparency as an important prerequisite on the path towards safe and effective human-supervisory control.



## Structure of the dissertation

This dissertation consists of two parts:

**Part I:** This part of the dissertation introduces the context, theoretical background, and research questions, as well as the methodological framework that was applied. In addition, the results and key findings from the relevant articles are presented, including a discussion on their implications. Finally, directions for future research, contributions and conclusions are indicated.

**Part II:** This part includes appendices in which additional background information to the studies is provided. In addition, the five publications relevant to this dissertation are included here. Article 1 is a systematic literature review, article 2 is an empirical study exploring operator goals, decisions, and SA requirements in a supervised autonomy setting. Article 3 elaborates on the use of a human information processing model as the basis for creating levels of transparency. Article 4 discusses the process for developing transparent HMIs for supervising an autonomous collision and grounding avoidance system, including the development process for generating realistic conflict scenarios on which the transparency levels could be evaluated. Finally, article 5 presents the results of an experimental evaluation in which the effect of transparency on key human performance variables was assessed.



# List of publications

## Appended articles

### Article 1:

van de Merwe, K., Mallam, S., & Nazir, S. (2024). Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review. *Human Factors*, 66(1), 180–208. <https://doi.org/10.1177/00187208221077804>

### Article 2:

van de Merwe, K., Mallam, S., Nazir, S., & Engelhardtson, Ø. (2024). Supporting human supervision in autonomous collision avoidance through agent transparency. *Safety Science*, 169, 13. <https://doi.org/10.1016/j.ssci.2023.106329>

### Article 3:

van de Merwe, K., Mallam, S., Engelhardtson, Ø., & Nazir, S. (2023). Towards an approach to define transparency requirements for maritime collision avoidance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 67(1), 483–488. <https://doi.org/10.1177/21695067231192862>

### Article 4:

van de Merwe, K., Mallam, S., Engelhardtson, Ø., & Nazir, S. (2023). Operationalising Automation Transparency for Maritime Collision Avoidance. *TransNav, International Journal on Marine Navigation and Safety of Sea Transportation*, 17(2). <https://doi.org/10.12716/1001.17.02.09>

### Article 5:

van de Merwe, K., Mallam, S., Nazir, S., & Engelhardtson, Ø. (2024). The Influence of Agent Transparency and Complexity on Situation Awareness, Mental Workload, and Task Performance. *Journal of Cognitive Engineering and Decision Making*, 18(2), 156–184. <https://doi.org/10.1177/15553434241240553>

## Other relevant publications

van de Merwe, K., Mallam, S. C., Engelhardtson, Ø., & Nazir, S. (2022). Exploring navigator roles and tasks in transitioning towards supervisory control of autonomous collision avoidance systems. *Journal of Physics: Conference Series*, 2311(1), 012017. <https://doi.org/10.1088/1742-6596/2311/1/012017>



# List of Tables

- Table 1. The research questions addressed in this dissertation. .... xiii
- Table 2. Summary of key point and contributions of this dissertation..... xv
- Table 3. Appended articles..... 6
- Table 4. The methodological choice for each research article. .... 22
- Table 5. Traffic situations and configuration used in the familiarisation and experimental trials. .... 31
- Table 6. Strategies to control validity and reliability in this dissertation. .... 36
- Table 7. Applying the PSW model to the information requirements from the task analysis. .... 43
- Table 8. Criteria for establishing a varied set of traffic situations. .... 44
- Table 9. Summary of predictions and results as discussed in Article 5. .... 53
- Table 10. Contributions of this dissertation. .... 70
- Table 11. The (condensed) results from the Goal-Directed Task Analysis (WP2), structured  
using the PSW model (WP3), and linked to HMI symbology (WP4)..... 85
- Table 12. Common information elements depicted on the radar screen. .... 92
- Table 13. Information specific to the collision avoidance system. .... 93
- Table 14. Information presented adjacent to the radar screen..... 96
- Table 15. Generic Level 1 SAGAT queries. .... 119
- Table 16. Generic Level 2 SAGAT queries. .... 121
- Table 17. Generic Level 3 SAGAT queries. .... 124
- Table 18. The traffic situations used for ranking participants' preferences..... 129
- Table 19. Definitions and examples that were read verbatim to the participants prior to ranking..... 130
- Table 20. Ranking preferences for transparency levels..... 130



# List of Figures

- Figure 1. A schematic overview of this dissertation. .... 5
- Figure 2. A model for human information processing (adapted from Lee et al. (2017) and Parasuraman et al. (2000))..... 10
- Figure 3. A model for human information processing with levels of Situation Awareness superimposed (adapted from Lee et al. (2017), Parasuraman et al. (2000), and Endsley (1995)). ..... 12
- Figure 4. Transparency as a property of the interface to provide observability and predictability of system behaviour..... 15
- Figure 5. Transparency as bi-directional communication between system and human. .... 16
- Figure 6. Transparency as an emergent property between systems and humans collaborating on a shared task. .... 16
- Figure 7. Transparency and explainability in a decision-making context (adapted from Endsley (2023b) and National Academies of Sciences, Engineering and Medicine (2022))..... 17
- Figure 8. A simple model of human information processing (adapted from Parasuraman et al., 2000)..... 19
- Figure 9. The Goal-Direct Task Analysis method (adapted from Endsley et al., 2003). .... 25
- Figure 10. The analysis framework employed in this study (van de Merwe, Mallam, Nazir, et al., 2024a). ..... 26
- Figure 11. Observations and interviews performed onboard a passenger ferry. .... 27
- Figure 12. The PSW model applied in a framework to derive transparency requirements for human supervised CAGA systems. .... 28
- Figure 13. Example of a form used in the workshops to validate the traffic situations. .... 29
- Figure 14. Experiments conducted at the TARG research lab at USN, onboard a passenger ferry, and at NTNU Ålesund respectively..... 30
- Figure 15. Independent and dependent variables. .... 30
- Figure 16. Screenshot from the ranking exercise from one of the participants..... 33
- Figure 17. Ensuring reliability, validity, and quality of the experiment. .... 35
- Figure 18. The generic framework used to map the changes between the baseline and supervision case..... 42
- Figure 19. Traffic situation without transparency information. .... 45
- Figure 20. Traffic situations with varying transparency levels. .... 46
- Figure 21. Mean scores for level 1, 2, and 3 SA, mental workload, and time to comprehension as a function of transparency and complexity. .... 49
- Figure 22. The development process as applied in this dissertation. .... 58
- Figure 23. A typical traffic situation depicted on the radar screen..... 91
- Figure 24. Traffic situation with a low level of transparency. .... 99
- Figure 25. Traffic situation with a medium (A) level of transparency..... 100
- Figure 26. Traffic situation with a medium (B) level of transparency. .... 101
- Figure 27. Traffic situation with a high level of transparency. .... 102
- Figure 28. Trial 1 - complexity = low, transparency = low, head-on (7HOLLT3). .... 103
- Figure 29. Trial 1 - complexity = low, transparency = medium (A), crossing (2CRLT32)..... 104
- Figure 30. Trial 1 - complexity = low, transparency = medium (B), head-on (2HOLLT31)..... 105
- Figure 31. Trial 1 - complexity = low, transparency = high, overtaking (10OTLLT321). .... 106
- Figure 32. Trial 1 - complexity = high, transparency = low, head-on (15HOHLT3)..... 107

Figure 33. Trial 1 - complexity = high, transparency = medium (A), crossing (21CRHLT32).....	108
Figure 34. Trial 1 - complexity = high, transparency = medium (B), head-on (11HOHLT31). ....	109
Figure 35. Trial 1 - complexity = high, transparency = high, overtaking (17OTHLT321).....	110
Figure 36. Trial 2 - complexity = low, transparency = low, head-on (9HOLTL3). ....	111
Figure 37. Trial 2 - complexity = low, transparency = medium (A), crossing (13CRLTL32).....	112
Figure 38. Trial 2 - complexity = low, transparency = medium (B), head-on (5HOLTL31).....	113
Figure 39. Trial 2 - complexity = low, transparency = high, overtaking (9OTLT321). ....	114
Figure 40. Trial 2 - complexity = high, transparency = low, head-on, (10HOHTL3).....	115
Figure 41. Trial 2 - complexity = high, transparency = medium (A), crossing (15CRHTL32).....	116
Figure 42. Trial 2 - complexity = high, transparency = medium (B) head-on (13HOHTL31). ....	117
Figure 43. Trial 2 - complexity = high, transparency = high, crossing (13OTHTL321). ....	118
Figure 44. The rating sheets used for the NASA-TLX (Hart & Staveland, 1988). ....	127
Figure 45. Weighting mental workload scores with pairwise comparisons (Hart & Staveland, 1988).....	128



# Abbreviations

---

Abbreviation	
AI	Artificial Intelligence
AIS	Automatic Identification System
ARPA	Automatic Radar Plotting Aid
BDI	Beliefs, Desires, Intentions
BIMCO	Baltic and International Maritime Council
BY	Buoy
CAGA	Collision And Grounding Avoidance
COG	Course Over Ground
COLREG	Convention on the International Regulations for Preventing Collisions at Sea
COVID	Coronavirus disease
CPA	Closest Point of Approach
CR	Crossing
CTA	Cognitive Task Analysis
DNV	(Not an acronym)
ECDIS	Electronic Chart Display and Information System
FV	Fishing Vessel
GDTA	Goal-Directed Task Analysis
GW	Give-Way
HAI	Human-Automation Interaction
HCD	Human-Centred Design
HF	Human Factors
HMI	Human Machine Interface
HO	Head-on
HRT	Human-Robot Transparency
ICJME	International Committee of Medical Journal Editors
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IMO	International Maritime Organisation
ISO	International Organisation for Standardisation
MITRE	(Not an acronym)
MV	Motor Vessel
NASA-TLX	National Aeronautics and Space Administration – Task Load Index
NC	No collision
NMA	Norwegian Maritime Authority
NTNU	Norwegian University of Science and Technology
OOTL	Out-of-the-loop
OT	Overtaking/overtaken
PRISMA	Preferred Reporting items for Systematic review and Meta-Analysis
PSW	Parasuraman, Sheridan & Wickens
RAM	Restricted in Ability to Manoeuvre
RM-ANOVA	Repeated Measured – Analysis of Variance
ROC	Remote Operations Centre
RQ	Research Question
SA	Situation Awareness
SAGAT	Situation Awareness Global Assessment Technique
SART	Situation Awareness Rating Technique

---

---

**Abbreviation**

---

SAT	Situation Awareness-based Transparency
SLR	Systematic Literature Review
SO	Stand-on
SOA	Speed Of Advance
SOLAS	Safety Of Life At Sea
SPAM	Situation Present Assessment Technique
TA	Task Analysis
TARG	Training and Assessment Research Group
TCPA	Time to Closest Point of Approach
UAV	Unmanned Aerial Vehicle
UNCTAD	United Nations Conference on Trade and Development
USN	University of South-eastern Norway
Z	Stationary object

---

## Definition of terms

Term	Definition
Agent	An autonomous entity having goal-directed behaviour in an environment using observation through sensors and execution actions through actuators (Russell & Norvig, 2022)
AI-enabled system	A system that contains or relies on one or more AI components (DNV, 2023)
Artificial intelligence	System capability of an engineered system to acquire, process and apply knowledge and skills (ISO, 2021)
Automation	The full or partial replacement of a function previously carried out by the human operator (Parasuraman et al., 2000)
Autonomy	Ability of a system to work for sustained periods without human intervention (ISO, 2020b)
Display equipment	Device capable of representing information visually (IEC, 2022)
Function	Specific purpose or objective to be accomplished, that can be specified or described without reference to the physical means of achieving it (IEC, 2019)
Human Factors	Scientific discipline concerned with the understanding of interactions among human and other elements of a system, and the profession that applies theory, principles, data and methods to design in order to optimize human well-being and overall system performance (ISO, 2011)
Human Machine Interface (also: user*, human-system*, man-machine*)	Set of all the components of an interactive system that provide information and controls for the user to accomplish specific tasks with the interactive system (ISO, 2020a)
Machine learning	Process using algorithms rather than procedural coding that enables learning from existing data in order to predict future outcomes (ISO, 2017)
Operator	Individual whose primary duties relate to the conduct of monitoring and control functions, usually at a control workstation, either on their own or in conjunction with other personnel both within the control room or outside (ISO, 2000)
Situation awareness	Situation awareness is the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future (Endsley, 1995)
Supervisory control	One or more human operators are continually programming and receiving information from a computer that itself closes an autonomous control loop through artificial effectors and sensors to the controlled process or task environment (Sheridan, 1992)
System	Combination of interacting elements organized to achieve one or more stated purposes (ISO, 2011)
Task	Activities required to achieve a goal (ISO, 1998)
Transparency (also: system*, agent*, display*, automation*)	The ability of a system to support understandability and predictability of its current and future actions (Endsley, 2017)
User	Person who interacts with a system, product or service (ISO, 2011)
Validation	Confirmation, through the provision of objective evidence, that the requirement for a specific intended use or application have been fulfilled (ISO, 2015)
Verification	Confirmation, through the provision of objective evidence the specified requirements have been fulfilled (ISO, 2015)



# Table of contents

- 1 Introduction ..... 1
  - 1.1 Background and context..... 1
  - 1.2 Research gaps, objectives, and questions ..... 2
  - 1.3 Structure of the dissertation..... 4
- 2 Theory and background ..... 7
  - 2.1 Achieving safe maritime collision avoidance..... 7
    - 2.1.1 The human factor in avoiding collisions ..... 7
    - 2.1.2 Towards human supervised autonomous collision avoidance..... 8
  - 2.2 Human Factors in supervisory control ..... 9
    - 2.2.1 Human information processing and Situation Awareness..... 10
    - 2.2.2 Challenges to Situation Awareness in supervisory control ..... 13
  - 2.3 Supporting supervisory control with transparency..... 14
    - 2.3.1 Definitions of transparency ..... 14
    - 2.3.2 Transparency and explainability..... 16
    - 2.3.3 Frameworks and models for transparency..... 18
- 3 Research methodology ..... 21
  - 3.1 Scientific research philosophy..... 21
  - 3.2 Methodological approach ..... 21
  - 3.3 Research methods and data analysis process..... 23
    - 3.3.1 Article 1: Systematic Literature Review ..... 23
    - 3.3.2 Article 2: Goal-Directed Task Analysis ..... 24
    - 3.3.3 Article 3: Human information processing model ..... 27
    - 3.3.4 Article 4: Human Machine Interface development ..... 28
    - 3.3.5 Article 5: Controlled experiment..... 29
  - 3.4 Recruitment process ..... 33
  - 3.5 Validity, reliability, and quality of research..... 34
  - 3.6 Research ethics..... 37
- 4 Results ..... 39
  - 4.1 Article 1: Understanding the overall context..... 39
    - 4.1.1 Domains, models, interaction types, and operationalisations..... 39
    - 4.1.2 Transparency, SA, mental workload, and task performance ..... 39
  - 4.2 Article 2: Identifying requirements for supervision ..... 40
  - 4.3 Article 3: Structuring requirements using the PSW model ..... 42

4.4	Article 4: Developing Human Machine Interface concepts .....	43
4.4.1	Developing traffic situations .....	43
4.4.2	Developing transparent HMIs for collision avoidance.....	44
4.5	Article 5: Evaluating with controlled experimentation .....	47
5	Discussion.....	51
5.1	Theoretical reflections .....	51
5.1.1	Situation Awareness .....	53
5.1.2	Mental workload.....	54
5.1.3	Task performance .....	55
5.1.4	Complexity .....	55
5.2	Practical reflections .....	56
5.2.1	Designing for transparency.....	56
5.2.2	Assessing transparency in autonomous ship concepts .....	59
5.3	Methodological limitations and reflections .....	60
5.3.1	Systematic literature review .....	61
5.3.2	Goal-Directed Task Analysis .....	61
5.3.3	Human information processing model .....	62
5.3.4	Human Machine Interface development.....	63
5.3.5	Controlled experiment .....	64
5.4	Recommendations for future work.....	65
6	Conclusions .....	67
6.1	Revisiting the research questions .....	67
6.2	Contributions.....	68
6.3	Concluding remarks .....	71
7	References .....	73
	Appendix A – Coupling the Goal-Directed Task Analysis, PSW model, and HMI .....	85
	Appendix B – Guide to Human-Machine Interface and symbology .....	91
	Appendix C – Examples of transparency levels.....	99
	Appendix D – Traffic situations used in the experiment.....	103
	Appendix E – Generic SAGAT queries .....	119
	Appendix F – NASA-TLX.....	127
	Appendix G – Ranking transparency levels.....	129
	Appendix I – Publications .....	131
	Appendix J – Statements of co-authorship.....	231

# **PART I**





# 1 Introduction

## 1.1 Background and context

Maritime transport is responsible for the majority of global trade. An estimated 80% of the trade volume is carried out by a global shipping fleet of over 100,000 vessels, with a manning force of more than 1.8 million seafarers, and a vessel market size of 168 billion USN (Fortune Business Insights, 2021; UNCTAD, 2023b, 2023a). However, despite decades of growth, the industry is faced with several key challenges. For example, the prospect of global disruption as a consequence of climate change is pushing the maritime industry towards developing sustainable solutions (IMO, 2023). Also, the lack of- and increasing demand for certified seafarers in the world threatens to affect manning levels and the subsequent flow of goods (BIMCO, 2021). Furthermore, with more than 800 total ship losses in the last decade, and more than 3000 shipping incidents reported in 2022 only, improving maritime safety is a continuous and ongoing endeavour (Allianz, 2023). Moreover, increasing geopolitical tension, cybersecurity risks, and changes in legislation and regulations threaten to affect global trade (Allianz, 2024). To address these challenges, the maritime industry is looking for ways to reduce its environmental footprint, become more attractive to personnel, improve its safety record, and enhance its resilience to adverse conditions whilst maintaining profitability. One of the areas that has received considerable attention as a candidate for decreasing operating costs, increasing efficiencies, reducing reliance on qualified seafarers, and enhancing safety, is autonomous shipping (Kretschmann et al., 2017; Kurt & Aymelek, 2022).

In the last decade, the industry has painted a vision of a future where autonomous ships, powered by artificially intelligent (AI) agents, execute their voyages without direct human involvement (Rolls Royce, 2016). These ships, without the need for human support facilities, may use novel designs, use cleaner fuels, and sail at more efficient speeds, thereby saving fuel and personnel costs (Kretschmann et al., 2017; Kurt & Aymelek, 2022). Furthermore, as removing personnel from sharp-end operations will also remove their exposure to risk, unmanned ships may reduce the number of casualties among seafarers (de Vos et al., 2021; Wróbel et al., 2017). Still, despite these purported benefits, the effect of ship autonomy on overall safety performance is unclear, and highly depends on the operational concept. For example, some developments foresee advanced technologies to be used as decision support tools in traditional navigational contexts (Aylward et al., 2022; Pietrzykowski et al., 2017). More visionary developments delegate seafarers a new role as operators in Remote Operations Centres (ROC) where multiple ships can be monitored and supervised (Alsos et al., 2022; Porathe et al., 2020; Rødseth et al., 2021). Regardless of the concept, in the path towards becoming a green, attractive, and safe industry, future seafarers are expected to interact with increasingly sophisticated and capable systems.

Within autonomous shipping, collision avoidance is an area that has received a significant amount of attention, as solving this complex and multi-faceted task is seen as an important step in realising autonomous shipping (see Akdağ et al., 2022; Zhang et al., 2021 for reviews). Here, success depends on the collision and grounding avoidance (CAGA) system's ability to develop solutions that are compliant with the Convention on the International Regulations for Preventing Collisions at Sea (COLREG; IMO, 1977). These "rules of the road" for seagoing vessels are central to successful collision avoidance manoeuvring. However, one of the primary challenges with COLREG compliant behaviour is the vagueness of how the rules are phrased (Stitt, 2002). For example, performing "apparent" collision avoidance manoeuvres according to "the practice of ordinary seamen", and made

in “ample time” may make sense to experienced sailors, but may be challenging for an algorithm to interpret (Porathe, 2019; Wróbel et al., 2022; Zhou et al., 2020). In addition, notwithstanding advances in machine learning models that may be able to replicate traditional ship behaviour, AI-enabled systems typically employ probabilistic models trained on large datasets to recognize patterns in data. Although such models can be powerful, the association between input and output parameters are not wholly predictable or observable (Christoffersen & Woods, 2002; Littman et al., 2021). Also, systems using machine learning-based algorithms lack a model of causation which limits their capabilities to discovering correlations between variables only, rather than causal relationship (Littman et al., 2021). This means that such systems have limited ability to predict how novel situations may develop, or how situations would have played out under different circumstances (Pearl & Mackenzie, 2018).

As AI-enabled systems “do not know what is possible in the world”, creating reliable and predictable systems deployable in complex environments remains a major challenge (Littman et al., 2021, p. 23). Given these limitations, such systems are currently not adequately fit to operate in novel and complex situations and therefore require careful management and supervision (National Academies of Sciences, Engineering and Medicine, 2022). For the foreseeable future, this means that humans can be expected to play a supervisory role to oversee system performance, direct their functioning, and ensure that their desired utility is achieved (Endsley, 2017). However, because of their probabilistic nature, such systems are less tractable and predictable compared to “traditional” control systems (Hollnagel, 2012). Consequently, humans may find it challenging to understand and evaluate their behaviour (Endsley, 2017, 2023a). Considering how AI-enabled systems may significantly affect how work is organised, including the role of the human herein, this means that new knowledge, methods, and tools, with regards to how humans may be supported in effectively interacting with these emerging technologies, are urgently needed.

## 1.2 Research gaps, objectives, and questions

Early research from studies addressing human supervisory performance in safety-critical domains have suggested that providing insight into a system’s internal information processes can support human supervisors in understanding and predicting its behaviour (e.g., Christoffersen & Woods, 2002; Sheridan & Verplank, 1978). More recently, “agent transparency” (J. Y. C. Chen et al., 2014), “system transparency” (Ososky et al., 2014), “display transparency” (National Academies of Sciences, Engineering and Medicine, 2022), “automation transparency” (Skraaning & Jamieson, 2021), or simply “transparency” has been proposed as a means to enhance understandability and predictability and support human supervision of highly automated systems (Endsley, 2023b; Endsley et al., 2003; Meister, 1999).

Transparency is a design principle based on the notion that when selected system-internal information is made available to the operator, its understandability and predictability is enhanced. That is, when humans interact with autonomous systems, transparency principles can be used to convey the system’s state, its modes, and limitations, and support understandability and predictability regarding its current- and future actions (Endsley, 2017). As the Human-Machine Interface plays a vital role in bridging the gap between autonomous systems and humans, a potential lack of transparency may make it difficult for an operator to grasp the capabilities of the system in terms of what it is doing and why. This, in turn, may affect the operator’s trust in the system (Lee & See, 2004). Conversely, systems which disclose their inner processes should enhance the operators’ ability to assess their performance, calibrate their trust, and encourage appropriate use (Lee & See, 2004), rather than misuse or disuse

(Parasuraman & Riley, 1997). Therefore, the ability of the system to provide feedback about its internal information processing, its decisions, and planned actions is important for supporting supervisory performance (Beck et al., 2007; Endsley, 2017; Endsley & Kiris, 1995; Parasuraman & Riley, 1997). This potential “free lunch” (Wickens, 2018), i.e., the ability of transparency to alleviate some of the challenges of human supervision of autonomous systems without reducing its benefits, warrants further investigation.

Recent studies have investigated the relationship between transparency and human performance variables. Bhaskara et al. (2020) and Rajabiyazdi and Jamieson (2020) identified and compared the evidence for agent transparency in the contemporary literature. The results indicated there is emerging evidence regarding human performance improvements, despite a concern for increased mental workload. However, results were not consistent in terms of the correlation between the degree of transparency and performance variables. In other words, more transparency did not consistently produce improved operator performance outcomes. Furthermore, the authors concluded that the validation efforts for the transparency models have been largely incomplete or have provided inconclusive evidence. Therefore, given that agents are increasingly deployed in safety critical contexts where insufficient performance of the human-machine system may have consequences for safety, ongoing and continuous efforts are needed to understand how agents can be made understandable to human supervisors (National Academies of Sciences, Engineering and Medicine, 2022). Specifically, focus should be given to determining which transparency information best supports SA and operator performance as well as which methods should be applied in designing transparent agents without overloading the supervisor with information (National Academies of Sciences, Engineering and Medicine, 2022). Although the scientific literature demonstrates the potential benefit of transparency, there is a continuous and ongoing need for knowledge, methods, and tools with regards to its applicability as a design principle in safety critical domains. Therefore, this dissertation aims to contribute to generating and advancing knowledge on how human supervisory control can be supported through agent transparency. To achieve this, it aims to answer the following overarching question:

*How does agent transparency support human performance in supervisory control?*

To contribute to answering the main question, specific sub-questions are posed. Each is addressed through several activities and reported in a series of research articles. The first activity aims to establish a broad overview of the relevant scientific literature regarding agent transparency and key human factors variables, identify knowledge gaps, and determine a path forward for the remainder of the dissertation. Combined, this activity aims to answer the following question:

*RQ1: What is the relationship between agent transparency and Situation Awareness, mental workload, and task performance?*

The second research activity aims to establish an understanding of the collision avoidance context by mapping and assessing goals, decisions, and tasks for conventional- and supervised collision avoidance. Furthermore, this activity aims to identify information requirements to make agents, capable of collision and grounding avoidance, transparent to their users. Together, this activity aims to answer the following question:

*RQ2: How is human performance achieved in conventional- and supervised maritime collision avoidance?*

The third research activity aims to understand how a model of human information processing can be used as an approach to define transparency requirements for autonomous systems. By disclosing the system's perceived information, internal reasoning, decisions, and planned actions, this approach aims to explore how this model can function as a framework to organise the information requirements as layers of transparency. Therefore, this activity aims to answer the following question:

*RQ3: How does a model for human information processing form the basis for agent transparency in the ship autonomy context?*

The fourth research activity aims to examine how the organised information requirements from the previous activities can be represented as levels of transparency in realistic transparent HMI concepts. Operationalised through realistic traffic situations and symbology representing the autonomous system's information processing, this activity aims to provide the groundwork for systematically studying the relationship between agent transparency and human performance in the collision avoidance context. Hence, this activity aims to answer the following question:

*RQ4. How should a maritime collision avoidance system be made transparent to a human supervisor?*

Finally, the fifth research activity aims to study the relationship between agent transparency and key human factors variables: SA, mental workload, and task performance. Through a controlled experiment with certified navigators, this activity aims to systematically evaluate the effect of individual levels of transparency and derive a comprehensive understanding of the factors contributing to human performance when supervising autonomous agents. Accordingly, this activity aims to answer the following question:

*RQ5. What is the relationship between agent transparency and Situation Awareness, mental workload, and task performance in maritime autonomous collision avoidance manoeuvring?*

### **1.3 Structure of the dissertation**

In this dissertation, the first chapter provides the general background to the research, the research gaps, objectives, and research questions (this chapter). The second chapter provides the theoretical foundation to the work, addressing central concepts, theories, and relevant contexts. The third chapter depicts the research methodology, including chosen philosophies, methods, considerations regarding validity, reliability, quality of research, and ethics. The fourth chapter discusses the results and discussions from each of the research articles. Chapter five presents the overall conclusions, scientific and practical contributions of this work. Finally, relevant details are provided in the appendices, including the five publications integral to this dissertation.

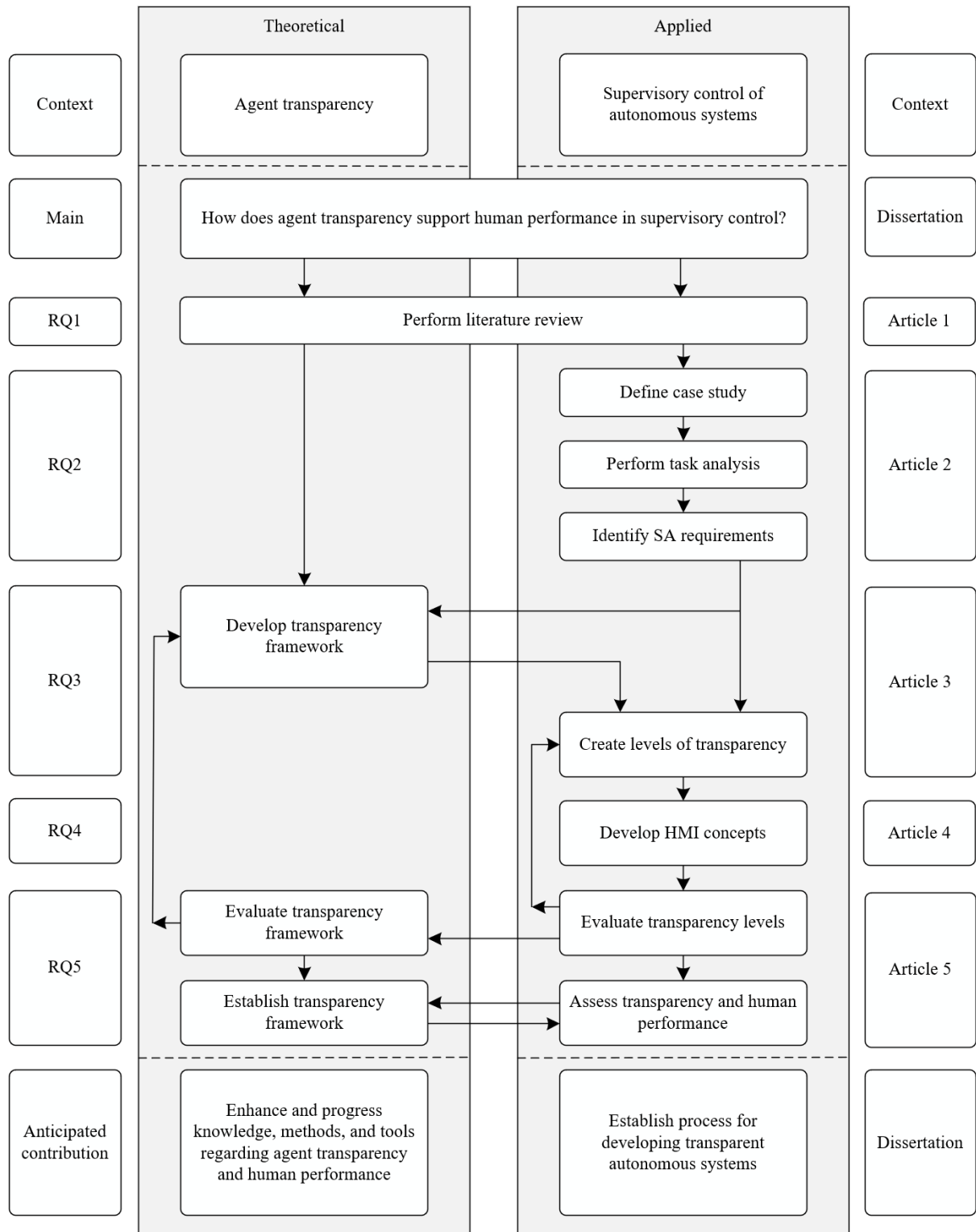


Figure 1. A schematic overview of this dissertation.

This dissertation aims to combine a theoretical and applied approach in answering the overall research question. As depicted in Figure 1, the main research question is answered through a series of activities, each addressed through its individual research questions, elucidating and combining theoretical insights with practical applications. The literature review provides the theoretical foundation on which the further activities are based. Here, the state-of-the-art of the scientific evidence on agent transparency and human performance is mapped (RQ1). Subsequently, the GDTA focuses on nautical

collision avoidance manoeuvring, based on in-situ observations and interviews, and where subject-matter experts (navigators) play a vital role. This analysis provides a set of SA requirements on which subsequent activities are built (RQ2). Next, a theoretical model is adapted to the nautical collision avoidance context, such that the SA requirements are structured and transformed into transparency requirements for human supervised autonomous collision and grounding avoidance systems (RQ3). Furthermore, graphical HMI concepts are developed, with the aid of a navy-certified navigator, to transform the transparency requirements into graphical symbology for presentation onto a radar display depicting realistic traffic situations (RQ4). Finally, levels of transparency are developed that are experimentally evaluated with certified navigators on key human factors variables (RQ5). The culmination and integration of this body of work is ultimately used to contribute to answering the overall research question.

The results of the research activities performed in this dissertation are presented in five peer-reviewed research articles in recognised journals and conferences (see Table 3). The work in this dissertation follows a sequential progression from the start of the research in Article 1 to the experimental results in Article 5.

*Table 3. Appended articles.*

<b>No.</b>	<b>Title</b>	<b>Authors</b>	<b>Journal/Conference</b>	<b>Status</b>	<b>RQ</b>
1	Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review	Koen van de Merwe Steven Mallam Salman Nazir	Human Factors	Published	1
2	Supporting human supervision in autonomous collision avoidance through agent transparency	Koen van de Merwe Steven Mallam Salman Nazir Øystein Engelhardtzen	Safety Science	Published	2
3	Towards an approach to define transparency requirements for maritime collision avoidance	Koen van de Merwe Steven Mallam Øystein Engelhardtzen Salman Nazir	Proceedings of the Human Factors and Ergonomics Society Annual Meeting	Published	3
4	Operationalising Automation Transparency for Maritime Collision Avoidance	Koen van de Merwe Steven Mallam Øystein Engelhardtzen Salman Nazir	TransNav, International Journal on Marine Navigation and Safety of Sea Transportation	Published	4
5	The Influence of Agent Transparency and Complexity on Situation Awareness, Mental Workload, and Task Performance	Koen van de Merwe Steven Mallam Salman Nazir Øystein Engelhardtzen	Journal of Cognitive Engineering and Decision Making	Published	5

## 2 Theory and background

This section introduces and elaborates on each of the knowledge areas introduced in Chapter 1: collision avoidance in the maritime context, human factors in supervisory control, and agent transparency.

### 2.1 Achieving safe maritime collision avoidance

#### 2.1.1 The human factor in avoiding collisions

Considering the international nature of the shipping industry, safety at sea is largely regulated by the International Maritime Organisation (IMO). The IMO is a specialised agency of the United Nations which is, amongst others, responsible for developing regulations to improve safety of the shipping industry (IMO, 2019). Its scope embodies a variety of maritime safety aspects, including technical requirements related to ship stability, fire safety, life-saving appliances, ship design, operational requirements for search and rescue, radiocommunication, and navigation. Within these requirements, the “human element” is recognized as central for safety of life at sea. For example, in navigation, the International Convention for the Safety of Life at Sea (SOLAS) stipulates the role of the ship’s Master to ensure that the intended voyage is planned appropriately, using the latest information, to allow for safe passage of the ship (IMO, 1974:2000). In addition, the International Convention on Standards of Training, Certification, and Watchkeeping stipulates the watchkeeping arrangement and principles to be observed for seagoing ships. This includes that a “safe and continuous watch or watches appropriate to the prevailing circumstances and conditions are maintained on all seagoing ships at all times” (IMO, 1978:2010; Chapter VIII/2). To support the crew in this, the regulations dictate that measures shall be in place with the aim to minimize the potential for human error by ensuring that the bridge design facilitates the navigational tasks to be performed by the bridge crew, that convenient and continuous access to essential information is available, and that this information is presented in a clear and unambiguous manner (IMO, 1974:2000). In case of vessel encounters during the ship’s voyage, the Convention on the International Regulations for Preventing Collisions at Sea stipulates the rules for how navigators shall perform collision avoidance manoeuvring (IMO, 1977).

Collision avoidance is internationally regulated through the collision avoidance rules, i.e., the COLREGs. These rules were developed to provide the “rules of the road” for sea-going ships. The aim of the COLREGs is to increase safety by providing a set of rules that make for predictable vessel traffic behaviour. With the rules, navigators are provided with guidance on how to manage vessel encounters such that collisions can be prevented. The COLREGs are made up of 41 rules divided into six parts: Part A - general aspects, part B - the steering and sailing rules, part C - lights and shapes, part D - sounds and light signals, part E - exemptions, and part F - verification of compliance. For the purpose of this dissertation, parts A and B are the most relevant as these provide detailed descriptions regarding collision avoidance manoeuvring.

Part A describes the general application of the rules and makes explicit the ultimate responsibility of the ship’s navigator to avoid dangers of navigation and collision. Part B provides the steering and sailing rules for vessels in any condition of visibility and for vessels in sight of one another. Here, the rules state the requirements for how to manoeuvre and determine priority when encountering other vessels on collision course. That is, Rule 5 states the requirements for maintaining a proper lookout to ensure a full appraisal of the situation is made such that vessels and objects are detected early.

Furthermore, Rule 6 stipulates the requirements for determining the speed appropriate to the circumstances. Rule 7 describes the process for determining collision risk, and Rule 8 describes the actions for avoiding collision, including the execution of route and/or speed changes. Rules 11 to 18 describe the rules for collision avoidance when vessels are in sight of one another and Rules 13 to 15 specifically deal with three typical ship encounter situations: overtaking, head-on, and crossing. Finally, Rules 16 to 18 describe the circumstances under which a ship shall give-way (perform an avoidance manoeuvre), when it shall stand-on (hold course and speed), and the hierarchy of responsibilities between vessels.

As much as the rules were developed to provide specific guidance for navigators in handling ship encounters, their interpretation is not straightforward, and experience is required when applying them (Stitt, 2002; Wang et al., 2021; Weber, 1995). For example, Rule 8 “Action to avoid collision” states: “any action to avoid collision shall be taken in accordance with the Rules of this Part and shall, if the circumstances of the case admit, be positive, made in ample time and with due regard to the observance of good seamanship” (IMO, 1977). This rule can be interpreted as when a navigator identifies that own ship engages in a collision scenario with another ship, the avoidance action shall be performed without delay, and be clearly visible to the other ship. However, this rule does not specify what “ample time” means, nor the interpretation of “positive”, or “good seamanship”. Furthermore, Rule 2 “Responsibility” states: “In construing and complying with these Rules due regard shall be had to all dangers of navigation and collision and to any special circumstances, including the limitations of the vessels involved, which may make a departure from these Rules necessary to avoid immediate danger” (IMO, 1977). Here, the rule can be interpreted as to allow the navigator to take any measure necessary to avoid immediate danger, even if this means deviating from the COLREGs. Clearly, these two examples provide evidence of the interpretative nature of the COLREGs and the experience necessary to apply them. Nevertheless, the COLREGs applies “to all vessels upon the high seas and in all waters connected therewith navigable by seagoing vessels” (IMO, 1977) and is thereby an essential element in ensuring safety in maritime navigation; now and on the road towards autonomous ships (IMO, 2021).

### **2.1.2 Towards human supervised autonomous collision avoidance**

Rules and regulations are currently under development at national and international levels to prepare for, and accommodate, the arrival of ships with autonomous ship functions. That is, IMO is developing a code to regulate the operation of maritime autonomous surface ships (IMO, 2022). In Norway, the Norwegian Maritime Authority (NMA) has published guidance for documentation requirements and principles to be applied when seeking approval for operating unmanned or partially unmanned operations (NMA, 2020). Furthermore, several classification societies are working on developing guidelines for the approval of autonomous and remotely controlled ships (e.g., American Bureau of Shipping, 2022; Bureau Veritas, 2019; DNV, 2021). Even though developments towards established rules and regulations are ongoing, national authorities may already provide temporary approval of autonomous ship concepts for testing purposes and limited to national waters only (e.g., NMA, 2020), based on the current IMO regulations for the approval of “alternatives and equivalents” (IMO, 2013).

The “guidelines for the approval of alternatives and equivalents as provided for in various IMO instruments” (IMO, 2013, p. 1) provides a temporary alternative path to approving ships for which the standard regulations are lacking or are insufficient, e.g., for autonomous ships. Central to this



approach is a risk-based method based on which a ship's design is evaluated against overall risk criteria. This means that, through a structured design and approval process, the submitter of an autonomous ship concept provides data and documentation to the relevant regulatory body for evaluation. Assuming sufficient evidence for risk mitigation is provided, this approach allows for deploying technological solutions based on alternative designs. As such, if an equivalent level of safety can be demonstrated, e.g., in terms of life safety criteria, the environment, and damage to ship structures and systems, approval may be provided on a case-by-case basis (although for operation in the national waters of the given administration only).

This approach provides the basis for guidelines published by classification societies on autonomous and remotely operated ships (e.g., DNV-CG-0264; DNV, 2021). For this approach, the class society plays the role of intermediate between autonomous ship concept submitter and national authorities and performs concept reviews, provides the submitter with specifications, analysis, and test scopes, and performs reviews of the submitter's tests. Class societies may also perform their own tests to independently verify claims made by the submitter. Throughout this process, the national authority provides the framework for approval basis and performs design and test reviews. Consequently, this interaction between concept submitter, classification society, and national authority aims to support and facilitate the final approval of the ship autonomy concept.

Central to the approval of autonomous ships is that the degree of autonomy and the division of function control is contingent on the principle that “the combined human/machine capabilities [...] should be the same or better than the conventional capabilities. This in order to achieve an equivalent or better level of safety” (DNV, 2021, p. 52). This means that the concept of operations provided by the submitter should clearly describe the operational tasks that the vessel will perform, and the extent to which these will be automated. For example, some tasks may be fully automated with no human involvement, some tasks may be highly automated, but require human supervision, whereas other tasks may be executed by humans with automation as support. Considering the potential changes in tasks, roles, and responsibilities for humans within the ship autonomy domain, it is important to support humans in this and mitigate any potential challenges to human performance.

In the context of collision avoidance, this means that the concept submitter should provide evidence indicating that the overall system delivers an equivalent level of safety, or better. Furthermore, “special attention should be placed on the timing aspects, and the ability of the human to establish sufficient situational awareness so that correct actions can be taken within reasonable time” (DNV, 2021, pp. 27, 28). For these tasks, the interface between the system and the human plays a vital role in communicating the required information for effective human supervision. In situations in which humans are expected to intervene and assume control of the vessel because of “system-limitations or failures, [...] ample time [should be allowed] for the human to get the required situational awareness in order to be able to make good decisions” (DNV, 2021, p. 42). Hence, to ensure equivalent safety in concepts where autonomous functions perform tasks previously performed by humans, special focus should be given to supporting the cognitive processes required for adequate and effective human supervisory control (Sheridan, 1992).

## **2.2 Human Factors in supervisory control**

The construct “Situation Awareness” is a key variable associated with supervisory control (Endsley, 1995). Therefore, a discussion around this construct, including factors affecting it, is useful within the

context of this dissertation such that its connection to human supervisory control and autonomous collision avoidance is illustrated.

### 2.2.1 Human information processing and Situation Awareness

Within the field of applied cognitive psychology, scientists have aimed to create conceptual models to understand how humans process information, perform cognitive tasks, and make decisions. This knowledge can be used to develop systems that support operators in achieving their goals. Figure 2 depicts a model conceptualising how humans perceive and analyse information, make decisions, and perform actions in four discrete stages: information acquisition, information analysis, decision selection, and action implementation. (Lee et al., 2017; Parasuraman et al., 2000). In this model, the first stage represents the acquisition and registration of multiple sources of information, pre-processing of information, orienting of sensory receptors, and selective attention to information sources. The second stage represents conscious perception and manipulation of processed and retrieved information in working memory. Here, based on the information from the initial stage, associations between information elements and inferences are made prior to generating decisions and conclusions. The third stage represents the stage where decision alternatives are created and decided upon. The fourth stage represents the action implementation resultant of the decision. As feedback from actions is perceived again as stimuli for the information processing system, the cycle starts anew.

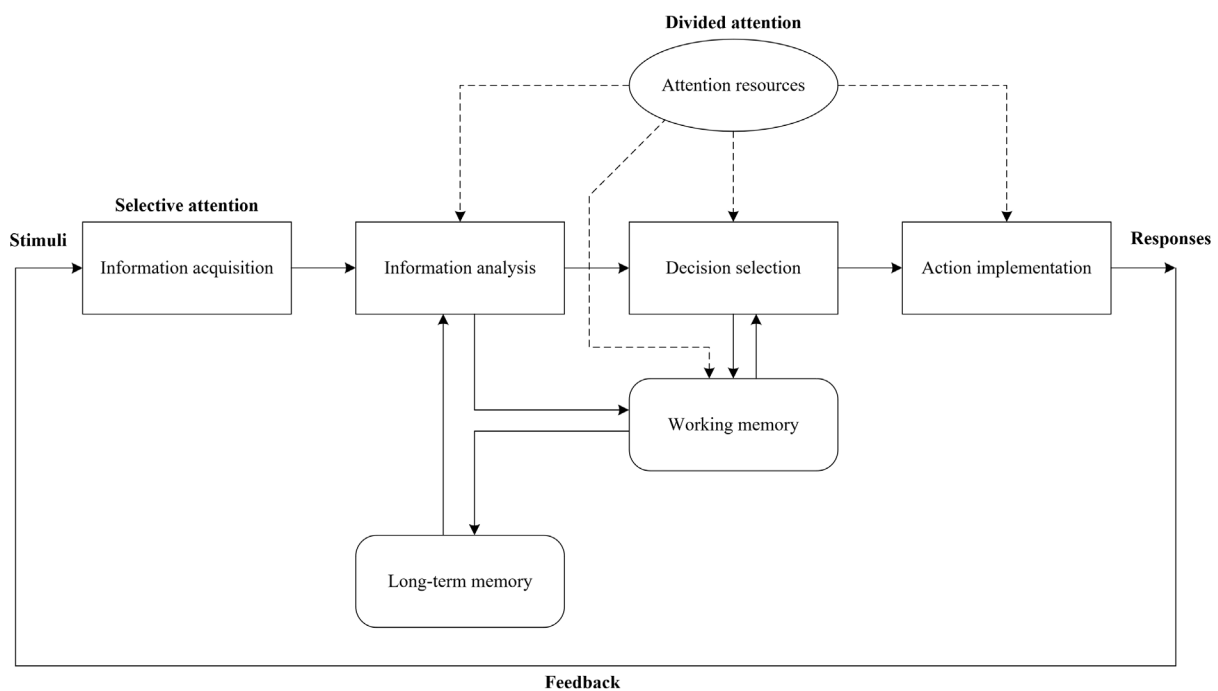


Figure 2. A model for human information processing (adapted from Lee et al. (2017) and Parasuraman et al. (2000)).

In this model, the roles of working-, long-term memory, and attentional resources are key factors in the perception and analysis of information, decision making and response execution. Working memory refers to the transient and vulnerable form of information storage capable of containing a limited amount of information. Here, information in the form of images, symbols, locations, etcetera, can be mentally maintained and manipulated (Wickens & Carswell, 2021). Considering the finite space of working memory, the amount of information that can be processed by working memory is limited (Cowan, 2010; Miller, 1994). Therefore, maintaining information active in working memory,

whilst limiting distraction by other competing information elements, is important for achieving and maintaining performance (Kintsch (1970), in Haberlandt, 1999). However, it is also cognitively demanding. In terms of attentional resources, selective attention focuses on relevant incoming stimuli, focused attention typifies the effort to maintain focus on relevant stimuli without becoming distracted, whilst divided attention enables the processing of multiple stimuli across the processing stages: information acquisition, analysis, decision selection, and action implementation (Styles, 1997; Wickens & Carswell, 2021). Since attentional resources are also limited, dividing attention between more than one activity is mentally demanding too (Kahneman, 1973). Consequently, as demands increase, the quality of action execution may degrade (or fail), more efficient and less resource consuming methods may be used, and focus may be shifted to higher priority tasks rather than lower priority ones (Hancock et al., 2021). As demands exceed capacity, performance will break down (Wickens, 2008). This implies that attention and memory processes play important roles in information processing in dynamic environments.

In complex and constantly evolving environments, such as shipping, action execution is highly dependent on the human's ability to make accurate and timely decisions. The operator's constant awareness of what is happening around him/her and what this information means now and in the future is called "Situation Awareness" (Endsley et al., 2003). A common definition of SA is "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1988, 1995). In this definition, SA is divided in three processing steps:

- Level 1: The perception of elements in the environment, within a volume of time and space
- Level 2: The comprehension of their current status
- Level 3: The projection of their status in the near future

In terms of information processing, Endsley's model of SA includes explicit processing stages for information perception, comprehension, and future predictions (as depicted in Figure 3). In this model, SA can be described in terms of the *knowledge* that is produced, and the *processes* to produce that knowledge, in order to make decisions and perform actions (Endsley, 1995; van Doorn et al., 2017). For each level of SA this means:

- Level 1: Perceptual knowledge – unprocessed knowledge about elements in their environment
- Level 2: Comprehended knowledge – an understanding of their meaning and relationships
- Level 3: Projected knowledge – insight into the predicted future state, given the dynamic environment, in relation to operational goals

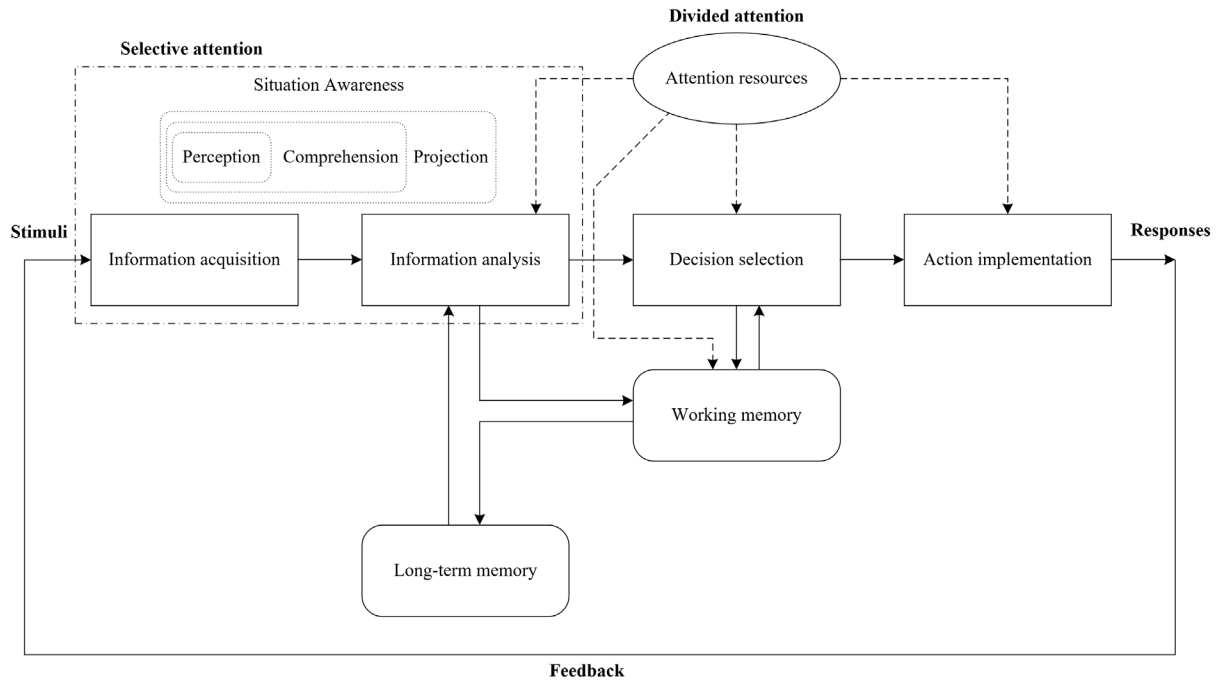


Figure 3. A model for human information processing with levels of Situation Awareness superimposed (adapted from Lee et al. (2017), Parasuraman et al. (2000), and Endsley (1995)).

The continuous cognitive processes that operators perform to achieve, acquire, and maintain SA knowledge is called “SA assessment” (Endsley, 1995). This includes providing selective attention to information, dividing attention between processing steps, maintaining relevant information elements in working memory, and applying mental models stored in long-term memory. In this context, mental models are conceptual analogues of the external world which allows for its understanding and prediction (Mogford, 1997). In terms of SA, mental models are important for directing attention to relevant information, developing an understanding of this information, and projecting this information into the future. Mental models provide an essential mechanism in human’s ability to integrate substantial amounts of information, form an understanding of their meaning, and project their effects in the future. In the context of collision avoidance, this means that navigators need to have mental models of their task environment and the systems they operate to actively perceive, make sense of, and predict the status of potential objects such that quality decisions can be made on how to avoid collisions.

For autonomous collision avoidance, this implies that operators also need effective mental models of the system under supervision. This includes understanding the system’s limitations, capabilities, reliability, functioning, and an understanding of its logic and components (Endsley, 2023b). Therefore, the operator’s mental model of the autonomous collision avoidance system is of particular relevance in understanding if the system is behaving correctly or whether intervention is needed. This means that, to effectively assess the performance of the system, operators need to have “task SA” (i.e., SA of the collision avoidance task), as well as “system SA” (i.e., SA of the collision avoidance system).

For human supervised autonomous collision avoidance, system SA implies that the operator has an understanding of how the system is performing (Endsley, 2023b). For example, if the system encounters a situation outside its operational design domain, the operator may be required to intervene (Rødseth et al., 2021). This means that operators should be aware of whether the system is able to

perform its function in the current situation, the sufficiency of what the system knows about the collision situation, and the impact of the system's actions on the situation. However, the degree to which an operator can achieve SA depends on individual-, task-, and system factors. For example, individuals may differ in their ability to achieve and maintain SA due to innate abilities or differences in training. Also, differences in tasks complexities influence an operator's ability to achieve and maintain SA. Furthermore, poorly designed system interfaces may hinder the operator in perceiving and evaluating critical information. Moreover, the effort associated with maintaining task- and system SA in parallel may lead to increased mental workload due to limited availability of attentional resources and working memory. Finally, the fact that operators are less involved in the system's information and decision-making loop has consequences for the effectiveness of human oversight (Endsley et al., 2003).

## 2.2.2 Challenges to Situation Awareness in supervisory control

The “ironies of automation” (Bainbridge, 1983), the “out-of-the-loop (OOTL) performance problem” (Endsley & Kiris, 1995), and the “automation conundrum” (Endsley, 2017) are three human performance challenges commonly associated with supervisory control of automation. The ironies in automation are described as the paradoxical expectations system designers have between an *envisioned* automated system, where “unreliable” human behaviour is replaced by “reliable” automation, and an *implemented* system, where humans are given (an arbitrary set of) left-over tasks unable to be automated in addition to compensating for system reliabilities (Bainbridge, 1983). Although automation undisputedly leads to performance improvements, it may also lead to complacency on the part of the operator (the assumption that “all is well”; Parasuraman & Manzey, 2010; Wickens et al., 2015), automation bias (the assumption that the system is probably right; Mosier & Skitka, 1996), reduced vigilance (because of depletion of mental resources ; Finomore et al., 2013; Warm et al., 1996), and misuse and disuse of the system (because of over- and underreliance on the system; Parasuraman & Riley, 1997). Finally, when automation fails, humans may not be completely up to date with the current state of the system, resulting in unreasonably high workload in trying to recover the situation (Onnasch et al., 2014). That is, given the limited information processing capacity of humans, because of attentional and memory limitations, when task demands exceeds cognitive supply mental overload occurs and task performance is affected (Wickens et al., 2013).

As discussed earlier, decision making and action execution depends on the operator's ability to perceive, interpret, decide, and act in a continuously evolving information loop. When automation takes over (part of) the information processing activities, operators are no longer part of the loop. As information perception, comprehension, and projection are essential elements in obtaining and maintaining SA, this means that being outside of this loop has consequences for SA (Endsley & Kiris, 1995). For supervised collision avoidance, where system oversight depends on a comparison between task- and system SA, challenges can be expected when the operator is not able to “follow along” what the system was doing when it failed. Having to take over from a failing system can lead to high mental workload, when suddenly being confronted with the responsibility of the control task with limited time to perform it (Onnasch et al., 2014). Considering that it takes mental effort to acquire SA knowledge and continuously perform SA assessment, high workload may interfere with obtaining and maintaining system- and task SA. Although sufficient mental capacity is not a guarantee for good SA, high mental workload due to tasks competing for the same mental capacities may lead to the operator only paying attention to a subset of information which could result in sub-optimal SA (Endsley, 1995).

Supervisory control is also affected by the reliability and robustness of the system. For system with high reliability, a loss of skills, as a result of a lack of experience with manual control, may aggravate the operator's ability to adequately supervise the system (Endsley & Kiris, 1995). A conundrum exists where humans are less likely to take over manual control of a system, the more reliable and robust this system becomes (Endsley, 2017). As human tasks shift towards supervisory control with the advancement of technology, it may become increasingly difficult for humans to fully understand what the system is doing. In contrast with active information processing in manual control, supervisory control typically involves passively monitoring a system (Metzger & Parasuraman, 2001). Here, the lack of cognitive engagement in the task leads to reduced information processing performance and ability to retain critical task information in working memory (Endsley, 2017). Also, a change in feedback provided by the system affects the human's ability to understand its decisions and actions, especially when essential elements are occluded, not available, or lost in the information noise (Moacdieh & Sarter, 2017).

Although a range of design recommendations have been developed on how to support human-automation interaction (Endsley, 2017; Endsley et al., 2003; Sheridan, 2021; Wickens & Carswell, 2021), adequate and appropriate feedback is a central element for humans to create mental models of the system under supervision (Norman, 1990). However, what constitutes "adequate and appropriate feedback" depends on the function and task distribution between humans and systems, and the operational context in which the system is deployed. For collision avoidance systems, feedback should be compatible with the operators' information processing steps by, for example, providing SA knowledge directly to the operator (van Doorn et al., 2021). This means that, for operators given the task of supervising collision avoidance systems, it should provide relevant information in a way that does not interfere, but rather support, the cognitive processes needed for supervision. That is, depending on which of the information processing steps are automated and the level of sophistication, a collision avoidance system may provide information about what it perceives in its vicinity, its interpretation of this information related to its goals, and its proposed solution. Therefore, notwithstanding optimisation of individual-, task-, and other factors, design should focus on facilitating system SA by considering the amount of information made available, the degree of integration in interfaces needed for task SA, and the degree of information competition as a consequence of maintaining task- and system SA.

## 2.3 Supporting supervisory control with transparency

### 2.3.1 Definitions of transparency

In their work in the context of undersea teleoperations, Sheridan and Verplank highlighted the need for operators to understand "when the computer activity should be apparent to the operator" to be able to monitor and diagnose its behaviour (1978, pp. 9–4). Also, Norman (1990) highlighted the need for adequate feedback to support human supervisory performance to alleviate some of the performance issues associated with human supervision. Furthermore, Christoffersen and Woods (2002) discussed the need for observability and directability in making systems cooperative, especially for systems with high degree of automation. Here, they discussed that feedback to humans should be event based (highlighting changes and events), future oriented (including anticipatory reasoning), and pattern-based (allowing for quick detection of abnormalities). Finally, Lee and See (2004) discussed the need for systems to display information about its purpose, processes, and how it performs in relation to building trust. Based on these initial ideas, the concept of "transparency" (Endsley et al., 2003), "agent

transparency” (J. Y. C. Chen et al., 2014), “system transparency” (Ososky et al., 2014), “display transparency” (National Academies of Sciences, Engineering and Medicine, 2022), or “automation transparency” (Skraaning & Jamieson, 2021) has emerged as a strategy to support humans in supervisory control.

In the literature on transparency several definitions can be extracted. First, some researchers define transparency as a property of the interface, with emphasis on observability and predictability of system behaviour (see Figure 4). Second, several research have expanded on this definition by addressing bi-directional communication between human and system, thereby allowing for observability and predictability of humans by systems (see Figure 5). Finally, some researchers have interpreted transparency as an emergent property between teams of humans and systems (see Figure 6).

Meister (1999) argued that by visualizing system internal functioning, human users may appreciate the meaning of the system’s current and future actions. Here, transparency refers to “the extent to which internal system functioning is made apparent to the human operator” (1999, p. 136). J.Y.C. Chen et al. (2014, p. 2) defined transparency as the “descriptive quality of an interface pertaining to its abilities to afford an operator’s comprehension about an intelligent agent’s intent, performance, future plans, and reasoning process”. Furthermore, in her review of human-automation research, Endsley defined transparency as “making apparent what the system is doing, why it is doing it, and what it will do next” (2017, p. 19). In this definition, the intention of transparency is to provide operators with understandability of its actions, and predictability of its future actions. Finally, Skraaning and Jamieson (2021, p. 1) describe transparency as a “design principle espousing that the responsibilities, capabilities, goals, activities, and/or effects of automation should be directly observable in the human-system interface”. Common to these definitions is that transparency is seen as a property of the system interface. Here, transparency is a design principle that can be applied in the development process of creating human machine interfaces of autonomous systems. In addition, in these definitions transparency is limited to one-way communication between systems and humans (see Figure 4). That is, transparency is a design principle applied to systems such that their users can observe and predict the behaviour of the system.

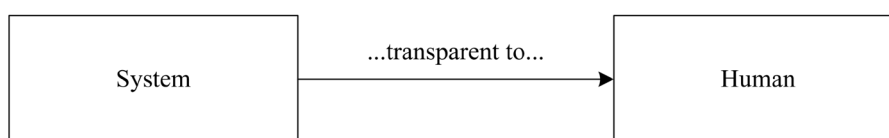


Figure 4. Transparency as a property of the interface to provide observability and predictability of system behaviour.

Some researchers have expanded the scope of transparency to include bi-directional communication between systems and humans. Lyons (2013) developed a model for human-robot interaction describing transparency as information a system needs to present to users prior, during, or after interactions. In addition, the system also needs to be able to have an awareness of the human’s cognitive states and be able to communicate this back to the human (see Figure 5). This “robot-to-human” and “robot-of-human” transparency are seen as essential elements in effective human-robot teaming. Thus, transparency is described as a means to establish shared intent and shared awareness between humans and systems.

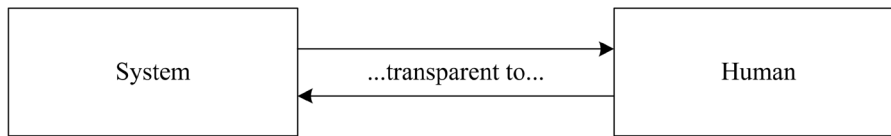


Figure 5. Transparency as bi-directional communication between system and human.

J.Y.C. Chen et al. (2018) expanded on their original definition of transparency to include bi-directional communication of information. In addition, the model of transparency was broadened to include teams of humans and systems working together to achieve a common goal. Here, “both the human and the agent share their goals, reasoning, and projections to achieve their goals as a team, [this way] both the human and the agent maintain transparency regarding their contributions to a shared task. Interestingly, here transparency is not solely defined as a property of the system interface, but (also) as an *activity* that needs to be continuously performed by actors working towards a common goal (see Figure 6). Similarly, Ososky et al. (2014, p. 2) describe that transparency, “within the domain of collaborative robotics, is not solely a characteristic or feature of a robotic asset; rather, it is an emergent characteristic of the human–robot system”. Thus, by performing collaborative activities towards a common goal, exchanging goals, reasoning, and intent, transparency *emerges*.

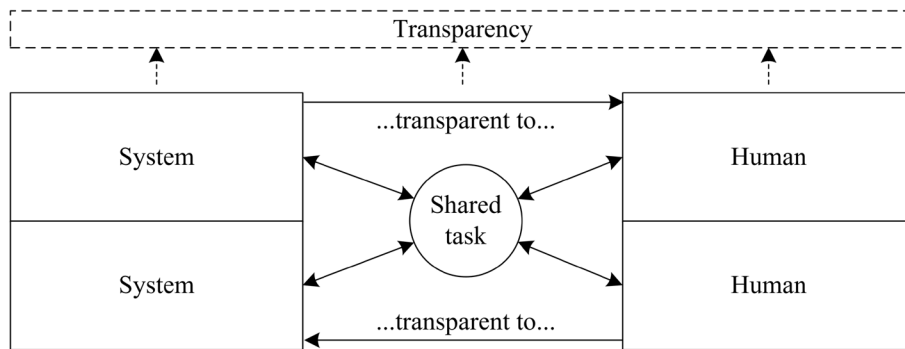


Figure 6. Transparency as an emergent property between systems and humans collaborating on a shared task.

Within the autonomous shipping domain, activities point towards automating tasks currently performed by navigators. Here, humans are delegated supervisory control tasks where monitoring, adjusting, and supervising autonomous ships are central elements. Although future technological developments may include monitoring of human performance as input to agents, this research does not assign this role to the collision avoidance system. Therefore, this dissertation adopts the first viewpoint of transparency, as depicted in Figure 4.

### 2.3.2 Transparency and explainability

When discussing AI-enabled systems, transparency and explainability are frequently mentioned in tandem. When operators interact with agents that provide recommendations or perform actions that have safety critical consequences, insight into the agent’s reasoning is important in effective supervisory control (Warden et al., 2019). Therefore, such agents should be able to provide “explanations” and be “transparent” about their decisions and actions. However, despite appearing to have the same meaning, there are important distinctions between these two concepts.

Explainability aims to support understandability of AI-enabled systems by providing operators with explanations of how agents derive their conclusions or recommendations. For example, an agent may



provide the background for its recommendations by providing insight into which elements were included in its analysis, their relative importance, and how these relate to its recommendation. Thus, explainability “provides information in a backward-looking manner on the logic, process, factors, or reasoning upon which the system’s actions or recommendations are based” (National Academies of Sciences, Engineering and Medicine, 2022, p. 31). In other words, explainability contributes to understanding the logic used by the agent, its capabilities, and limitations by providing *retrospective* information about its processing behind its decisions. This way, explainability supports the operator’s comprehension of how the agent works, when it will work, and when it will not work (Endsley, 2023b). Based on this understanding, operators can develop accurate mental models of the agent (see Figure 7).

Transparency “provides a real-time understanding of the actions of the AI system as a part of Situation Awareness” (National Academies of Sciences, Engineering and Medicine, 2022, p. 31). Transparency supports humans by affording *prospective* information about how the agent performs *in real-time* and informs what its future actions will be. Here, information from the agent is shared to support the operator’s understanding of how the agent is currently performing and which actions it will take in the near future. This means that transparency contributes to perceiving, comprehending, and projecting the decisions and actions of the agent (see Figure 7). Therefore, whilst explainability supports the development of mental models of the agent, transparency supports the development SA of the agent (Endsley, 2023b).

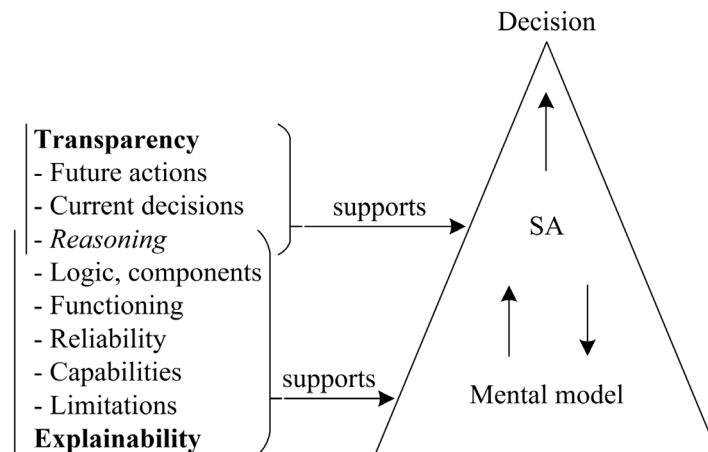


Figure 7. Transparency and explainability in a decision-making context (adapted from Endsley (2023b) and National Academies of Sciences, Engineering and Medicine (2022)).

Operators that interact with agents that provide non time-critical decisions, actions, and recommendations will benefit from understanding the agent’s reasoning in order to build and maintain accurate mental models. Likewise, operators interacting with agents that provide real-time information about their decisions and future actions, will also benefit from the agent’s reasoning in order to build and maintain SA. In other words, explainability supports and maintains the mental models underpinning agent SA, whereas transparency directly supports SA of the agent in its task environment by providing current and prospective information (Endsley, 2023b). Consequently, agent reasoning plays a key role in both concepts. As this dissertation is concerned with supporting effective oversight of agents performing decisions and actions in real-time, this dissertation investigates the application and effects of affording reasoning, decisions, and future actions to human supervisors, i.e., transparency.

### 2.3.3 Frameworks and models for transparency

Research on transparency and explainability has resulted in several models and frameworks for how to develop transparent systems. Among these, the SA-based Agent Transparency model (SAT; J. Y. C. Chen et al., 2014) and the Human-Robot Transparency model (Lyons, 2013) have received considerable attention. In addition, the Coactive System Model based on Observability, Predictability, and Directability is relevant in this context (M. Johnson et al., 2014).

The SAT model combines various theories into an overall model representing three levels of transparency, including the three levels of SA (Endsley, 1995), the Beliefs, Desires, Intentions (BDI) Agent Framework (Rao & Georgeff, 1995), the 3P framework depicting the agent's Purpose, Past performance, and its Processes and algorithms (J. Y. C. Chen et al., 2014; Lee & See, 2004). The model consists of three layers representing system-internal information into a single model. At the first level of the SAT model, the operator is provided with basic information about the agent's current state and goals, intentions, and proposed actions. Also, the agent's purpose, performance, and progress in reaching its goals are represented here. At the second level, the operator receives information about the agent's reasoning process behind those actions and the environmental and other constraints that the agent considers when planning those actions. At the third level, the agent affords operators with information regarding the agent's projection of its future state, such as predicted consequences, likelihood of success/failure, and any uncertainty associated with the aforementioned projections. Also, the agent's limitations are afforded there, including the likelihood or error and history of performance. Finally, although not formally a level of transparency in this model, uncertainty information may either be included in the third level or added as a separate one. Whilst originally a model depicting the structure of information flow from the agent to the operator, in later years, this model has been expanded to include bi-directional transparency between multiple agents and humans (J. Y. C. Chen et al., 2018). Considering the distinction between transparency and explainability (see Figure 7), the SAT model appears to integrate both elements into a single model, thereby supporting the maintenance of mental models as well as system SA.

Lyons (2013) developed a model for transparency based on bi-directional communication between robot and human. The Human-Robot Transparency model (HRT) consists of two parts. First, the robot-to-human part of the model describes the information the robot should provide to the operator to afford transparency. This part consists of several sub-elements describing and structuring the type of information that constitutes robot transparency. This includes an intentional part where the purpose and intent of the robot is conveyed, including why the robot exists, and how the robot is programmed to interact with humans. Furthermore, the task model affords operators the robot's understanding of its tasks, goals, progress, and awareness of its capabilities and errors. Moreover, the analytical model communicates the robot's underlying analytical principles used by the robot to make decisions, especially relevant in situations with high degree of uncertainty. Finally, the environmental model, communicates the robot's understanding of the dynamics of its surrounding environment, including its limitations given the environmental conditions. Second, the human-to-robot part of the overall model consist of a teamwork element affording the state of the robot's role in the human-robot team. In addition, the human state model allows a robot to convey its understanding of the operator's state and intervene in the operator's actions if the system deems this necessary. Similar to the SAT model, the HRT model integrates explainability and transparency elements, in addition to adopting the bi-directional definition of transparency.

The Coactive System Model based on Observability, Predictability, and Directability from M. Johnson et al. (2014) describes both an approach and model for developing transparent systems based on interdependence between agents and humans. Here, the principle of observability is used to make “pertinent aspects of one’s status, as well as one’s knowledge of the team, task, and environment observable to others” (2014, p. 51). This includes the capability of the agent and human to observe and interpret each other’s signals. Predictability implies that one’s actions should be sufficiently predictable such that others can reasonably rely on them when considering own actions. This mutual predictability is an essential element in the relationship between agents and humans. Finally, directability implies the ability to direct the behaviour of others and be directed by other team-member, human or agents. By defining requirements associated with observability, predictability, and directability, designers can identify which information needs to be shared, who to share it with, and when to share it. In line with the SAT-, and the HRT model, this model adopts the bi-directional definition of transparency. However, the model is not explicit about what kind of information should be shared between agent and human beyond information supporting observability, predictability, and directability. It is therefore unclear whether this model supports explainability, transparency, or both.

This dissertation investigates the feasibility of an alternative model for representing transparency that would suit the context of autonomous collision avoidance and the anticipated role of the operator herein. As discussed earlier, this dissertation adopts the stance that transparency is a design principle aimed at enhancing agent understandability and predictability for human supervisors. Transparency is thus a property of the agent and is concerned with the information flow from the agent to the operator. Also, this dissertation is concerned with supporting effective real-time oversight of agents performing safety critical decisions and actions. Although the role of explainability is recognised, this dissertation is interested in studying the effect of real-time information provision rather than post-hoc explanations. Considering the continuous cognitive processes that operators perform to acquire, achieve, and maintain SA knowledge, a model is needed that should be compatible with these processes. The information processing model, depicted in Figure 2 and simplified below, represents how humans perceive and analyse information, make decisions, and perform actions in four discrete stages: information acquisition, information analysis, decision selection, and action implementation. In addition, Figure 3 depicts how the role of SA fits within this model by specifying how the processes of SA assessment underpins decision making by constantly updating and maintaining the SA knowledge represented in the perception, comprehension, and projection stages. Thus, an agent designed in accordance with this model, i.e., by providing operators with insight into its information perception, analysis, decision-making and actions should be compatible with the information needs of the operator supervising that agent.

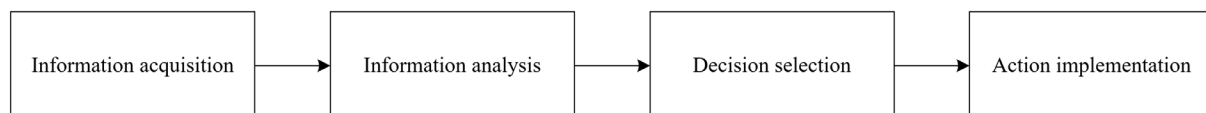


Figure 8. A simple model of human information processing (adapted from Parasuraman et al., 2000).

Parasuraman et al. (2000) describe that, in addition to representing a simple model of human information processing, this model could be used to represent the types and levels of human interaction with automation. That is, when determining the distribution of functions and tasks between humans and automated systems, the authors proposed to use this model to understand in which stage of the information processing cycle automation is deployed and to which degree. This way, the model

represents an approach to understanding the interaction between humans and agents and can thereby be useful in system design. For example, in designing support systems for human decision making in collision avoidance, the distribution of functions and tasks can be based on this model. That is, depending on the degree of sophistication, systems may be delegated the task of performing “information acquisition” from the environment around the ship, but not for any of the other information processing stages. In such systems, only information is provided to the operator without meaningful processing or interpretation. For systems delegated the task of “information analysis”, algorithms may be applied to the data to allow for prediction of the information elements over time. For collision avoidance systems, this implies the capability to predict how the situation around own ship will develop over time, including target ships, objects, land, and etcetera. Here, the implications of the predictions and their relation to own ship play a vital role in determining collision risk. For systems with even further sophistication, the risk picture may be used to derive decision alternatives upon which the risk can be reduced, i.e., “decision selection”. Here, the system may provide the operator with one or more decision alternatives that provide a future track intended to solve the collision risk situation. Finally, systems may also be delegated with the task of “action implementation”. Here, the system may execute an avoidance manoeuvre, determined by the system’s acquisition of information, analysis, and decisions, without human involvement. As such, this simple information processing model can be used as a basis to develop systems that afford operators with SA knowledge of the system’s information perception, processing, decision making, and action execution that is compatible with the operator’s information processing, as discussed earlier.

## 3 Research methodology

This chapter details the dissertation's research philosophy, approaches to data gathering, and considerations regarding the quality of its results. In addition, methodological challenges and ethical considerations are discussed.

### 3.1 Scientific research philosophy

This dissertation is situated within the “Human Factors” (HF) discipline. HF is the scientific discipline concerned with the interaction between humans and other elements of a sociotechnical system (Meister, 1999). The goal of HF is to support humans in their work environment such that errors can be reduced, productivity can be increased, and safety enhanced (Human Factors and Ergonomics Society, 2023). HF is defined as the “scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data and methods to design in order to optimize human well-being and overall system performance” (International Ergonomics Association, 2000; ISO, 2019, p. 2). In other words, HF is concerned with the research and application regarding the arrangement of human and machine elements, that interact and function concurrently, and organized into a whole to accomplish a specified goal (Meister, 1999). By positioning this research within the HF discipline, several philosophical approaches become relevant.

Post-positivism, i.e., the philosophical stance that objective information cannot be observed in its completeness and that approximations are needed, lends itself well as a philosophy for HF research (Howell, 2013). For example, constructs such as mental workload, and SA are challenging to measure directly and are therefore typically measured by use of proxies (e.g., questionnaires, behavioural-, and psycho-physiological measurements). In addition, as HF is concerned with solving real-world problems, the pragmatist philosophical viewpoint is equally relevant considering its focus is on generating practical knowledge in specific contexts, enabling successful action through theories and knowledge, and solving problems to inform future practices (Saunders et al., 2019). Considering the applied nature of this dissertation, i.e., the role of agent transparency in autonomous shipping, the application of a pragmatic philosophy, in addition to a post-positivistic one, is warranted. Combining these philosophies means that this dissertation aims to develop theories and generate knowledge based on objective, valid, and reliable methods that can be applied in practice. Furthermore, in line with the post-positivistic philosophy, the role of the researcher was aimed to be neutral and independent of the gathered data. However, according to the pragmatist philosophy, it is recognised that researcher biases, methodological, and practical limitations may influence the research results. As this dissertation combines the two philosophies, the researcher requires mitigating measures that aim to minimise these effects and optimise objectivity in the data where applicable. These are discussed below.

### 3.2 Methodological approach

This dissertation deployed a mix of qualitative and quantitative methods to address the relationship between agent transparency and human performance (R. B. Johnson et al., 2007; Plano Clark & Ivankova, 2016). Quantitative methods were used to assess the relationship between transparency and agent performance in an experimental setting (Article 5). In addition, qualitative methods were used throughout the dissertation, i.e., to systematically map the literature using the PRISMA method

(Article 1), to understand and analyse collision avoidance manoeuvring using a Goal-Directed Task Analysis (Article 2), to structure data according to a context-specific model of human information processing (Article 3), to develop and design transparency concepts using a human-centred design process (Article 4), and to capture navigator preferences using a ranking approach (Article 5). By combining quantitative and qualitative methods, this dissertation considered a mixed-method approach appropriate to explore facets of the relationship between agent transparency and human performance (Anguera et al., 2018; R. B. Johnson et al., 2007). Each of the primary and secondary methods is depicted in Table 4 and discussed further below.

Table 4. The methodological choice for each research article.

No.	Title	Methodological approach	Primary methods	Secondary methods	Input data
1	Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review	Qualitative	PRISMA	Workshop	Scientific publications
2	Supporting human supervision in autonomous collision avoidance through agent transparency	Qualitative	Goal-Directed Task Analysis	In-situ interviews In-situ observations Workshop	9+2 navigators COLREGs Procedures
3	Supporting human supervisory performance through information disclosure: Establishing transparency requirements for maritime collision avoidance	Qualitative	Modelling	-	GDTA
4	Operationalising Automation Transparency for Maritime Collision Avoidance	Qualitative	Human Machine Interface development	Workshop (x2)	1+2+2 navigators
5	The Influence of Agent Transparency and Complexity on Situation Awareness, Mental Workload, and Task Performance	Mixed-method	Controlled experiment	SAGAT NASA-TLX Time recording Interviews	34+3 navigators

### 3.3 Research methods and data analysis process

This section discusses the methods and data analysis processes that were applied in this dissertation. Detailed descriptions can be found in the method sections of the respective appended articles. However, for readability purposes, relevant sections of the articles are repeated and summarised below. This means that, given the similarities between the descriptions below and the respective sections in the articles, some similarity in wording of sentences and paragraphs should be expected. For clarity, references to the appended articles are provided where relevant.

#### 3.3.1 Article 1: Systematic Literature Review

The first article performed a systematic literature review to establish the overall research context by mapping the state-of-the-art regarding agent transparency and human performance (van de Merwe, Mallam, & Nazir, 2024). Literature reviews are an essential part of any scientific enquiry as in the quest to extend the boundaries of scientific knowledge, knowing where the boundaries are is valuable knowledge (Xiao & Watson, 2019). That is, by mapping out the scientific evidence for a specific field, gaps in the knowledge can be identified and new directions can be chosen. Literature reviews add value compared to the knowledge generated in single studies in that the former integrates and synthesizes findings from multiple studies. Testing an overall theory or hypothesis based on multiple information sources provides stronger evidence than basing a theory on a single study only. Also, evidence sourced from multiple single studies implies that variation in methodological approaches, intervention strategies, and contexts are incorporated in the literature review. This, in turn, provides a broader and more varied information basis compared to single study results only. Furthermore, literature studies are valuable for checking whether consistency exists between scientific studies or whether there are disagreements among the results. Finding variations among the results may point towards gaps in the knowledge or shed light on the applicability of the results. Finally, literature reviews may highlight strengths and weaknesses in the evidence and argue for further research (Petticrew & Roberts, 2006).

Despite the strength of literature reviews to provide the status-quo of a specific field through a meta-analysis, the method is not without its weaknesses (Booth et al., 2016). First, a lack of a clear method may hamper the study's interpretability. For example, imprecise research questions, methods for identifying literature, and approaches to synthesis, makes it difficult for readers to understand the study's results. Second, in deciding which studies to include in the review and which ones not to include, literature reviews run the risk of suffering from selection bias. For example, when studies are chosen that predominantly support the researcher's hypothesis, internal validity may suffer, impacting the defendability of the results. Finally, a lack of a clear link between the data and the study's conclusions may hamper auditability of the results. Without a method and description of how the data from the individual studies is synthesized, it is unclear whether the review's conclusions are derived from the data or from the researcher's a priori assumptions. Systematic literature reviews (SLR) aim to alleviate these issues by deploying a rigorous methodological approach to data gathering, synthesis, and reporting.

The primary rationale for using SLRs is the potential for increased accuracy and improved reflection of reality (Mulrow, 1994). SLRs differentiate from non-systematic ones in their application of explicit scientific principles aimed at reducing biases. In addition, the application of explicit methods allows for increased replicability and improved transparency in terms of its conclusions. That is, in contrast with non-systematic literature reviews, SLRs follow a transparent approach that makes explicit the

review's choices in terms of data selection, analysis, and conclusions. In addition, the SLR answers more specific research questions than the traditional review (Petticrew & Roberts, 2006). Whereas the non-systematic review may provide a detailed and well-grounded overview of the scientific literature, the lack of explicit methods or protocol may threaten the study's comprehensiveness or balance in its selection or discussion of the evidence.

In the first article (van de Merwe, Mallam, & Nazir, 2024), the Preferred Reporting Items for Systematic review and Meta-Analysis protocol (PRISMA) was used to provide a structured approach for gathering, filtering, and reporting on findings in the literature (Moher et al., 2009, 2015). The PRISMA method uses a three-step approach to report its findings: a structured approach to data gathering, defining explicit criteria to filter and reduce data, and clear expectations regarding data analysis. For data gathering, a set of inclusion criteria were established based on the PICOC approach (Population, Intervention, Comparisons, Outcomes, and Context) (Booth et al., 2016; Petticrew & Roberts, 2006). That is, the first article was interested in identifying studies performed on users in the safety critical domain, where transparency principles were tested in simulated and/or operational environments, and where its effect on SA, and/or mental workload, and/or operator performance metrics were reported. In addition, exclusion criteria were established for the initial screening, thereby omitting non-English articles, articles from outside the time-period, non-peer reviewed, or grey literature, and non-experimental studies on transparency. Finally, literature was searched in three databases: Scopus (with ScienceDirect for the full-text journals), IEEE explore, and Web of Science based on a search string. For data reduction, the initial set of 1714 articles derived from the database search was reduced to 1575 after removal of duplicates. These were subsequently filtered based on the inclusion and exclusion criteria. This left 59 articles for full-text review. A workshop was held in which a subset of 25 articles was reviewed by the doctoral researcher and two supervisors. Each participant performed an independent assessment of the articles. 42 articles were excluded with reasons, resulting in 17 articles for the qualitative analysis. For data analysis, the results from each of the individual articles was extracted including the domain in which transparency was studied, the sample size, which (if any) transparency model was used, the Human-Automation Interaction type (HAI), how transparency was operationalized, and the comparisons that were made in the experimental study. For each of the articles the results were extracted, including SA effects of using the automation in the study, the effect on mental workload, and the behavioural/performance measures employed in the study.

### **3.3.2 Article 2: Goal-Directed Task Analysis**

In the second article, a GDTA was used to establish the specific research context and derive SA requirements for human-supervised autonomous collision and grounding avoidance (van de Merwe, Mallam, Nazir, et al., 2024a). Task analysis (TA) is a method commonly used to describe and evaluate the interaction between humans and other actors within a system, e.g., between humans or between humans and machines (Kirwan & Ainsworth, 1992; Stanton, 2006). The method is typically used by human factors engineers, designers, and ergonomists to identify, organise, and analyse what humans are required to do to achieve their goals. By creating a representation of the human involvement in a system, this information can be used to ensure there is compatibility between the goals of the system and the human capabilities such that overall goals can be achieved (Stanton et al., 2013).

Whereas TA is typically concerned with the physical activity performed within the system, cognitive task analysis (CTA) is interested in the mental activities. Understanding the physical activities that



need to be performed to achieve a system goal can be useful when designing step-by-step guidance on task execution or when designing task zones where work is associated with space or location requirement (ISO, 2000). However, modern computer-based systems using screen-based interaction methods put increasing demands on the cognitive skills of the operators. CTA emphasises the mental activities required to achieve system goals and is useful for describing and representing the cognitive elements that underlie goals, decisions, and judgements. In other words, CTA aims to understand how cognition enables humans to perform their tasks, and then uses this understanding to create support systems to assist humans in performing their tasks even better (Stanton et al., 2013).

The second article used a GDTA to derive requirements to support SA when operators supervise autonomous collision avoidance systems (van de Merwe, Mallam, Nazir, et al., 2024a). GDTA is a type of CTA that focuses on the goals, decisions, and information needs of humans when performing a task (see Figure 9). GDTA is a technology-agnostic technique, which means that the analysis aims to determine what operators ideally would like to know to perform a task, without specifying with which technology this information is made available (Endsley et al., 2003). In addition, the analysis aims to understand how the operator integrates the information to derive decisions. This way, systems can be designed that support the cognitive needs of the operator and thereby enhances decision making and performance. The method typically uses a hierarchical structure to visualise how goals, sub-goals, decisions, and information needs are associated. Based on information sources, in which interviews and observations with Subject-Matter Experts (SMEs) play a leading role, the hierarchy is created and populated. Final validation of the hierarchy's content is performed with an independent group of SMEs.

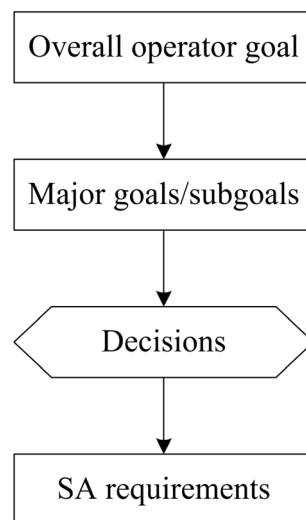


Figure 9. The Goal-Direct Task Analysis method (adapted from Endsley et al., 2003).

The GDTA in article two was based on four input sources: in-situ interviews with navigators, in-situ observations on ship bridges, an appraisal of the COLREGs, and a review of company procedures (van de Merwe, Mallam, Nazir, et al., 2024a). In addition, the results of the analysis were validated with two independent navigators (see Figure 10).

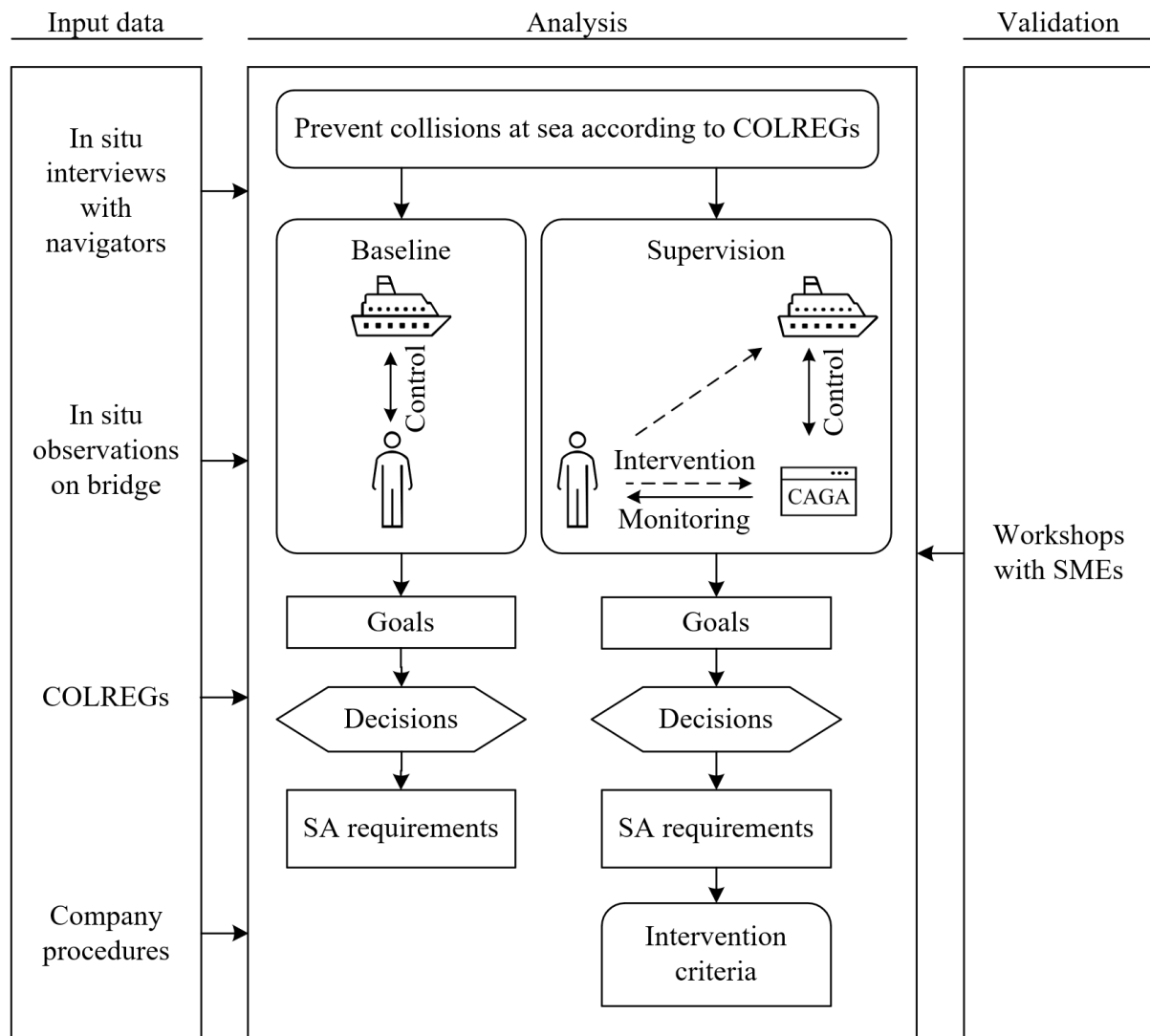


Figure 10. The analysis framework employed in this study (van de Merwe, Mallam, Nazir, et al., 2024a).

First, nine navigators were interviewed whilst on-duty on the ship bridges of passenger ferries. An interview guide was developed addressing the actions navigators perform to obtain a picture of their surroundings, how they determine safe speed, how they determine collision risk, and how collisions are avoided. In addition, the interview guide addressed the navigator's potential interaction with an autonomous collision avoidance system. For this part, a series of questions based on a modified MITRE Human-Machine Teaming Systems Engineering Guide (MITRE, 2018) was used to identify SA requirements when humans team with advanced automation to perform a task. Second, in-situ observations were performed whilst interviewing the navigators and were conveniently used as examples and objects of enquiry during the interviews. Therefore, potential collision situations that arose during the visits were observed, noted, and discussed in detail (see Figure 11). Third, an appraisal of the COLREGs was performed to identify goals, decisions, tasks, and information needs provided in the rule descriptions. The COLREGs describe, to a degree, the tasks to be performed in ship-to-ship encounters (IMO, 1977). As such, the information already embedded in the rule descriptions was used to understand how navigators establish an awareness of the traffic, estimate safe speed, determine collision risk, and decide on which actions are needed to avoid collisions. Fourth, documents provided by the ferry operator were reviewed to identify any specific information and amendments to the information above. Finally, once the GDTA was established, its results were

validated with two independent navigators in workshops. Here, the task analysis was reviewed, and corrections, amendments, and adjustments were made where necessary.



*Figure 11. Observations and interviews performed onboard a passenger ferry.*

### 3.3.3 Article 3: Human information processing model

Article 3 (van de Merwe et al., 2023b) discusses the application of a model to define layers of transparency based on the information requirements established in Article 2 (van de Merwe, Mallam, Nazir, et al., 2024a) concerning conventional- and supervised collision avoidance. The aim for establishing these requirements was to support the development of transparent Human Machine Interfaces (Article 4; van de Merwe et al., 2023a) and support the supervision of collision and grounding avoidance systems. However, depicting all information identified in the GDTA on an HMI would not be prudent as this would likely lead to an excess amount of information for the operator. Therefore, an approach for organising this information was applied using a model of human information processing, i.e., the model by Parasuraman, Sheridan, and Wickens (PSW; 2000).

The PSW model is both a conceptual model for human information processing and a pragmatic means to describe four classes of automated functions. As such, the compatibility of information processing between human and system, as represented in this model, combined with the model's practicality and face validity, made this model attractive to use as a basis for developing transparent automation. Therefore, based on the mapping of goals, decisions and information needs identified in the GDTA (Article 2; van de Merwe, Mallam, Nazir, et al., 2024a), the information requirements were structured based on the model's information processing stages adapted to the collision avoidance context. As per DNV's guidelines for autonomous and remotely operated ships, the following categories were adopted: condition detection, condition analysis, action planning, and action control (DNV, 2021). Based on this contextualised model, a layered approach to transparency was developed in which

information from the task analysis was structured. This resulted in a set of information elements that is unique to each of the information processing stages (see Figure 12).

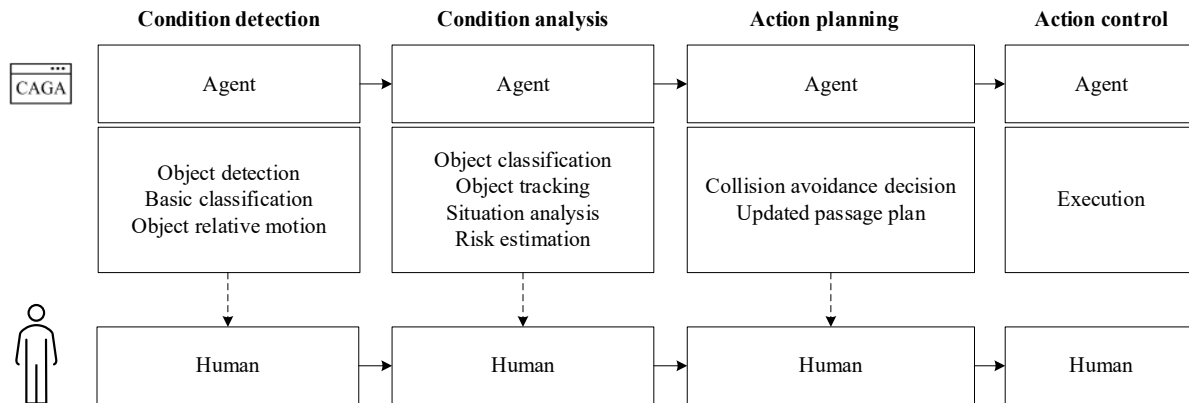


Figure 12. The PSW model applied in a framework to derive transparency requirements for human supervised CAGA systems.

For example, in the “condition detection” stage, the agent conveys which objects it has detected, performs basic classification of these objects, and determines objects’ relative motion. In the “condition analysis” stage, the agent conveys its object classification, tracking, analysis of the situation, and risk estimation. In the “action planning” stage, it depicts its decision regarding collision avoidance and its updated passage plan. Finally, in the “action control” stage the action is executed. However, since there is limited information processed at this processing step, i.e., there is only action execution and monitoring, there is limited information displayed for this stage. Consequently, the contextualised PSW model served as a framework for structuring SA requirements depicting the collision avoidance system’s internal information processing. To examine the results of the GDTA, see Appendix A – Coupling the Goal-Directed Task Analysis, PSW model, and HMI.

### 3.3.4 Article 4: Human Machine Interface development

Article 4 (van de Merwe et al., 2023a) describes the process for developing concepts for transparent HMIs based on the information requirements identified in Article 2 (van de Merwe, Mallam, Nazir, et al., 2024a), and the transparency model described in Article 3 (van de Merwe et al., 2023b). As HMIs are typically system components capable of handling the interaction between humans and systems, the HMI supports human-machine interactions by providing relevant feedback to support SA and allowing for appropriate input commands to support action execution. Since design decisions can have major impact on the user experience, care was taken to develop transparent HMIs that would not interfere in the experimental evaluations (Kirk, 2013). Therefore, the approach to developing accurate, representative, and integrated HMIs was performed in two parts.

First, realistic and representative traffic situations were developed to provide the context for the CAGA system. An iterative approach was used where a navy-certified navigator with five years of navigational experience developed realistic traffic situations based on a set of criteria. These situations were subsequently reviewed in a workshop with two independent navigators and assessed for their degree of realism, complexity and likelihood of occurrence. Also, the navigators were asked to identify if the situation depicted a traffic conflict, if yes, what type, and which manoeuvre own ship should perform. Based on feedback provided by the navigators, the situations were updated and finalised (see Figure 13).

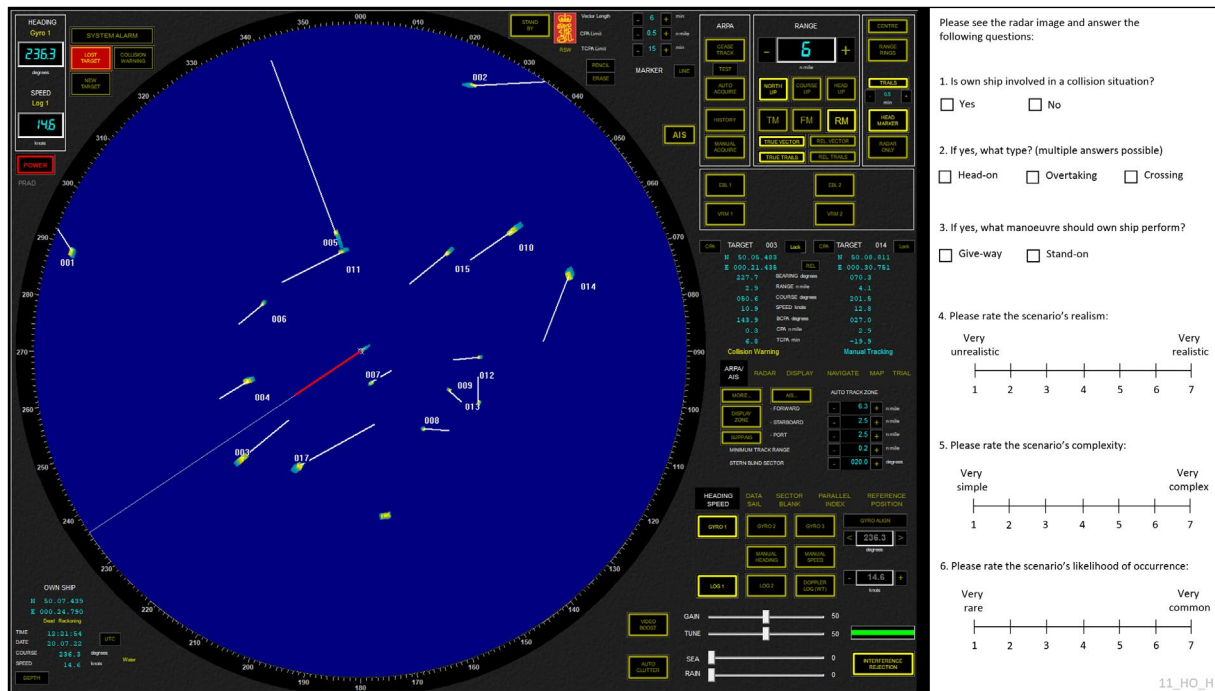


Figure 13. Example of a form used in the workshops to validate the traffic situations.

Second, symbology was developed representing the CAGA system's decisions, planned actions, and reasoning based on the following principles (ISO, 2019). First, HMIs were developed such that the information pertaining to CAGA was suitable to the task. In this context, the transparency information pertaining to the CAGA system was limited to providing the information needed to understand the CAGA system's decisions, planned actions, and underlying reasoning. Second, emphasis was put on integrating transparency information in the user's primary task display. For collision avoidance, the Automatic Radar Plotting Aid (ARPA) is used by navigators to provide "continuous, accurate and rapid situation evaluation" (IMO, 1979, p. 2). This meant that the radar display was chosen as the interface for integrating transparency information from the CAGA system. In practice, this meant placing symbology on and around the radar display similar to regular ARPA symbology. Third, the symbology was designed to be as self-explanatory as possible. This was aimed to be achieved by developing graphical elements representing transparency information based on the IEC 62288 standard for maritime navigation and radiocommunication equipment and systems such that the same "look-and-feel" as the existing ARPA symbology was created (IEC, 2022). In addition, as the graphical elements on the ARPA make limited use of text, i.e., information was primarily presented using symbols, the transparency information followed the same principles. Using an iterative design approach, the symbology was developed with a navy-certified navigator and validated with two independent navigators. Here, two workshops were held where navigators were asked to explain and describe, using a talk-aloud protocol, which information they perceived from the CAGA system. Their answers were compared to the intentions of the design, the designs were subsequently updated and finalised. A detailed explanation of the developed symbology, their explanations, including an example, is depicted in Appendix B – Guide to Human-Machine Interface and symbology.

### 3.3.5 Article 5: Controlled experiment

Article 5 describes the execution of a controlled laboratory experiment that was conducted to evaluate the relationship between agent transparency and human performance variables (van de Merwe,

Mallam, Nazir, et al., 2024b). A full-factorial repeated measures (or within-subjects) design was used with two independent variables: complexity (2 levels) and transparency (4 levels).

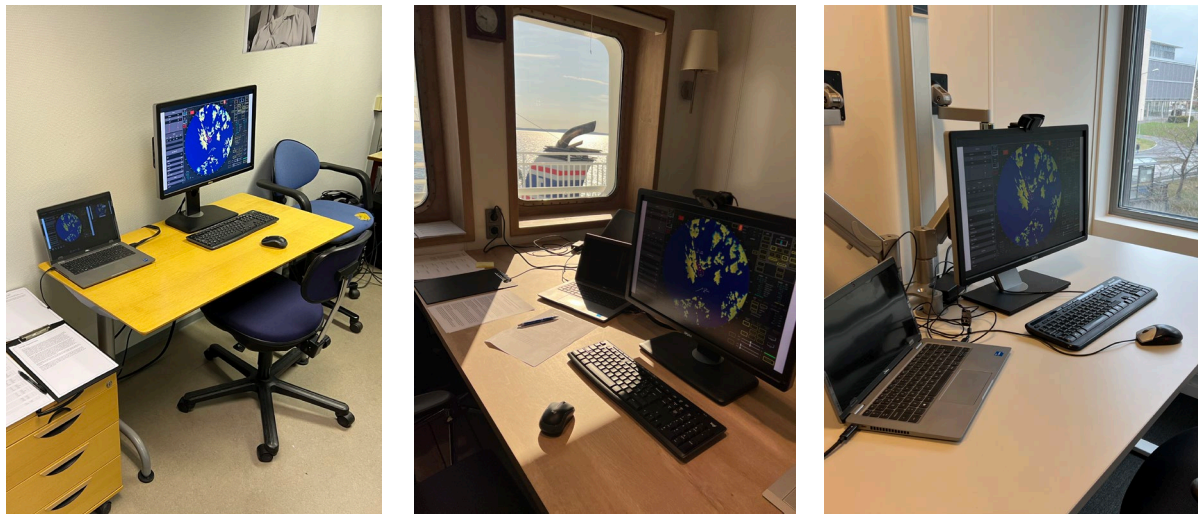


Figure 14. Experiments conducted at the TARG research lab at USN, onboard a passenger ferry, and at NTNU Ålesund respectively.

In this experiment, 34 navigators holding a deck officer license (32 males and 2 females) took the role of a supervisor of a ship equipped with an autonomous CAGA system (see Figure 14). They were tasked with observing and understanding a traffic situation depicting own ship in conflict with a target ship and own ship’s proposed solution to resolve it. Out of a pool of 70 traffic situations (see Table 8), 16 unique traffic situations were used in the experiment (in addition to four for the familiarisation) with two levels of complexity and four levels of transparency, distributed across two sessions (see Table 5).

The order of the traffic situations was randomised, resulting in each participants receiving a different order. As dependent variables, the effects on SA, mental workload, and task performance were measured and averaged across the two sessions (see Figure 15). Finally, participant preferences pertaining to levels of transparency were recorded through a ranking exercise as part of a semi-structured interview.

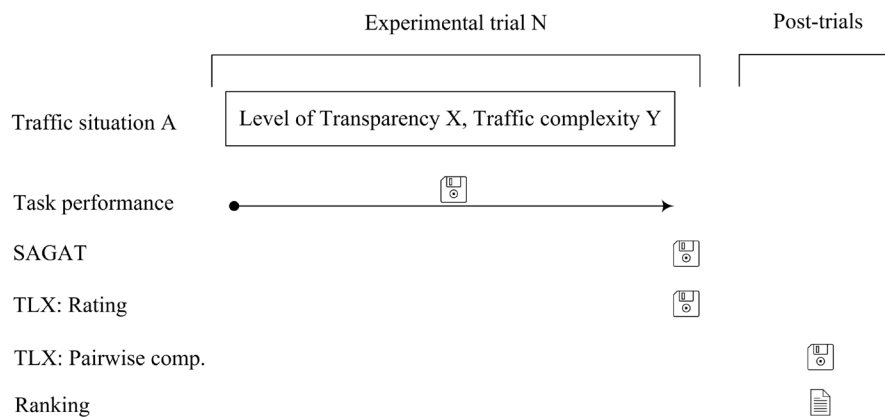


Figure 15. Independent and dependent variables.

Table 5. Traffic situations and configuration used in the familiarisation and experimental trials.

Situations A-D were used in the familiarisation, one to eight were used in trial one, and nine to sixteen in trial two. Key: HO = Head-on, CR = Crossing, OT = Overtaking/overtaken, NC = No collision, RAM = Restricted in Ability to Manoeuvre. The reference column can be used to locate the specific traffic situation in Appendix D – Traffic situations used in the experiment.

No.	Transp. level	Complexity	Collision situation	Own ship priority	Change	Target RAM?	Geo.	Ref.
A	Low	Low	HO	GW	Route	No	Sea	1_HO_L
B	Medium (A)	High	CR	SO	N/A	No	Sea	17_CR_H
C	Medium (B)	High	HO	GW	Route	No	Sea	16_HO_H
D	High	Low	OT	GW	Speed	No	Land	11_OT_L
1	Low	Low	HO	GW	Route	No	Sea	7_HO_L
2	Medium (A)	Low	CR	GW	Route	No	Sea	2_CR_L
3	Medium (B)	Low	HO	GW	Route	No	Sea	2_HO_L
4	High	Low	OT (target)	GW	Route	Yes	Sea	10_OT_L
5	Low	High	HO	GW	Route	No	Sea	15_HO_H
6	Medium (A)	High	CR	GW	Speed	No	Sea	21_CR_H
7	Medium (B)	High	HO	GW	Route	No	Sea	11_HO_H
8	High	High	OT (target)	GW	Route	Yes	Sea	17_OT_H
9	Low	Low	HO	GW	Route	Yes	Sea	9_HO_L
10	Medium (A)	Low	CR	GW	Speed	No	Land	13_CR_L
11	Medium (B)	Low	HO	GW	Route	No	Sea	5_HO_L
12	High	Low	OT (target)	SO	N/A	No	Sea	9_OT_L
13	Low	High	HO	GW	Route	No	Sea	10_HO_H
14	Medium (A)	High	CR	GW	Route	No	Sea	15_CR_H
15	Medium (B)	High	HO	GW	Route	Yes	Sea	13_HO_H
16	High	High	OT (own)	GW	Route	No	Sea	13_OT_H

The Situation Awareness Global Assessment Technique (SAGAT) was used to measure participant SA through queries concerning three levels of SA: the agent's information perception, its comprehension, and its decisions and planned actions (Endsley, 2000). The SAGAT was chosen as the preferred method because of its track record of providing an objective assessment of a person's SA compared to subjective measurements such as the Situation Awareness Rating Technique (SART; Endsley et al., 1998), the Situation Present Assessment Technique (SPAM; Endsley, 2021) and other methods (Gawron, 2019a; Stanton et al., 2013). The SAGAT avoids errors related to poor meta-awareness about one's own SA (participants may not be aware of what they do not know) and prevents measurements being influenced by the participants' perception of their own performance. However, compared to subjective rating scales, the SAGAT requires effort and in-depth knowledge of the scenarios to develop a varied pool of meaningful queries. Still, the strength of the SAGAT in terms of its sensitivity, validity, and reliability were contributing factors for choosing this technique (Endsley, 2000, 2021; Endsley et al., 1998). This means that for this experiment, a set of generic SAGAT queries were developed for each level of SA (see Appendix E – Generic SAGAT queries). These were subsequently tailored to each traffic situation, and were administered directly after each experimental trial, one at a time and in order of level of SA.

Mental workload was measured with the NASA-Task Load Index (Hart & Staveland, 1988). The NASA-TLX was chosen over other methods of workload, such as stand-alone performance measurements, secondary task performance, or psychophysiological measures, because of the sensitivity of the method and the fact that it is well-known and widely used (Gawron, 2019b; Hart, 2006; Stanton et al., 2013). In this method, workload is measured across six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration level. First, participants rate their perceived level of workload for each of these sub-dimensions on a 10-point rating scale, and subsequently, a pair-wise comparison technique is applied in which the subjective importance of each dimension is assessed. The result of this comparison is a set of weights for each of the sub-dimensions. The final workload score is created by multiplying each dimension's score with its weight and summing the results. In this experiment, the NASA-TLX was administered after the final SAGAT query was answered. However, as setting the weights after each trial is somewhat time-consuming and as the type of task is constant across the experiment, a version of the NASA-TLX was used where participants only performed pairwise comparisons once, and only after all experimental trials were performed. As such, the weights derived from the pairwise comparison applied to all workload scores for the individual trials (Gawron, 2019b). See Appendix F – NASA-TLX for the dimensions and pairwise comparisons.

Task performance was operationalised as the time required to comprehend the traffic situation (i.e., time-to-comprehension; TTC). In psychological research, an abundance of indices is available to measure human performance (Gawron, 2019a). Indicators include measures of accuracy (e.g., number of correct responses, number of errors, deviations), time indices (e.g., search time, reading speed, time to complete), domains specific measures (e.g., aviation, driving, control rooms), and team performance measures (e.g., team communication, team effectiveness measures). In addition, one can measure performance through psychophysiological measurements (e.g., heart rate, eye tracking, galvanic skin response), psychological measurements (e.g., rating scales, self-report measurements), 3<sup>rd</sup> party assessments (e.g., expert judgements), and primary task performance (e.g., speed and accuracy of task completion) (Coleman, 2019). For this experiment, a primary task performance measurement (time to comprehension) was used to assess the time it took participants to evaluate and understand the traffic situation including the information provided by the CAGA system. The time it took participants was self-guided and consisted of participants deciding that the traffic situation and the visualised solution was sufficiently understood. The time measurement started when the traffic situation was displayed and ended upon a key press by the participant after which the screen was blanked. Time was measured in seconds with no time limit imposed. However, to instil participants with a sense of urgency, participants were told they had a 90 second time-limit to evaluate the traffic situation after which the radar image would disappear automatically. No time keeping device was available to the participants and, in practice, there was no time limit imposed by the researchers to avoid a ceiling effect in the measurements.

Finally, the participants' preferences were recorded during a post-experiment interview. Preference was operationalized through a ranking exercise in which participants were asked to rank the four levels of transparency on two dimensions: observability and predictability (MITRE, 2018). Understanding the participant's perceptions regarding the system's transparency is interesting as it provides a subjective dimension to the aforementioned human performance variables, considering the role of user perceptions in technology acceptance (Davis, 1989; Venkatesh et al., 2003). A think-aloud protocol was used to record the participant's verbal reasoning of the ranking (Eccles & Aarsal, 2017).



Ranking was recorded to supplement the other dependent variables and obtain qualitative feedback on the transparency levels (see Figure 16 and Appendix G – Ranking transparency levels).



Figure 16. Screenshot from the ranking exercise from one of the participants.

### 3.4 Recruitment process

In this doctoral research, a total of 50 participants were recruited for all of the activities and divided across three articles (Articles 2, 4, and 5). The recruitment of the participants was primarily performed using snow-ball sampling techniques based on the professional networks from the doctoral researcher and the supervisors from DNV and USN.

For the second article, the Goal-Directed Task Analysis, nine navigators were interviewed on the topic of collision avoidance manoeuvring onboard a passenger ferry (van de Merwe, Mallam, Nazir, et al., 2024a). This included two for which video interviews were performed due to the resurgence of COVID-19 restrictions. In addition, two independent navigators from the university's maritime education program performed an independent validation of the results of the GDTA. These participants were identified through the doctoral researcher's and the supervisors' professional network. For the fourth article, a navy-certified navigator was recruited to assist in the development of the traffic situations and HMIs (van de Merwe et al., 2023a). During the development of the traffic situations, a workshop was held for which two independent navigators provided structured feedback on the traffic situations. In addition, a separate workshop was held to validate the developed HMIs with two licensed navigators. For the fifth article, three navigators were recruited to perform pilot testing prior to commencement of the experiment and 34 navigators were recruited for the experiment (van de Merwe, Mallam, Nazir, et al., 2024b). These participants were recruited through various means using both convenience and snowball techniques based on the doctoral researcher's- and supervisor's professional networks within DNV, USN, and externally.

In terms of distribution between males and females, the maritime industry is a male dominated industry. This gender imbalance made it challenging to obtain an equal distribution in this doctoral

research. Throughout the studies, only two females were recruited: none for Article 2 and 4, and two in Article 5. Therefore, no scientific enquiries were performed to include gender differences in the analysis. In terms of the distribution of all participants across the articles, Article 2 used navigators from ferry operators only, Article 4 used navigators from nautical training institutes and DNV only, and Article 5 used navigators from various ferry operators and nautical training institutes.

### **3.5 Validity, reliability, and quality of research**

Because the aims of research are to draw conclusions about an effect and to make generalisations to other settings of interest, validity and reliability are central components herein (Kirk, 2013). Validity refers to the degree to which “experimental results lead to an intended conclusion from the data” (Ritter et al., 2013, p. 2). Reliability is “concerned with the extent that an experiment can be repeated or how far a given measurement will provide the same results on different occasions” (Howell, 2013, p. 2). There are various types of validity, but in general, three types are of greatest interest here: internal, external, and statistical conclusion validity (Kirk, 2013).

Internal validity refers to how well the research design explains its outcomes or whether the results were influenced by other factors. In qualitative research, internal validity can be enhanced by applying scientific rigour in developing a theoretical framework, selecting approaches to data gathering and participant sampling, concurrently collecting and analysing data, ability to shift between micro- and macro perspectives to develop theories (Malterud, 2001; Morse et al., 2002). In addition, the use of member check and cross-checking activities with independent subject matter expert are means to enhance the validity of the results (Malterud, 2001). In quantitative research, internal validity can be enhanced using experiments to control for externally influencing factors. This way, precise answers to the question of which causes lead to which effects can be answered (Coleman, 2019).

External validity refers to how well the research findings are generalisable to and across populations and settings. In qualitative research, external validity is also referred to as transferability, which is the “range and limitations for application of the study findings, beyond the context in which the study was done” (Malterud, 2001, p. 484). Purposeful sampling, based on a theoretical framework, in which participants with in-depth subject knowledge are obtained from a relevant population, is one of the ways to enhance transferability in qualitative research. In quantitative experimental research, high internal validity may have consequences for external validity, i.e., generalizability. Because experiments require a tight control of extraneous variables to investigate cause-effect relationships, the study’s ability to predict the same effect in real-life situations may suffer (Jarvie & Zamora Bonilla, 2011). Nevertheless, strategies such as sampling participants from a relevant population to perform experimental tasks illustrative of real-world tasks can strengthen the study’s ability to reflect real-world effects (Ritter et al., 2013).

Finally, statistical conclusion validity refers to the ability to make valid inferences from data without being affected by random error or improper statistical procedures (Kirk, 2013). Threats to this type of validity include low statistical power, violation of assumptions of statistical test, fishing for significant results, poor reliability of measures, poor reliability of experimental implementation, and poor control of individual differences between participants. Strategies such as obtaining a sufficiently large sample size to enhance power (Boruch, 1997; Ritter et al., 2013), appropriate use of statistical methods (Tabachnick et al., 2019), reliable and valid measurement instruments (Coleman, 2019), a standardised approach to execution of the experiment (Ritter et al., 2013), a repeated measures design (Coleman,

2019), and randomisation of stimuli (Coleman, 2019; Jarvie & Zamora Bonilla, 2011), are effective means to control statistical conclusion validity. In addition, pilot tests, as depicted in Figure 17, are important to ensure the technical, procedural, data quality, and analytical elements of the experiment (Ritter et al., 2013).

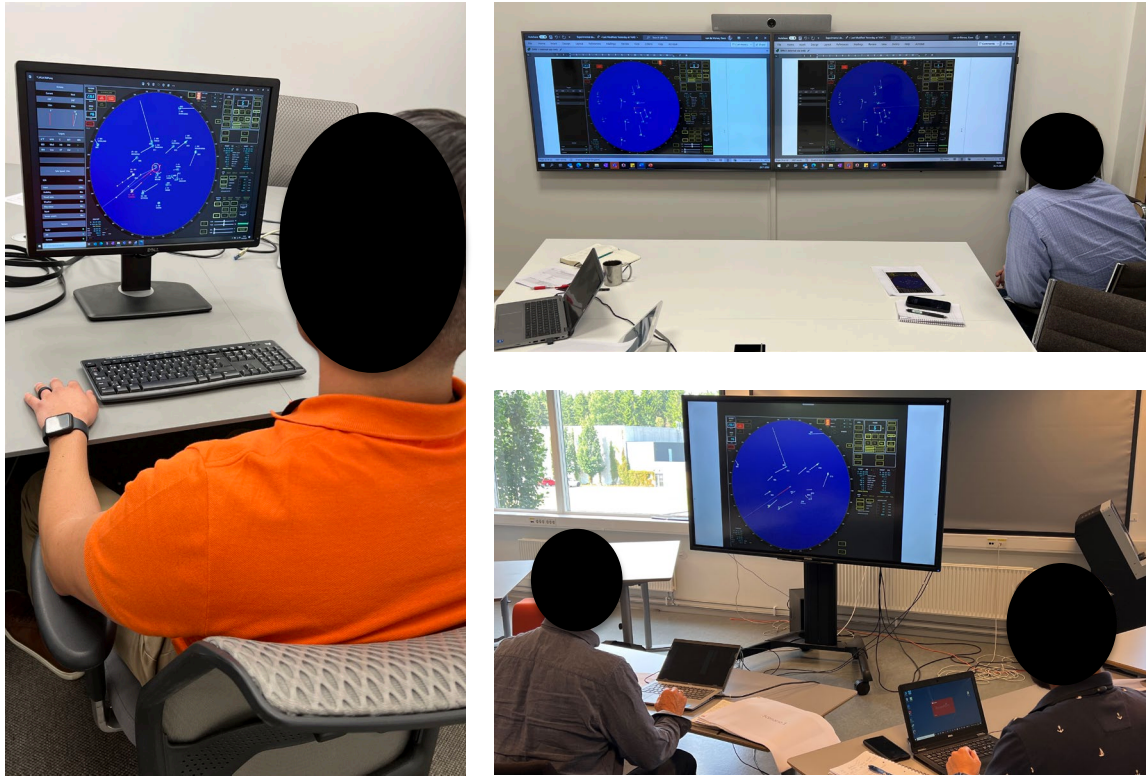


Figure 17. Ensuring reliability, validity, and quality of the experiment.

*Clockwise: Five pilot experiments were performed prior to the start of the experiment, two independent navigators validated the transparency designs, and two independent navigators validated the traffic situations.*

Reliability is concerned with the repeatability of a study and the extent that a measurement provides the same result on different occasions, in other words, stability of results (Howell, 2013). Reliability is often associated with research from the positivistic and post-positivistic philosophies, e.g., experimental research. Here, reliability refers to “the reproducibility of a measurement and describes the extent to which a particular method of measurement will yield the same result repeatedly on a trait presumed to be stable” (Boruch, 1997, p. 6). Establishing reliability in non-experimental studies is more challenging but nonetheless more easily realised when a structured, positivistic approach to the research is chosen (Howell, 2013). For example, when studying the collision avoidance task in Article 2, multiple qualitative methods provided input to the GDTA. This multi-method approach aimed to offset each method’s strength and weaknesses, triangulate the data, and complement the various methods and contributed to the data’s validity and reliability. In applying this approach, the comprehensiveness and accurateness of the study’s results and conclusions are enhanced (Moon, 2019; Plano Clark & Ivankova, 2016).

Table 6 depicts the strategies deployed in this dissertation to address the studies’ validity and reliability.

Table 6. Strategies to control validity and reliability in this dissertation.

	<b>Article 1</b>	<b>Article 2</b>	<b>Article 3</b>	<b>Article 4</b>	<b>Article 5</b>
<b>Method</b>	PRISMA	GDTA	Modelling	HMI development	Controlled experiment
<b>Validity</b>	<p>Applied an established method for systematic literature reviews</p> <p>Performed data synthesis based on clear research questions</p> <p>Three researchers performed the full-text selection</p>	<p>Used multiple data gathering methods: In-situ interviews, in-situ observations,</p> <p>COLREGs, procedures</p> <p>Reviewed results with two independent navigators</p>	<p>Applied an established model for human information processing</p> <p>Contextualised model based on published guidelines for autonomous shipping</p>	<p>Used navy-certified navigator in development of traffic situations and HMIs</p> <p>Applied iterative approach</p> <p>Reviewed traffic situations and HMIs with independent navigators</p>	<p>Performed pilot testing</p> <p>Used realistic traffic situations</p> <p>Transparency layers integrated in ARPA</p> <p>34 licensed navigators participated</p> <p>Participants performed relevant task</p> <p>Measured with established methods (SAGAT, NASA-TLX, time)</p> <p>Stimuli were randomly assigned</p> <p>Performed quality control measures prior to data analysis: checking data for normality, outliers, missing values</p> <p>Performed data analysis using established statistical methods (RM-ANOVA)</p>
<b>Reliability</b>	<p>Systematically searched for relevant literature based on explicit criteria</p> <p>Reviewed full-text articles based on in- &amp; exclusion criteria</p> <p>Specified the analytical process</p>	<p>Used multiple data gathering methods: In-situ interviews, in-situ observations,</p> <p>COLREGs, procedures</p> <p>Applied an established task analysis method</p>	<p>Established categorisation criteria for generating transparency levels</p>	<p>Established explicit criteria for developing varied traffic situations</p> <p>Correlated GDTA and model based on categorisation criteria to generate levels of transparency</p>	<p>Established experimental procedure</p> <p>Used validated experimental software</p> <p>Automated execution of experiment</p> <p>Automated data capture</p>

### 3.6 Research ethics

This dissertation followed the Norwegian national research ethics guidelines for ethical research. For each individual study involving external participants, the Norwegian Agency for Shared Services in Education and Research evaluated and approved the dissertation's approaches to gathering and managing personal data (reference nr. 579620 and 986652). In addition, all personal data was managed according to the General Data Protection Regulation.

In this dissertation, informed consent was obtained from participants for the data gathering in Article 2 (interviews and workshops), Article 4 (two workshops), and Article 5 (experiment). For each occasion, participants were given a written form that explained the purpose of the project, what was expected of the participants, how personal data was managed, stored, and used in the research and publications. Participants were informed that participation was voluntary and that they could withdraw at any given moment and without reason. Participants also had the right to access their own personal data, request its deletion, request rectification, receive a copy, and send a complaint to the relevant authorities. Finally, participants were provided with contact information of the doctoral researchers' supervisors, the data protection officer, and the Norwegian Agency for Shared Services in Education and Research. All participants consented to the information provided and signed the form.

For the recruitment of participants for the workshops (Articles 2 and 4) and the experiment (Article 5), personal information, i.e., name, email address and telephone number, were gathered. However, to separate personal information from research data, a coding system was used to identify participant data without linking it to personal data. The key that linked the participant's personal data to the research data was stored separately and was deleted after the project's end data (28.02.2024).

Finally, all articles followed the Recommendation for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals, i.e., the Vancouver Recommendations (International Committee of Medical Journal Editors, 2023, p. 2; ICMJE). The ICMJE states that authorship is based on the following criteria:

1. "Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND
2. Drafting the work or reviewing it critically for important intellectual content; AND
3. Final approval of the version to be published; AND
4. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved."

All the persons named in the articles in this dissertation have satisfied these criteria. Additional persons, institutions, or companies that have significantly contributed to the content of the work have been identified in each of the articles' acknowledgement sections. For statements regarding authorship in this dissertation, see Appendix I – Statements of co-authorship.



## 4 Results

This section presents the results from the studies performed in this dissertation. Detailed descriptions can be found in the Results sections of the respective appended articles. However, for readability purposes, relevant sections of the articles are repeated and summarised below. This means that, given the similarities between the descriptions below and the respective sections in the articles, some reproduction of texts, figures, tables, as well as similarity in wording of sentences and paragraphs should be expected. For clarity, references to the appended articles are provided where relevant.

### 4.1 Article 1: Understanding the overall context

In Article 1, the primary findings from a systematic literature review are discussed (van de Merwe, Mallam, & Nazir, 2024). The results are summarised below. However, for a detailed overview of the results, including extensive tables, see Tables 1 and 2 in Article 1 in Appendix H – Publications.

#### 4.1.1 Domains, models, interaction types, and operationalisations

The article presents the domains in which transparency has been researched, the different transparency models applied by the experimental studies, the various human-automation interaction types and tasks employed by the studies, and the ways in which transparency has been operationalised in the literature.

The literature review found that transparency has been primarily researched in the military domain. 53% of the studies in the SLR focused on UAV operations, ground troop support, and military aviation, 12% of the studies were performed within the autonomous driving domain, whilst the other domains were related to civil defence (12%), civil aviation (12%), nuclear (6%), and robotics (6%).

Two transparency models were prevalent in the literature review: the (Dynamic) Situation Awareness-based Transparency model (J. Y. C. Chen et al., 2014, 2018) and the Human-Robot Transparency model (Lyons, 2013). In the 17 articles analysed for full-text review in the SLR, eight studies (47%) used the SAT model and one study (6%) used the HRT model. For the remaining eight studies (47%), no particular model was used as the basis for HMI- and experimental design.

Three categories of human-automation interaction types were identified in the studies. In six studies (35%), participants were tasked with *responding to proposals* provided by the agent, in five studies (30%), participants were required to *supervise automation*, i.e., monitor, respond to, and manually operate, and in six studies (35%), participants were required to *monitor only*, i.e., no manual interaction was required.

Finally, most of the studies in the SLR implemented transparency information on a graphical user interface. Exceptions include Skraaning and Jamieson (2021) who provided verbal feedback to nuclear control room operators about the state of the system, in addition to implementing graphical elements. Also, Bhaskara et al. (2021) gave participants in the higher transparency conditions formulae used by a recommender system on a sheet of paper. Otherwise, the studies used graphical elements to convey the system's inner reasoning, including icons, colour and opacity coding, text prompts, and graphs.

#### 4.1.2 Transparency, SA, mental workload, and task performance

The literature review also gathered empirical evidence from experimentally controlled studies on the effect of agent transparency and SA, mental workload, and task performance. Nine out of 17 studies

measured the effect on SA, 15 out of 17 studies measured mental workload, and 10 out of 17 studies measured task performance.

The findings indicate a neutral to positive effect of transparency on SA. However, the results vary between measurement technique and Human Automation Interaction type (HAI). Improved SA scores were found for the studies by Roth et al. (2020; SAGAT scores only), T. Chen et al. (2014, 2015), Skraaning and Jamieson (2021; experiment 2) and Selkowitz et al. (2017). Neutral effects were found in the studies by Roth et al. (2020; SART scores only), Guznov et al. (2020), Skraaning and Jamieson (2021; experiments 1 and 3), Selkowitz et al. (2015, 2017), Wright et al. (2020), Pokam et al. (2019).

The results indicate a predominantly neutral effect of agent transparency on mental workload. Ten studies found no effect between agent transparency and mental workload as measured with various subjective and objective indicators, two studies found an increase, and four studies found a decrease. Two studies measured workload with different methods and found different results, i.e., Skraaning and Jamieson (2021) and Selkowitz et al. (2017). Nevertheless, most of the studies in the review indicate that providing transparency information did not affect the participants to such an extent that this led to increased mental workload. In addition, adding transparency information did not lead to reductions in mental workload either.

Five studies found indications of improvements on task performance indicators such as correct use and rejections of agent generated proposals (Bhaskara et al., 2021; Mercado et al., 2016; Stowers et al., 2020), response time (Stowers et al., 2020), improved goal completion (T. Chen et al., 2015; Skraaning & Jamieson, 2021), and performing detection and verification activities (Skraaning & Jamieson, 2021). However, six studies found reductions in performance, such as, response time (Bhaskara et al., 2021; T. Chen et al., 2015; Roth et al., 2020; Skraaning & Jamieson, 2021), separation conflicts (Göriztlehner et al., 2014), and verification activities (Sadler et al., 2016). Finally, eight studies found no difference in task performance, e.g., Roth et al (2020), Sadler et al. (2016), and Wright et al. (2020).

## 4.2 Article 2: Identifying requirements for supervision

Article 2 presents and discusses the results from the GDTA in terms of identified information requirements for conventional and human supervised collision avoidance (van de Merwe, Mallam, Nazir, et al., 2024a). The results are summarised below, but for a detailed discussion of the results, see the Results section of Article 2 in Appendix H – Publications, especially Figures 5 to 8.

The results indicate a change towards increased cognitive activities required to verify agent performance. In the conventional collision avoidance case, i.e., the baseline, decisions and actions are performed by the navigator. Here, the navigator perceives the environment, analyses relevant information, determines collision risk, makes decisions given the situation, and executes avoidance manoeuvres. In the supervised collision avoidance case, i.e., the supervision case, these tasks are outsourced to the CAGA system, and it is now the system that perceives, analyses, decides, and takes actions. In this scenario, the operator is left with supervising the performance of the CAGA system. Thus, the supervision of the system entails ascertaining that CAGA makes a full appraisal of the situation, proceeds at a safe speed, determines collision risk, and performs avoidance actions in accordance with the Rules.



The results section of Article 2 depicts and exemplifies the change in SA requirements when shifting from the baseline case to the supervision case. That is, in conventional collision avoidance the navigator requires a range of information elements to be able to perform the tasks associated with performing the look-out function (COLREG Rule 5), determining safe speed (Rule 6), determining risk of collision (Rule 7), and taking actions to avoid collisions (Rule 8). For example, for Rule 5 – Look-out this includes location information of vessels, terrain, and objects, relative motion of targets, (time to) closest point of approach (CPA/TCPA) for targets, bow cross range and time, course of ground of targets. For Rule 6 – Safe speed, this includes information on current speeds, meteorological conditions, detected targets on ARPA and ECDIS, known vessel characteristics, look-out and navigational charts, water level contours, and limitations to radar equipment. For Rule 7 – Risk of collision, this includes detected targets, true- and relative vectors for targets, bearing or COG changes for targets, and CPA/TCPA for targets. Finally, for Rule 8 – Actions to avoid collision, this includes knowledge of Rule 18 – Responsibilities between vessels, changes to bearing with target vessels, availability of sea room, CPA with target vessels, and visual estimation of distances.

In the supervision case, the operator of the collision avoidance system needs to be able to verify that the system is performing the collision avoidance function according to its performance requirements (see Figure 18). This means that the system needs to be able to provide the relevant information necessary to allow for its verification. To satisfy Rule 5 – Look-out, this means depicting the CAGA system's detection of vessels, terrain, objects, and its estimated collision risk. For Rule 6 – Safe speed, the CAGA system should depict its chosen safe speed, the parameters this is based on including their individual effects, and any uncertainties and limitations to the data. For Rule 7 – Risk of collision, CAGA should depict which objects it has detected in the short to long range, an evaluation of their type, size, and activity, which objects pose a collision risk, and the status of its relevant sensors. Finally, for Rule 8 – Actions to avoid collision, CAGA should indicate its intended actions, the time to perform these actions, the magnitude of its planned course and/or speed changes, its identified limits to sea room and obstructions, its CPA to target vessels during the planned avoidance manoeuvre, and its determination of vessel priority based on Rule 18 – Responsibilities between vessels.

A supervisor may intervene in the CAGA system if its performance is not in accordance with its expected standard. For example, the supervisor may intervene when its detection of targets-, speed-, estimated collision risk-, intended actions-, time to perform these actions-, or determination of vessel priorities are incorrect. Also, when there are large uncertainties in its input data and there is disagreement across and between its sensors, the supervisor may intervene. Thus, Article 2 identified that the supervisor requires independent information to be able to evaluate the performance of the CAGA system, e.g., provided through conventional means. For example, on a bridge, a navigator may look out the window to cross-check the information from the collision avoidance system and may intervene using the ship's existing control options. In a remote-control centre, a supervisor may need access to information independent of the collision avoidance system.

For detailed results of the GDTA, see Appendix A – Coupling the Goal-Directed Task Analysis, PSW model, and HMI.

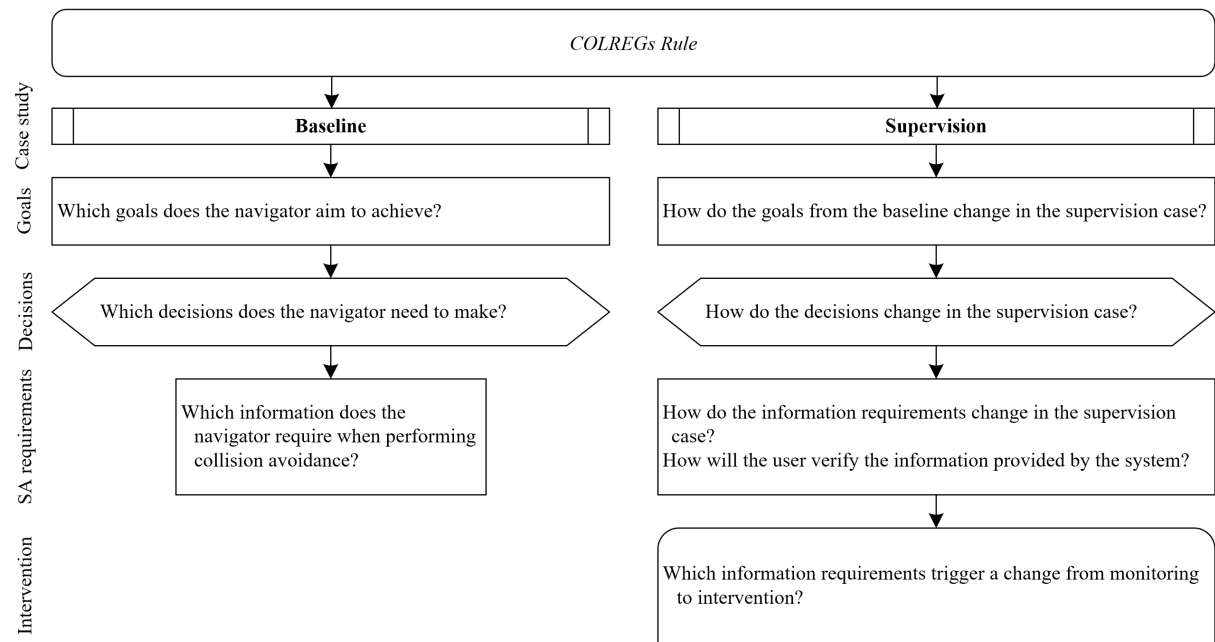


Figure 18. The generic framework used to map the changes between the baseline and supervision case.

### 4.3 Article 3: Structuring requirements using the PSW model

Article 3 (van de Merwe et al., 2023b) discusses the application of the PSW model to the SA requirements that were identified in Article 2 (van de Merwe, Mallam, Nazir, et al., 2024a).

The context-adapted transparency model, based on the original PSW model, describes the four stages of information processing as condition detection, condition analysis, action planning, and action control. In the *condition detection* stage, the CAGA system determines the presence of objects, provides basic information regarding these objects, and provides own ship relative motion with regards to these objects. In the *condition analysis* stage, target objects are tracked, classified, and future states are predicted. Based on this analysis, the collision avoidance system can determine the risk of collision with own ship. In the *action planning* stage, actions are planned based on the outcomes of the risk analysis with which the risk of collision can best be avoided whilst adhering to the COLREGs. This includes determining which ship has “right of way” and which changes are needed to course and speed (if any) to clearly indicate own ship’s intention to avoid collision. Finally, in the *action control* stage the collision avoidance system sends the parameters of the updated passage plan to the control- and machinery system for effectuation. Limited information processing is performed in this stage as it is primarily concerned with the execution of decisions made earlier.

In terms of mapping the SA requirements to this model, the results show that each of the information processing stages of the system can be depicted using a unique set of information parameters (see Table 7). Also, each set of information elements are additive in terms of that they convey information that other elements do not. For example, in the *condition detection* stage, the system may show which objects it has detected in the short and long range, target object type and size, whether the object is moving or stationary, and whether these objects are crossing, head-on, or overtaking. In the *condition analysis* stage, the CAGA system may depict the targets it follows over time, which ship types and manoeuvrability classifications it is considering, and target location predictions based on course and speed. In the *action planning* stage, the system conveys how the system sees the solution to the

collision risk situation. Here, the CAGA system may depict which ship requires to give-way or stand-on and what actions the system intends to perform, including if this means keeping current course and speed. In the final stage, *action control*, the execution of the plan is monitored only and any deviations from the original plan, is detected, analysed, decided, and planned in the preceding stages through a continuous information loop. Further details on how the data from the GDTA was structured according to the PSW model can be found in Appendix A – Coupling the Goal-Directed Task Analysis, PSW model, and HMI.

Table 7. Applying the PSW model to the information requirements from the task analysis.

Key: OT= overtaking/overtaken, HO=head-on, CR=crossing, GW=give-way, SO=stand-on.

Information processing stage	Information requirements for supervision (excerpts)
<b>1. Condition detection:</b> CAGA performs object detection, basic classification, object tracking, and status	<ul style="list-style-type: none"> <li>- Detected objects short &amp; long range</li> <li>- Identified target ship</li> <li>- Target object type and size</li> <li>- Identified target object as OT/HO/CR</li> <li>- Uncertainties in the radar/ sensor data</li> <li>- Status of sensors</li> </ul>
<b>2. Condition analysis:</b> CAGA performs object classification, tracking, situation analysis, and risk estimation	<ul style="list-style-type: none"> <li>- Objects that pose risk</li> <li>- Plotted objects</li> <li>- Risk object type and size</li> <li>- Risk object priority</li> <li>- Risk object course and speed</li> <li>- Risk object intended trajectory</li> <li>- Risk object conflict type</li> <li>- Safe speed parameters</li> </ul>
<b>3. Action planning:</b> CAGA decides on collision avoidance manoeuvring and determines an updated passage plan	<ul style="list-style-type: none"> <li>- Own ship priority (GW/SO)</li> <li>- Target vessel priority (GW/SO)</li> <li>- Own ship intended track and speed</li> </ul>
<b>4. Action control:</b> CAGA executes the plan	N/A: only action implementation

## 4.4 Article 4: Developing Human Machine Interface concepts

Article 4 (van de Merwe et al., 2023a) presents the process for developing traffic situations and HMI concepts for a CAGA system, including design examples, based on the structured SA requirements described in Article 3 (van de Merwe et al., 2023b).

### 4.4.1 Developing traffic situations

70 traffic situations were developed that aimed to represent realistic conflicts and reflect the variety of situations navigators could encounter in real-life (see Table 8). The situations were created on a desktop ARPA simulator from a popular equipment manufacturer and developed by a navy-certified navigator with five years of navigational experience. This pool of traffic situations provided the foundation for developing and applying transparency symbology, including experimental evaluation. To ensure consistent variation between the situations, the following criteria were established. First, situations were developed with high- and low levels of complexity. In low complexity situations own

ship was not restricted in its manoeuvring. In high complexity cases, own ship was “boxed-in” indicating there were some restrictions to performing avoidance manoeuvring thereby increasing the situation’s complexity. High complexity situations also had increased traffic density compared to low complexity situations. Second, the COLREG’s define three types of conflicts: crossing, head-on, and overtaking/overtaken. In principle, these three types represent all defined types of collisions one can encounter at sea. In addition, situations in which there were no collisions were developed. Third, according to COLREGs, when own ship encounters a target ship, it is either in a give-way- or in a stand-on situation. Give-way indicates own ship needs to perform an avoidance manoeuvre, stand-on indicates the target ship needs to manoeuvre. Forth, ships can have various restrictions, e.g., in their ability to manoeuvre. For this research, only target ships could be restricted in their ability to manoeuvre, not own ship. Finally, traffic situations could include land and/or open sea.

Table 8. Criteria for establishing a varied set of traffic situations.

Key: HO = Head-on, CR = Crossing, OT = Overtaking/overtaken, NC = No collision, RAM = Restricted in Ability to Manoeuvre. \*Note: in a head-on situation with one motorised target ship and no other exceptions, own ship cannot be stand-on.

Variant/Complexity	Head-on		Crossing		Overtaking		Total
	Low	High	Low	High	Low	High	
Type (HO/CR/OT)	5	5	4	4	4	4	26
Type (NC)	2	2	2	2	2	2	12
Own ship stand-on*	0	0	2	2	2	2	8
Restrictions target (RAM)	2	2	2	2	2	2	12
Geography (land)	2	2	2	2	2	2	12
Total	11	11	12	12	12	12	70

#### 4.4.2 Developing transparent HMIs for collision avoidance

To visualise the SA requirements, a set of symbology was developed that allowed the CAGA system’s information processing to be represented (see Appendix B – Guide to Human-Machine Interface and symbology, for an overview of the developed symbology and their explanations). Four concept illustrations are depicted below that use the same underlying traffic situation but add and vary layers of transparency information. Additional traffic situations depicting variations in traffic complexity and levels of transparency are included in Appendix D – Traffic situations used in the experiment. Note that the descriptions corresponding the illustrations are similar to those presented in Article 3. Still, for the purpose of readability and clarity in presenting the results of Article 3, these sections are repeated below.

Figure 19 depicts a traffic situation with own ship, in the centre of the radar screen, and several other ships within a 6nm range. In this case, own ship engages in a head-on situation with ship “003”. According to COLREG rule 14 “when two power-driven vessels are meeting on reciprocal or nearly reciprocal courses so as to involve risk of collision each shall alter her course to starboard so that each shall pass on the port side of the other” (IMO, 1977). However, ship “004” poses a hindrance to free avoidance manoeuvring for own ship and care should be taken that an avoidance manoeuvre does not result in a new collision situation.

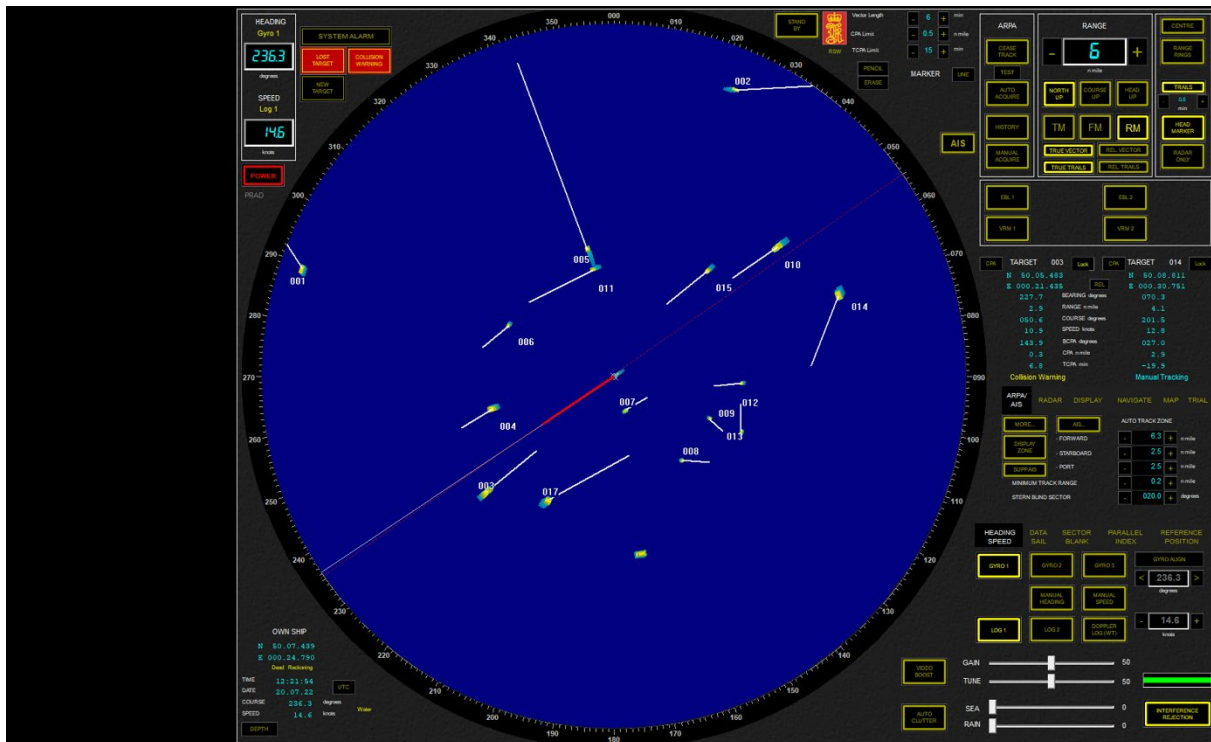


Figure 19. Traffic situation without transparency information.

In Figure 20, the same traffic situation is shown as above, but now with varying levels of transparency: low, medium (A), medium (B), and high. For larger versions, see Appendix C – Examples of transparency levels.

In the *low transparency variation* (top-left), own ship indicates its intended avoidance manoeuvre by drawing its planned track for the next three manoeuvring steps (each step corresponds to one vector length and equals six minutes). The system also depicts “GW” next to the own ship symbol which indicates it intends to give-way.

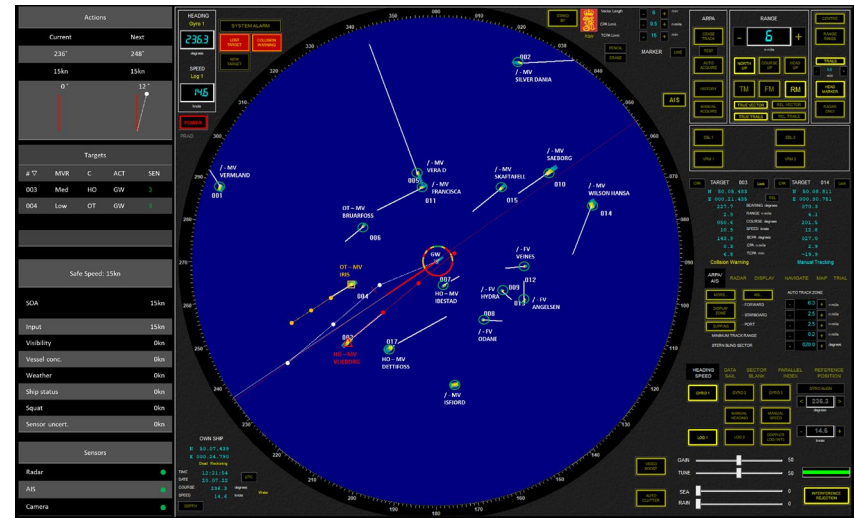
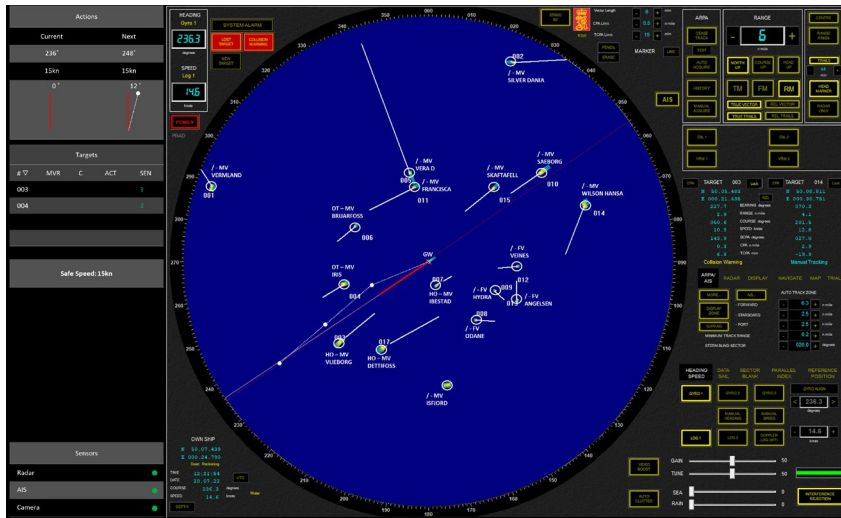
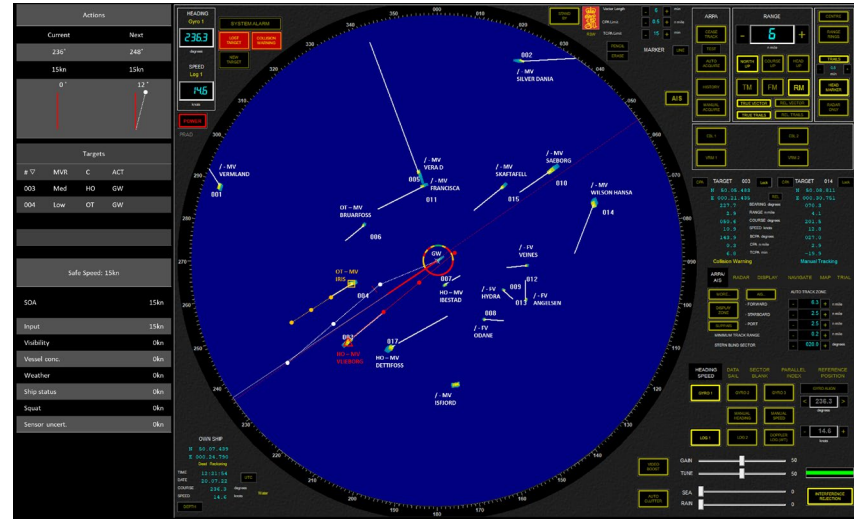
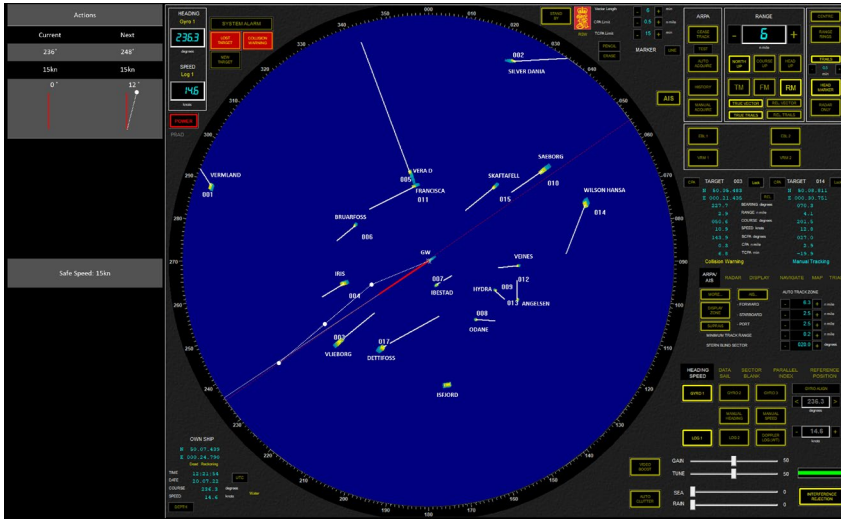
In the *medium (A) transparency variation* (top-right), it is depicted that the red target forms the highest risk, whilst the target in orange poses a potential risk during the avoidance manoeuvre. Additional indicators next to the target symbols indicate the type of conflict and vessel. Also, a risk circle depicts where own ship can manoeuvre within a one vector length. Finally, the factors influencing safe speed information is provided in table form on the left of the screen.

In the *medium (B) transparency variation* (bottom-left), all layers are depicted except the “information analysis” layer. Here, the system discloses its decisions, actions, and which information it has acquired. However, it does not provide information about how it analyses this information, e.g., which risks it has determined. This level of transparency was included to provide an alternative to the cumulative approach discussed above.

Finally, in the *high transparency variation* (bottom-right), all transparency information identified through the task analysis is provided on the HMI. Here, all targets have received identifiers (green circles), and initial classifications (ship types and relevant conflict type indicators). In addition, information regarding the status of the system’s sensors is provided in the tables to the left of the radar screen.

Figure 20. Traffic situations with varying transparency levels.

Top-left: Low, Top-right: Medium (A), Bottom-left: Medium (B), Bottom-right: High. For larger versions, see Appendix C – Examples of transparency levels.



## 4.5 Article 5: Evaluating with controlled experimentation

Article 5 presents the findings from a controlled experimental study analysing the relationship between agent transparency and key human performance variables in an autonomous conflict avoidance context (van de Merwe, Mallam, Nazir, et al., 2024b). Three dependent variables were measured (SA, mental workload, and task performance) as a function of transparency and traffic complexity. A detailed discussion of the results is depicted in the Results section of Article 5 in Appendix H – Publications. Note that the descriptions of the statistical results are predominantly similar to those presented in the Article 5. Still, in order to provide a complete picture of the experimental findings, the findings are repeated here. In addition, an overall and integrated presentation of the study's results is provided. The findings are also graphically depicted in Figure 21.

For level 1 SA, a main effect for transparency was found ( $F(3,31) = 9.374, p < .001, \eta_p^2 = .476$ ). The high transparency level ( $M_{high} = .71$ ) resulted in improved awareness of elements in the environment compared to the low transparency and the medium (B) condition ( $M_{low} = .46, M_{medium (B)} = .57$ ). No differences were found between the medium (A) condition and the other conditions ( $M_{medium (A)} = .60$ ). A main effect for complexity was found where traffic situations with high complexity indicate lower level 1 SA ( $F(1,33) = 30.347, p < .001, \eta_p^2 = .479; M_{low} = .70, M_{high} = .47$ ). A weak and non-significant interaction was found between the transparency and complexity conditions ( $F(3,31) = 2.885, p = .051, \eta_p^2 = .218$ ).

For level 2 SA, a main effect for transparency was found ( $F(3,31) = 10.572, p < .001, \eta_p^2 = .506$ ). The SAGAT level 2 scores for transparency level medium (A) ( $M_{medium (A)} = .84$ ) are higher than the low- and high condition ( $M_{low} = .65, M_{high} = .64$ ). Also, the scores for the medium (B) condition are higher than the scores for the low condition and did not differ from the medium (A) condition ( $M_{medium (B)} = .85$ ). This indicates that comprehension of the elements in the traffic situation was best when transparency was at a medium (A)- or at a medium (B) level. Furthermore, a main effect of complexity on level 2 SA was found ( $F(1,33) = 24.713, p < .001, \eta_p^2 = .428; M_{low} = .82, M_{high} = .67$ ). This indicates that a lower level 2 SA was achieved in high complexity cases compared to low complexity ones. Finally, a significant interaction between complexity and transparency was found for level 2 SA ( $F(3,31) = 3.206, p < .037, \eta_p^2 = .237$ ) showing significant differences in SA scores for medium (A) transparency and complexity.

For level 3 SA, a main effect for transparency was found ( $F(3,31) = 4.362, p < .011, \eta_p^2 = .297$ ). The scores on SAGAT were highest for the high transparency condition ( $M_{high} = .85$ ) and significantly higher than the low- and medium (A) transparency conditions ( $M_{low} = .73, M_{medium (A)} = .69$ ). This indicates that participants were best able to predict the future state of the elements in the environment with the highest level of transparency. No difference between the medium (B) transparency level and the other levels was found ( $M_{medium (B)} = .77$ ). A main effect for complexity was found in which the low complexity level resulted in higher scores on the SAGAT compared to the high complexity level ( $F(1,33) = 38.594, p < .001, \eta_p^2 = .539, M_{low} = .85, M_{high} = .67$ ). This means that the participants were better able to predict the future state of elements in the traffic situations when these were of low complexity compared to high complexity. No interaction between complexity and transparency was found for level 3 SA.

No main effect of transparency on mental workload was found. However, individual dimensions as measured through the NASA-TLX were analysed and showed an effect on the “Performance” sub-

dimension ( $F(3,28) = 7.791, p < .001, \eta_p^2 = .455$ ). Here, the participants reported they were more satisfied with “achieving the goals set by the experimenter” for the medium transparency level compared to the other levels (Hart & Staveland, 1988, p. 30). Also, a main effect for complexity on mental workload was found ( $F(1,33) = 21.964, p < .001, \eta_p^2 = .400; M_{low} = 55.29, M_{high} = 64.45$ ). This indicates that participants reported higher levels of workload in the high complexity cases compared to the low complexity cases. Finally, no interaction between complexity and transparency was found.

A main effect for transparency was found for mean TTC ( $F(3,22) = 24.73, p < .001, \eta_p^2 = .771$ ). The medium (A)-, medium (B)-, and high transparency levels ( $M_{medium (A)} = 52.62, M_{medium (B)} = 60.12, M_{high} = 60.80$ ) led to increased mean TTC compared to the low transparency condition ( $M_{low} = 38.40$ ). Also, medium (A)- and high transparency levels resulted in higher mean TTC compared to the low- and medium (B) levels. No difference in TTC was found between the medium (A)- and high transparency levels. For complexity, a main effect was found on the mean TTC ( $F(1,24) = 46.65, p < .001, \eta_p^2 = .66, M_{low} = 40.30, M_{high} = 53.14$ ). A high traffic complexity level resulted in increased mean times to comprehend the traffic situations for the participants. For the interaction between transparency and complexity no effect was found.

In terms of subjective ranking, a main effect of transparency was found ( $F(3,31) = 616.639, p < .001, \eta_p^2 = .984$ ). The medium (A)- and high transparency levels were preferred compared to the low- and medium (B) levels. The low transparency was rated the least preferred, followed by the medium (B) level, and a shared highest preference for the medium (A)- and high transparency level.

When integrating the results across the dependent variables, Figure 21 graphically depicts the graphs found in Article 5 but now in conjunction and side-by-side. For complexity, its effect in the dependent variables appear primarily uniform. As supported by the statistical data discussed earlier, higher complexity levels result in lower SAGAT scores compared to lower complexity levels, regardless of SA level. Also, for mental workload and task performance, higher complexity levels lead to increased mental workload and TTC respectively. Only for the combination between level 2 SA and the medium (A) transparency level, no differences were found.

For transparency, Figure 21 depicts that its effect differs from variable to variable and from level to level of transparency. For level 1 SA, high transparency provided the highest level 1 SA, for level 2 SA, medium (A) and medium (B) transparency levels was the highest, and for level 3 SA, the high transparency level provided the highest level 3 SA. This implies that for medium (A) to high levels of transparency increased levels of SA can be expected. This result aligns with the subjective rankings which indicated that participants preferred the medium (A)- and high transparency levels compared to the low- and medium (B) levels. However, these SA levels came with a task performance penalty; increased comprehension times were recorded with medium (A) and high levels of transparency. Nevertheless, for mental workload, no effects were found for any of the levels of transparency.



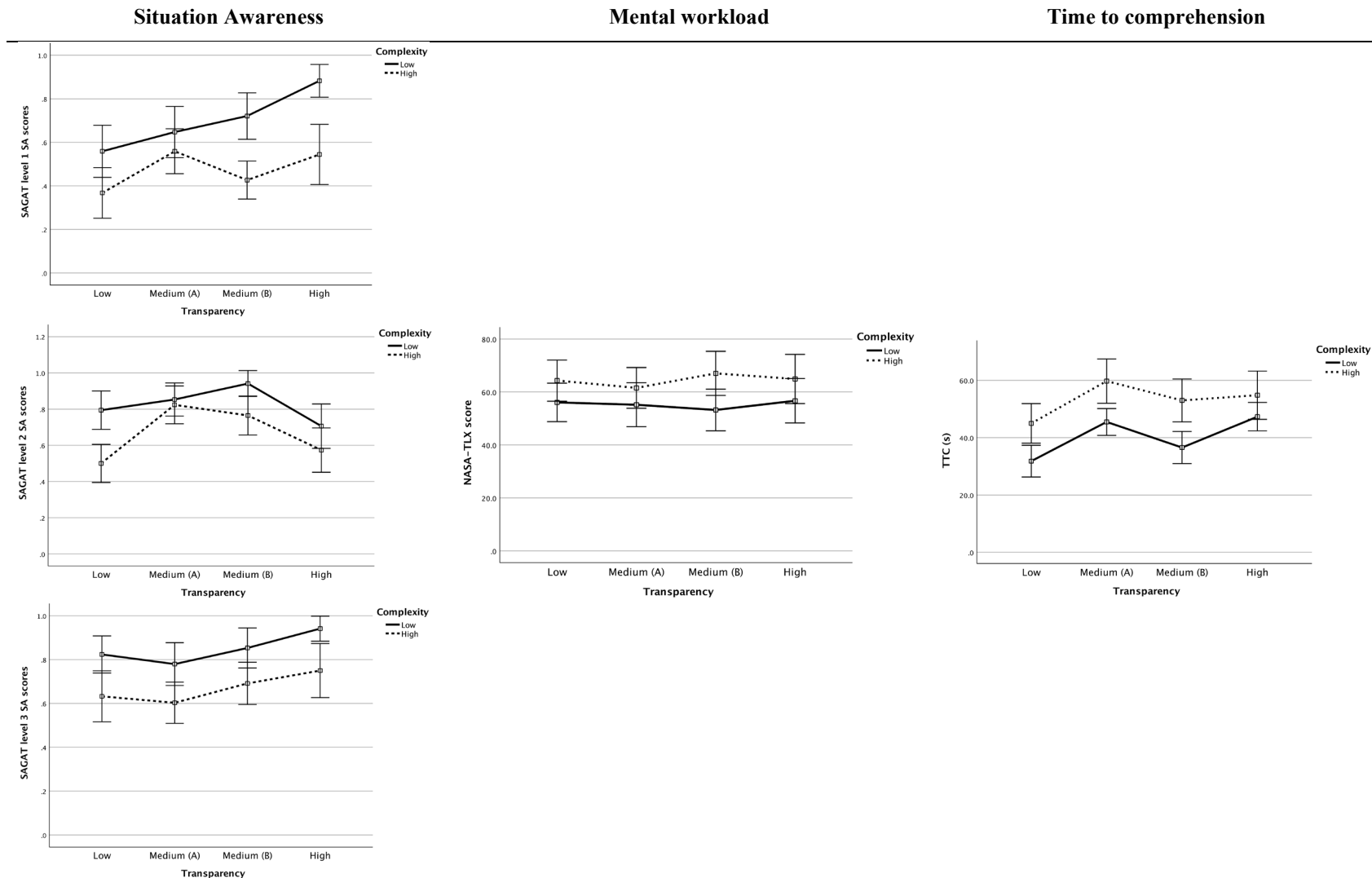


Figure 21. Mean scores for level 1, 2, and 3 SA, mental workload, and time to comprehension as a function of transparency and complexity.

Note the error bars represent the 95% confidence interval.



## 5 Discussion

The following section reflects on the results of the studies performed in this dissertation as well as the methods applied to generate these. As for other parts of this dissertation, relevant sections of the articles are repeated and summarised below. This means that, given the similarities between the descriptions below and the respective sections in the articles, some reproduction of texts, figures, tables, as well as similarity in wording of sentences and paragraphs should be expected. For clarity, references to the appended articles are provided where relevant.

### 5.1 Theoretical reflections

This section discusses the overall results from the studies performed in this dissertation. In the theoretical background, this dissertation argued that to ensure equivalent safety in “alternative and equivalent” concepts where autonomous functions perform tasks previously performed by humans, special focus should be given to supporting cognitive processes required for adequate and effective human supervisory control (DNV, 2021; IMO, 2013). This was considered relevant as supervisory control of autonomous agents is challenged by the problem of attaining and maintaining SA, resulting in the OOTL performance problem (Endsley & Kiris, 1995). One of the proposed alleviating measures to combat this was the application of transparency principles in the design of autonomous agents (Endsley et al., 2003; Meister, 1999). That is, by providing insight into the agent’s reasoning process, through providing direct SA knowledge on the agent’s display, human supervisors would be able to uphold their SA assessment process and consequently maintain agent SA.

To support this claim, emerging evidence was reported in terms of SA, albeit at a potential cost in mental workload (Bhaskara et al., 2020). However, as reported in a separate review on transparency models, validation efforts for the transparency models have largely been incomplete or have provided inconclusive evidence (Rajabiyazdi & Jamieson, 2020). Therefore, this dissertation started with obtaining an overview of the evidence regarding transparency, SA, mental workload, and task performance by systematically mapping and assessing scientific studies addressing these variables (Article 1; van de Merwe, Mallam, & Nazir, 2024).

The principal finding from Article 1 was that there is a promising effect of agent transparency on SA and task performance, without affecting mental workload, for tasks involving responding to proposals and supervision. However, the detailed findings are rather nuanced and probably the most interesting to discuss here are those studies in which SA, mental workload, and task performance were measured in conjunction. Good SA, without excessive mental workload, increases the probability for good operator performance (Endsley, 1995; van de Merwe et al., 2012; van Doorn et al., 2021). Therefore, studies that have measured these constructs in tandem may provide insight into the effect of transparency on the combined effect on these variables. In Article 1, five of the 17 studies measured these constructs together and for three of these improved SA and neutral or reduced workload were found, in combination with improved response times (T. Chen et al., 2015; Roth et al., 2020, for SAGAT only; Skraaning & Jamieson, 2021, experiment 1 and 2), goal achievement (T. Chen et al., 2015; Skraaning & Jamieson, 2021, experiment 1 and 2), and detecting process deviations and performing verifications (Skraaning & Jamieson, 2021, experiment 1 and 2). Others have found increased workload scores, but no effect for SA and task performance (Guznov et al., 2020). Finally, Skraaning and Jamieson (2021, experiment 3) found no effect for SA, workload, and task performance, and even reduced performance when participants were detecting and verifying events

from a plant-wide agent-like automation. Here, transparency information was made available on dedicated screens in the control room and the authors argued that this setup may have negatively affected performance measured. It also indicates that any benefits from transparency may be affected by agent type and how information is made available to operators. That is, information that is integrated in primary task displays reduces the effort needed to keep information elements in working memory which, in turn, benefits other tasks and information elements competing for the same resources (van Doorn et al., 2021; Wickens, 2008).

These findings indicate the intricate nature of the relationship between human performance and transparency. That is, the effects of transparency depend on the way transparency is defined, modelled, operationalised, and measured. Consequently, these results were interpreted as an incentive to acquire more knowledge around this construct. Also, since the literature review did not contain studies from the maritime industry, these findings were interpreted as an additional incentive to investigate the applicability of transparency principles for this domain. Therefore, this dissertation aimed to address these challenges, elucidate the transparency concept, and apply it to the maritime collision avoidance domain.

The articles appended to this dissertation highlight the approach, models, and methods that were applied throughout this research and their results are discussed earlier. Especially Article 5 addresses the relationship between transparency and the aforementioned key human performance variables, but now applied to the maritime domain (van de Merwe, Mallam, Nazir, et al., 2024b). Here, an experiment was performed that addressed levels of agent transparency applied to a collision avoidance system and where the same key human factors variables were measured: SA, mental workload, and task performance. For this study, and based on the findings from the literature review, it was predicted that transparency would have a positive effect on SA and task performance without affecting mental workload. An additional independent variable was added to evaluate potential interaction effects. For complexity, it was predicted that a negative effect on SA, mental workload, and task performance would occur. Finally, it was predicted that higher transparency levels could mitigate the effect of complexity on SA and task performance as increased transparency information would alleviate the task of building agent SA. No interaction effect was predicted for mental workload (see Table 9).

Table 9. Summary of predictions and results as discussed in Article 5.

Measure	Impact of transparency	Results match prediction?	Impact of complexity	Results match prediction?	Transparency x complexity (interaction)	Results match prediction?
Situation Awareness	Improved SA with higher transparency levels	Level of SA 1 - <b>Yes</b> 2 - <b>Yes</b> 3 - <b>Yes</b>	Reduced SA with higher complexity	Level of SA 1 - <b>Yes</b> 2 - <b>Yes</b> 3 - <b>Yes</b>	Higher transparency levels may negate effect of higher complexity	Level of SA 1 - No 2 - <b>Yes</b> 3 - No
Mental workload	No effect predicted	<b>Yes</b>	Increased mental workload with higher complexity	<b>Yes</b>	No interaction predicted	<b>Yes</b>
Task performance	Improved task performance with higher transparency levels	No	Reduced task performance with higher complexity	<b>Yes</b>	Higher transparency levels may negate effect of higher complexity	No

### 5.1.1 Situation Awareness

Although the precise effect of transparency on SA depends on the combination of level of transparency and complexity, the overall results indicate improved SA scores with increased levels of transparency (see Table 9). However, as elaborated on in Article 5, interpreting the detailed findings shows nuanced variations.

For level 1 SA, it was found that increased level 1 SA can be achieved with increased transparency. In Endsley's definition of SA (1995), level 1 SA is concerned with the perception of elements in their environment and provides the foundation for the higher levels of SA. This means that when the system depicts which information it has detected, this should support the participants' level 1 SA. Here, the results indicate that the highest level was achieved in the medium (A) and highest transparency condition. Interestingly, the medium (B) condition scored lower than the medium (A) and high transparency conditions. It was expected that when the system depicts which information it has detected, this should support the participants' level 1 SA. This may indicate that the information provided in the "condition analysis" step (absent in the medium (B) transparency condition yet present in the medium (A) condition) may have played a role in achieving improved level 1 SA. Possibly, the additional information regarding collision risk, e.g., risk objects, intended trajectories, and priorities, may have directed the participants' attention towards the ship's surrounding traffic and thus better able to achieve level 1 SA.

For level 2 SA, medium (A) and medium (B) transparency levels resulted in the highest SA scores. For medium (A) transparency, this was as anticipated as the system's analytical information is depicted at this transparency level, and having this information readily available on the HMI should support this type of SA. However, for the medium (B) transparency level, this type of information is not available, yet participants score equally well on level 2 SA. In addition, analytical information is

also available in the high transparency condition, yet the level 2 SA scores are significantly lower than the medium (A) and the medium (B) levels. For example, at the medium (A) level of transparency, the system depicts which objects it sees as posing a collision danger by extrapolating the objects' current vector and highlighting the level of risk using specific symbology and colours. This way, participants could directly perceive the outcomes of the system's risk analysis process and use this information to understand the system's reasoning. Perhaps, the information associated with the medium (B) condition was sufficient to deduce the system's internal reasoning. Also, it is possible the amount of information associated with the high transparency condition may have distracted the participants to such an extent that their level 2 SA was affected.

For level 3 SA, the highest level of SA was achieved with the highest level of transparency. Also, no differences were found between the low and medium (A) level of transparency. To support level 3 SA, the transparency levels provided participants with the CAGA system's future state prediction of own-ship and target objects. The future state of own ship, i.e., its future track and speed, was depicted for each level of transparency. The future state of target ships was depicted for the medium (A)- and high transparency levels but not for the other levels. Therefore, it was expected that level 3 SA would either be improved across all levels of transparency, or only for the medium (A)- and high level. However, considering that the highest level of level 3 SA is only achieved at the highest level of transparency, makes this result not straightforward to interpret. A possible explanation is the "completeness" of the information depicted in the highest level of transparency: its decisions and planned actions, its analysis, its perception of the environment, and its sensor states. Here, all transparency information is provided which may have resulted in an improved understanding of all parameters by the system. This way, participants may have used this complete picture to reason towards the correct answer when answering the level 3 SA query in the SAGAT.

Comparing these results to the results from the studies in Article 1, comparable results were found. For example, Roth et al. (2020) found that level 3 SA was most improved in the high transparency condition compared to the low condition, when participants were evaluating agent-generated proposals in an unmanned-manned helicopter teaming operation. Chen et al. (2014, 2015) found improvements in SA when participants were supervising unmanned aerial vehicles in a search operation. Also, Selkowitz et al. (2017) reported improved level 2 and level 3 SA when monitoring an autonomous robot, but not level 1. Still, despite some studies failing to identify a relationship between transparency and SA for supervision (Skraaning & Jamieson, 2021; experiment 3) and monitoring tasks (Pokam et al., 2019; Selkowitz et al., 2015; Wright et al., 2020), the overall results point toward a neutral to positive relationship between transparency and SA. The results from the experiment in Article 5 have strengthened these findings.

### **5.1.2 Mental workload**

In this study, no effect of transparency was hypothesized for mental workload, and the data indicates that none was found. However, for complexity, increased workload scores were found for all high complexity traffic situations. Finally, no interaction effects with transparency were found either.

When comparing these results from the studies in Article 1, limited effects of mental workload were also found for those experiments in which participants were tasked with monitoring an autonomous agent (e.g., Du et al., 2019; Selkowitz et al., 2015, 2017; Wright et al., 2020). In studies where participants were acting as a supervisor, either reductions in workload (T. Chen et al., 2014, e.g.,

2015; Skraaning & Jamieson, 2021; experiment 1 and 2), increases in workload (Guznov et al., 2020), or no effects were found (Skraaning & Jamieson, 2021; experiment 3). In studies where participants were asked to respond to system-generated proposals, no effect on mental workload was reported (e.g., Bhaskara et al., 2021; Loft et al., 2021; Mercado et al., 2016; Roth et al., 2020; Stowers et al., 2020). Across all the studies in Article 1 in which mental workload was measured, 17 out of 23 workload indicators did not show a relationship with transparency (van de Merwe, Mallam, & Nazir, 2024). Considering these results, this experiment does not change the overall conclusion that adding transparency information does not affect mental workload.

### 5.1.3 Task performance

The overall results for task performance show that operators take more time in building up a mental picture for the medium (A)- and high transparency conditions and less time in the low- and medium (B) transparency conditions. This was the case for both the low- and high complexity conditions indicating an equal effect of traffic complexity regardless of transparency level.

For this experiment, the results indicate a trade-off between the effort and performance where participants with higher level 1- and level 3 SA scores also used more time to comprehend the traffic situations, albeit without increased mental workload. Also, when participants took more time to analyse the traffic situations, the effect of complexity was negated for level 2 SA. In Article 5, the time to comprehension was driven by the instruction for the participants to “continue to the next step when you feel you have built up a sufficient understanding of the traffic situation”. Since the participants consisted of professional navigators that were tasked with supervising an autonomous ship, it may be argued that they had a professional interest in performing this task as accurately as possible. This means that when the CAGA system’s analytical information was depicted on the HMI, arguably the most safety-critical information the system is able to depict, they used their time to comprehend the system’s reasoning behind its avoidance action. Possibly, one of the main drivers of the increased comprehension time was that they were comparing the system’s analysis to their own. Conversely, when the CAGA system did not provide analytical information, no increase in TTC was found, possibly because there was less information to compare. Similar observations have indeed shown the increased need for comparing information when systems provide recommendations (Endsley, 2017). Given the potential safety critical role of humans in supervising autonomous ships, this indicates the importance of addressing the type of information in developing transparent CAGA systems, and not the amount only (van de Merwe, Mallam, Nazir, et al., 2024a).

### 5.1.4 Complexity

For complexity, the results are as anticipated, i.e., increased complexity results in reduced SA scores, increased workload, and reduced task performance. This result can be traced back to the information processing model and the processes of selective, divided, and working memory. As anticipated, high complex traffic situations, consisting of a larger number of ships and more complex collision situations, increase the burden on the human information processing system. Here, increased amounts of information are perceived, processed, and temporarily stored, compared to low complexity situations, and this has clearly affected SA scores, perceived mental workload, and task performance. However, this effect appears to be negated for the medium (A) transparency condition. Here, no differences were found between low- and high complexity traffic situations in terms of level 2 SA. This seems to indicate that transparency is able to alleviate some of the challenges associated with

obtaining and maintaining level 2 SA knowledge, i.e., the understanding of the meaning and relationships of perceived information elements, in high complexity cases.

Earlier researchers have debated the relationship between SA and mental workload and the need to consider both constructs in the assessment of human performance (Endsley, 1995; Parasuraman et al., 2008; Vidulich, 2000). According to Vidulich and Tsang, mental workload is generally characterised “in terms of the level of attentional demands on placed on the operator in course of performing required tasks”, and SA is “primarily associated with the informational content of the operator’s memory system during task performance” (2014, p. 95). Furthermore, they argue that mental workload and SA often compete for the same resources and are therefore both supported and limited by the these. As such, the more demanding and complex a task is, the more work is required to perform it and assess the situation. This implies that high mental workload could lead to poor SA because of the way the limits of the information processing system affect the amount of attentional resources that can be allocated to the information. Also, the amount of information that can be actively kept in working memory, as part of SA assessment, affects the degree to which SA can be obtained and maintained. However, mental workload and SA are influenced by other factors than information amount and complexity only.

Mental workload is influenced by factors such as level of automation, its degree of adaptiveness, and granularity of control, task difficulty, and time pressure when interacting with autonomous systems (Endsley, 2017; Galy et al., 2012). SA is influenced by, amongst others, workload, engagement, mental models, and the automation interface (Endsley, 2017). In the case of this dissertation, the HMI was manipulated to include transparency information, integrated in the primary task display and congruent with relevant HMI standards for the radar display, such that system internal information could be disclosed to the operator. As the participants were tasked with making sense of the information provided by the collision avoidance system, it could be anticipated that the additional information burden provided by complexity would negatively affect SA, mental workload, and task performance. This was indeed the case for these variables. However, for level 2 SA, it was found that providing the system’s analytical information directly on the HMI, the participants were freed from processing the additional information elements imposed by high complexity traffic situations, resulting in improved scores. Similar to Van Doorn et al. (2021), this “display-based” information processing, resulted in improved performance compared to “memory-based” information processing, at least for level 2 SA.

## 5.2 Practical reflections

Besides routed in theoretical knowledge, the activities and results of this dissertation have a strong practical connection to the developments of autonomous systems and the role of the human herein. Reflecting on the work, two aspects are highlighted: designing for transparency, and assessing transparent designs.

### 5.2.1 Designing for transparency

The research performed in this dissertation followed a development process by establishing the research and application context, identifying SA requirements, developing concepts, and evaluating designs. Although this dissertation should not be considered a graphical design-oriented study, the novelty of the ship autonomy field and the lack of readily available CAGA systems meant a design process still was needed to produce the transparency levels and integrate these into HMIs. Considering



that the overall research objective of this dissertation was to evaluate how transparency can support human performance in the context of autonomous ships, the path towards achieving this aim reflects a Human-Centred Design (HCD) process.

HCD is an approach where capabilities and human needs are taken into account in system design (Endsley et al., 2003). In contrast with technology-centred design, HCD considers the user's tasks, context, and capabilities to inform system design, e.g., for designing human machine interfaces. The aim of this type of design is to create more effective systems, by reducing errors and improve productivity, whilst increasing user acceptance and satisfaction. To achieve a HCD process, the following main activities should be performed: The system's context of use is specified and understood, user requirements are specified, design solutions are produced, and finally, the design is evaluated (ISO, 2019). Through a series of iterations, where user requirements and design solutions are updated and evaluated, a final solution is produced that meets the user requirements.

As depicted in Figure 22, Article 1 specified and described the dissertation's overall context through a systematic literature review, i.e., transparency in safety critical domain (van de Merwe, Mallam, & Nazir, 2024). Here, a mapping of the scientific context was performed by addressing how transparency has been researched and which effects are reported concerning the relationship between transparency and key human factors variables. Article 2 established, described, and analysed the specific context for this dissertation, i.e., collision avoidance within the maritime domain (van de Merwe, Mallam, Nazir, et al., 2024a). Here, in-situ observations and interviews with navigators onboard ship bridges, document reviews, and COLREG assessments formed the basis for identifying cognitive tasks and information requirements for conventional and supervised autonomous collision avoidance. Subsequently, Article 3 applied a contextualised model for human information processing to structure and organise these requirements such that autonomous collision avoidance systems could be made transparent to their users (van de Merwe et al., 2023b). Article 4 used these structured requirements to develop HMI design concepts (van de Merwe et al., 2023a). Based on the model for information processing, symbology was developed and integrated in a conventional radar display to create levels of transparency for a selection of traffic situations. Navigators were used to independently evaluate the designs throughout their developments. Finally, Article 5 closed the HCD development loop by experimentally evaluating the effect of transparency on key human factors variables. This way, the outcome of this research was connected to the dissertation's wider context, i.e., transparency in safety critical domains.

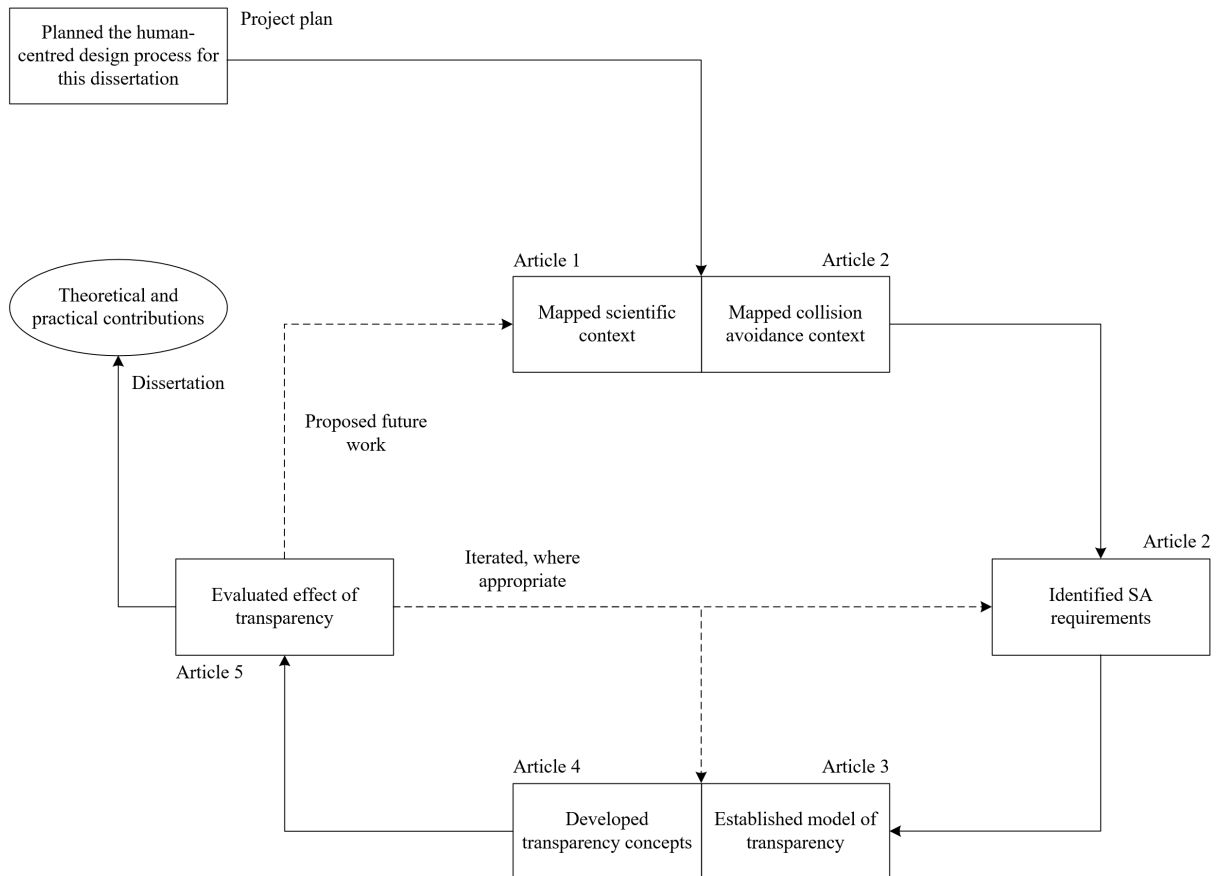


Figure 22. The development process as applied in this dissertation.

In executing the development process for this dissertation, a critical element in designing for transparency was the choice of the underlying model. As discussed, the PSW model was adapted and contextualised to be used as a model for transparency (Parasuraman et al., 2000; van de Merwe et al., 2023b). By defining layers of transparency, the structure of the transparency concepts was shaped. In other words, the steps in the information processing model defined the amount of transparency layers that were developed in the concept phase and evaluated in the experiment. Had a different model been chosen, different transparency concepts would have been produced.

For example, as described in the literature review in Article 1, Stowers et al. (2020) used a cumulative approach based on the SAT model in which each subsequent level of transparency added new information to the previous one, i.e., similar to the approach used in this dissertation. However, here, unmanned vehicle capabilities were conveyed through the size of icons depicting the speed of the vehicle (level 1+2). In addition, the vehicle's time to meet its goals was added to convey the vehicles reasoning behind its actions (level 3). Finally, uncertainty information was added as a separate and final layer as visualised through changes in the opacity of the vehicle's icon (level 3 + uncertainty). Alternatively, in the study performed by Pokam et al. (2019), the HRT model by Lyons (2013) formed the basis for the design. This model assumes that, for human-agent systems to be transparent, information should flow from the agent to the human, and vice versa, depicting the agent's intent, tasks, analysis, and its understanding of its environment. Based on these principles, levels of transparency were developed and evaluated in a driving simulator. In this study, the HRT model was used to define information elements and classify these according to a model of information

processing: information acquisition, information analysis, decision making, and action execution. Although similar to the model in this dissertation, here, five HMIs were developed depicting degrees of transparency of the autonomous vehicle. The first level of transparency served as the control condition and did not depict any system internal information to the participants. The second level provided information about the system's information acquisition and action execution. Third level added the systems analytical information, the fourth added its decision making (but did not depict its action execution), and the fifth level provided full insight into the system's information processing levels. Finally, the literature review identified several studies that did not base their designs on any model of transparency. Here, transparency was implemented based on how the concept was interpreted and which definition was adopted. For example, Skraaning and Jamieson (2021) interpreted transparency as the observability of responsibilities, capabilities, goals, activities, and/or effects of automation in the human-machine interface. In three separate nuclear control room experiments, transparency was operationalised as adding information about automation activities on process screens, provided verbal feedback about system status, and depicted the state of the automation's status, actions, and progress along procedural steps.

As the literature review has shown, research regarding the utility of transparency as a design principle is ongoing (van de Merwe, Mallam, & Nazir, 2024). Although the outcome of a HCD process is highly context dependent, the process itself is matured, standardised, and widely used. However, the process for deriving transparent designs supportive of effective supervisory control is less so. This means that research is needed to consolidate definitions, models, and design processes for transparency. Nevertheless, this dissertation has shown that, despite the current state of transparency research, the HCD process is an effective method for developing transparent systems and providing operators with insight into what the system is doing, why it is doing it, and what it will do next (Endsley, 2017). This means system designers, tasked with developing transparent systems, have an established and viable process at their disposal to develop systems supportive of supervisory control through transparency. However, for the moment, they should be mindful of the choices around the transparency model and the consequences this has on the design of their systems.

## **5.2.2 Assessing transparency in autonomous ship concepts**

As introduced earlier, guidance developed by the relevant authorities and classification societies aim to support concept submitters in seeking approval for the approval of autonomous ship concepts (American Bureau of Shipping, 2022; Bureau Veritas, 2019; DNV, 2021; NMA, 2020). For example, as described in DNV's class guidance for autonomous and remotely operated ships, the distribution of functions between humans and machines should result in equal or better capabilities than conventional solutions (DNV, 2021; IMO, 2013). This means that "the concept of operations should clearly describe all the operational tasks that the vessel will undertake that will be either fully or partly automated" (DNV, 2021, p. 27) as well as how human-system integration will be designed (American Bureau of Shipping, 2022). Also, "when a human is in charge of decision making, the location of the decision maker should be clearly described" (DNV, 2021, p. 27). For autonomous ships concepts, this will typically be on-board, from a remote location, or a combination. Furthermore, special focus shall be placed on the ability for the human operator to establish sufficient SA such that corrective actions can be taken (Bureau Veritas, 2019). To achieve this, "the interface between the system and the human users shall be designed according to best practises for user interfaces and with defined responsibility modes for the operator (DNV, 2021, p. 42).

This means that concept submitters, aiming to provide autonomous collision avoidance services, should propose solutions that define the function and task distribution between the collision avoidance system and the human operator. In addition, submitters should develop HMI solutions that support operators in their task of supervisory control, regardless of the location this task is performed from. Finally, the concept submitter should provide evidence indicating that the overall system delivers an equivalent level of safety, or better, in order to receive approval of the autonomy concept by a relevant authority. Given the application of intelligent agents, with advanced sensory-, information processing-, decision making-, and action capabilities, this dissertation argued that insight into the agent's reasoning processes is a crucial element to consider in the development of such concepts. Likewise, this means that the relevant authority should establish clear expectations and acceptance criteria towards the approval of such concepts with regards to transparency.

Based on the work performed in this dissertation, criteria may include expectations with regards to the provision of clear, unambiguous, and timely information for the supervisor with regards to insight into the system's decisions, planned actions, and the reasoning for these. For collision avoidance systems, examples include depictions of own ship's intended avoidance manoeuvre, when to execute the manoeuvre, its assessment of priority (give-way or stand-on), which actions are expected of the target ship, which objects and type it has detected, and any restrictions to manoeuvrability. Furthermore, the system should provide its interpretation of collision risk and the reasons for performing these actions, e.g., because of critical CPA, TCPA, or other risk indicators, such as the available sea room. Moreover, the collision avoidance system should state which COLREG situation is appropriate given the situation, e.g., head-on, overtaking, or crossing, and for which target ship this is relevant for. Finally, given the dynamic nature of maritime collision avoidance, the relevance of this information should be evaluated over time, i.e., how long the information depicted on the HMI is relevant for, to allow for sufficient time for supervision and possible intervention.

In addition to these *product*-focussed requirements, the concept approver should also state clear expectation regarding the development *process* for creating transparent systems. As mentioned above, the interface between the autonomous system and the user shall be designed according to best practices for user interfaces. The HCD process provides an established, standardised, and widely used method for managing the development of hardware and software components of interactive systems (ISO, 2019). The rationale for application of HCD principles is to develop systems that enhance human-automation interaction, reduce the probabilities for errors, and increase productivity (Human Factors and Ergonomics Society, 2023; ISO, 2019; Lee et al., 2017). This dissertation has applied the HCD process and provided evidence for the value of transparency as a design principle to support supervisory control of autonomous collision avoidance. This means that, in addition to setting requirements to transparency as an end-product, concept approvers should set expectations on the process for how to achieve transparency.

### 5.3 Methodological limitations and reflections

The activities performed in this dissertation were planned, executed, analysed, and documented with the aim to produce high-quality knowledge. However, reflecting on the work, a number of limitations are discussed that should be considered in interpreting the dissertation's results and conclusions.

### 5.3.1 Systematic literature review

Although the use of SLRs has the potential for increased accuracy and improved reflection on the research field, it is not without its limitations (Mulrow, 1994). Performing a SLR means making choices for each of the steps of the method which have the potential to influence the accuracy and outcomes of the study. This includes, defining eligibility criteria, information sources, search terms, data selection process, data analysis process, and data reporting (Moher et al., 2015). For example, by limiting the information sources to only include peer-reviewed journal articles, one should be aware of the potential for missing potentially relevant data published in channels outside of this criterium, e.g., reports from research institutes, or non-peer reviewed project reports. Also, whilst setting eligibility criteria limits the amount of data, it may also exclude and overlook other relevant publications, e.g., from non-safety critical domains. Similarly, selection and analysis processes used for synthesising the data may exclude relevant information elements, e.g., which variables to report on. Finally, choices on how to report on the data may influence the results of the SLR, e.g., to perform a narrative synthesis of the data or perform a quantitative meta-analysis (Paré et al., 2015). The effect of these choices means that making different choices throughout the process will likely affect which data is represented, which results are synthesised, and how these are reported.

In this dissertation, a search string was defined and applied to three relevant databases that aimed to cover the breadth and depth of the research topic. Only research articles written in English that performed experimental studies within the safety-critical domain were included. Based on a full-text review, a narrative synthesis was chosen that reported on the overall outcome of the results. That is, as several of the studies did not report sufficient statistical data to allow for a meta-analysis, a qualitative synthesis approach was taken to analysis, summarise, and report on the findings from the literature. Although these considerations are made explicit here, limitations are inherent to performing research in general and SLRs in specific (Booth et al., 2016). That is, throughout research activities, choices are required that will affect how research is performed and will, to an extent, shape the outcomes of the research activities. The strength of the PRISMA method is that it requires these choices to be made explicit such that fellow researchers are made aware of these and can take these into account in their interpretation of the study's outcomes (Moher et al., 2009, 2015). Also, by being transparent and explicit about the choices, others may attempt to reproduce the study by following the same approach and using the same search terms, databases, inclusion- and exclusion criteria, selection process, and etcetera. This dissertation, and the appended publication, has made explicit the criteria and choices that were made in assimilating the research data into an overview of the research regarding the relationship between transparency and key human factors variables. In addition, three researchers performed the full-text selection, thereby strengthening the validity of the study's results (van de Merwe, Mallam, & Nazir, 2024).

### 5.3.2 Goal-Directed Task Analysis

This dissertation used various data sources as input to perform a GDTA. Among these, in-situ observations and -interviews were used where the collision avoidance task was studied onboard the bridges of passenger ferries. Here, a semi-structured interviewing technique was used that consisted of questions pertaining to the collision avoidance task, in addition to questions probing the use of a hypothetical autonomous collision avoidance system.

Nine licensed navigators were interviewed with a mean experience 30.6 years. Although critics may argue that "the more interviewees, the better", in this study, the number of interview subjects was

driven by “information power” (Malterud et al., 2015). This concept suggests that the more information a sample holds, the lower the number of participants are needed to meet the goal of the analysis. To determine the information power of a sample, Malterud suggests considering the following aspects: the specificity of the study’s aim, the specificity and experience of the participants, the amount of a-priori information available, the quality of the dialogue, and finally, the specificity of the case under investigation. For this study, this meant that, first, this study’s scope was considered narrow as it was limited to understanding collision avoidance manoeuvring only. Second, the population was limited to licensed navigators only, a highly specialised nautical position. Third, collision avoidance manoeuvring is highly regulated through internationally enforced rules and regulations and ample knowledge, experience, and documentation is available. Fourth, the interviews, based on a semi-structured protocol, were performed onboard ship bridges and were augmented by in-situ observations, which supported the dialogue between the interviewer and the participants. Finally, the scope of this study was to gain an in-depth understanding of the collision avoidance task only and was therefore limited and practical in scope. Based on this reasoning, it was deemed that the number of interview participants provided a sufficient information basis for the analysis. Also, as discussed earlier, two independent navigators reviewed the results of the GDTA, thereby strengthening the quality of the data.

Due to complexities involved in ensuring obtaining consent from participants and ensuring anonymity of data gathering, it was decided not to perform voice or video recording of the interview data in Article 2. Because of the operational environment in which this study took place, there was little control over when personnel exited or entered the ship’s bridge. In addition, recording of voice data would inevitably lead to unintentionally recording radio communications for which consent was challenging to obtain. Finally, as each interview was performed while the participants were on duty and question could only be asked when they were available, each interview lasted between four to six hours. This made voice or video recording also a practical challenge. It was therefore decided to record the data by taking notes instead. Because of the extensive time available for the interviews, pauses were used to write out the notes, identify gaps, and prepare for the next questions. In addition, observations of collision avoidance manoeuvring were used to cross-check the interview notes and revisit earlier questions. This way, the quality of notetaking was enhanced without the need for voice recording and subsequent transcribing. Finally, to ensure validity of the captured data, the results of the GDTA were validated with two independent navigators (van de Merwe, Mallam, Nazir, et al., 2024a).

### **5.3.3 Human information processing model**

It is recognised that the operationalisation of the agent’s inner reasoning as levels of transparency, including the ways in which it can be traversed, is highly dependent on the choice of model and what is considered vital information. Here, it was considered that “action planning” was the most critical information and therefore this would be presented as the first layers of transparency. Also, the choice of an information processing model dictated the number of layers of transparency and type of information represented by these. Although the application of this approach is sensible in this context, it is recognised that other models will produce different operationalisations. For example, if the SAT model was chosen as the basis for transparency, the first layer of transparency would represent “What’s going on and what is the agent trying to achieve?”, the second layer “Why does the agent do it?”, and the third layer “What should the operator expect to happen?” (J. Y. C. Chen et al., 2014, p. 2). Here, the first layer would represent the agent’s purpose, processes, and performances. The second

layer would represent its reasoning process and constraints. The third layer would represent its projections to a future or end state, and its potential limitations, including its likelihood of error and history of performance. Although there is overlap between this model and the model chosen for this dissertation, their differences would likely have resulted in alternative operationalisations of transparency. Future work could explore developing designs based on different transparency models to evaluate which of these would have most merit given the respective context and task.

### 5.3.4 Human Machine Interface development

As discussed in Article 1, operationalisation of transparency depends on the context and the task (van de Merwe, Mallam, & Nazir, 2024). For example, when comparing the graphical design of transparency concepts for the control room experiment by Skraaning and Jamieson (2021) and Wright et al. (2020) it is clear designs were chosen that were aligned with the user interface philosophy in which transparency elements were integrated. In this dissertation, there were several principles the design and implementation of transparency elements adhered to, i.e., compatibility of transparency elements with existing design principles and integration of information in primary task displays. The aim of this approach was to avoid overloading the navigators with transparency information and thereby artificially confounding the results of the experiment. In other words, the aim was to limit the effect of transparency information on working memory capacity, selective, and divided attention (Lee et al., 2017; Wickens et al., 2013). This was achieved by employing a navy-certified navigator to develop traffic situations and HMI symbology based on a set of explicit design criteria, iterative design process, and independent validation with two navigators.

In developing transparent and integrated HMI concepts, based on the contextualised information processing model, decisions were made related to how to graphically represent the various information layers dictated by the model. As the model consist of four steps: condition detection, condition analysis, action planning, and action control, there a several ways in which these steps could be represented. The most straightforward approach would be a cumulative approach starting with the first step of the model (condition detection) and subsequently add information from the next steps as layers of transparency. This would have meant that the minimum transparency layer would represent the agent's perception of its environment. The next step would provide analytical information, whereas the last step would provide the agent's decisions and planned actions. (Note that in the last step "action control", no information is processed and is thereby not represented as a transparency layer.) However, considering collision avoidance is about avoiding collisions, it was deemed that such an approach would be little supportive of the supervisor's understanding of the agent's decisions and planned actions. That is, as the agent's plans for performing avoidance actions would only be made visible with all other information it was processing, this approach would potentially result in increased demand on the supervisor's information processing capabilities and would be in disagreement with the aforementioned design principles. Therefore, it was reasoned that a more fruitful starting point for providing transparency to supervisors would thus be the "action planning" step of the information processing model and not the "condition detection" step. This way, the supervisor would be informed of the agent's most crucial information, i.e., its decisions and planned actions to avoid a collision, and the additional underlying information would be made available by going "backwards" through the model.

### 5.3.5 Controlled experiment

In Article 5, an experiment was performed that aimed to assess the effect of transparency on SA, mental workload, and task performance. Through experiments, the existence of a causal relationship between variables could be determined by manipulating independent variables, measuring their effect on dependent variables, whilst controlling for extraneous variables. In contrast with quasi-, natural-, and field studies, controlled experiments are better able to tease out the effect of one variable on another in terms of its type and size (Coleman, 2019; Kirk, 2013). However, results from controlled experiments may not be transferable to situations other than those that were tested (Kirk, 2013). For this experiment, although efforts were made to ensuring the experiment was as representative as possible for the autonomous navigational context, the experiment used static images to represent traffic situations. Participants were asked to try to comprehend the information provided on the HMI and subsequently answer SA queries based on the traffic situation. In terms of the results, it can be debated whether the use of static images, into which the participants were “dropped in”, is a representative way to approximate real-life collision avoidance situations. When on the ship’s bridge, navigators are constantly processing dynamic information on own ship’s manoeuvring in relation to other traffic, obstacles, and land. In this experiment, navigators were exposed to a collision situation that they had to make sense of rather than act on. Also, since the information was static, temporal information regarding target ships’ movements over time, e.g., when changing course or speed, was not possible. Consequently, despite the efforts to create realistic conflict situations, the fact they were static reduced the realism of the situations. As such, when interpreting the results for application in a real-world setting, this needs to be considered. Future work should focus on implementing transparency in real-time simulation facilities comprising dynamic traffic situations and (simulated) collision avoidance systems.

As discussed earlier, critics may argue that more participants would have provided better experimental results. That is, for statistical analysis, the study’s sample size and its distribution are important determinants of the robustness of the outcomes. In this experiment, the number of participants for the experiment was based on statistical requirements for data analysis. A requirement for ensuring robustness of results when using parametric statistical inference techniques, such Analysis of Variance (ANOVA), is a normally distributed sampling distribution (Vogt, 2005). Sampling distributions tend to approach a normal distribution from sample sizes of 30 and more according to the Central Limit Theorem (Tabachnick et al., 2019; Vogt, 2005). To ensure this, efforts were made to maximise recruitment of licensed navigators. As professionals can be challenging to obtain for academic research, it was aimed to lower the threshold for recruitment as much as possible. Therefore, the doctoral researcher travelled to locations most suitable for the participants to perform the study, including onboard passenger ferries and at various national nautical training institutes such that interference between their participation and their professional activities was minimized (see Figure 14). This way, a sufficient sample size of 34 navigators was obtained and robustness of statistical results was ensured. Furthermore, a range of strategies were used to ensure quality of the experimental data, including pilot testing of the experimental setup and procedures, the choice and quality of the traffic situations, integration of transparency information in the primary task display, the choice of established and validated measurement techniques, performed quality control measures prior to data analysis, and applying established statistical methods for analysing the data. Finally, the experiment followed a strict procedure, established experimental software, and was predominantly automated such that the effect of the experimenter on the data was minimised. In all, these strategies contributed to the validity and reliability of the data, and the trustworthiness of the experimental results.



## 5.4 Recommendations for future work

This dissertation has assumed a supervisory role of an operator that monitors, controls, and potentially intervenes in a collision avoidance system. The effectiveness of this role, i.e., the challenges to supervisory control, was discussed earlier and was summarised as “ironies of automation” (Bainbridge, 1983), “OOTL performance problem” (Endsley & Kiris, 1995), and “automation conundrum” (Endsley, 2017). This dissertation investigated the role of transparency in supporting supervisory control through transparency. However, it is recognized and emphasised that transparency is only a part of the puzzle. Research on supervisory control of highly automated systems has uncovered there are many factors that affect human supervisory performance of which agent-human communication is only one. Individual-, task-, and system factors, comprising of skills, training, competence, complexity, level and type of automation, granularity of control, automation reliability, competing task demands, and distribution of roles and responsibilities between agent and human all contribute to successful automation oversight and interaction performance (Endsley, 2017). Considering these constraints and given the developments within the autonomous shipping domain, there is a continuous and urgent need to explore how teams of agents and humans can work together (National Academies of Sciences, Engineering and Medicine, 2022). Here, the focus should be on creating meaningful human work where the combined capabilities of humans and agents can be exploited. Although it may be assumed that such teams will be more effective than each of its constituents alone, future work needs to ensure that humans can understand and predict the behaviours of the agent, develop trust relationships, make accurate decisions based on input from the agent, and exert timely and appropriate control over the agent when needed (National Academies of Sciences, Engineering and Medicine, 2022). To this end, transparency has an important role to play in terms of supporting task-, agent-, and system SA (Endsley, 2023a).

The PSW model, as applied in this dissertation, is a simple, yet pragmatic representation of human information processing. In their original article, Parasuraman et al. (2000), commented on the limits of this model as a “theoretical structure of the human cognitive system”, but rather aimed to “propose a structure that is useful in practice” (Parasuraman et al., 2000, p. 288). Based on these considerations, this model was applied in this dissertation and proved to be a pragmatic framework for developing transparent HMIs. However, despite its simplicity, this representation does not cover the whole breadth of the contemporary knowledge that other transparency models have aimed to address (J. Y. C. Chen et al., 2014; Lyons, 2013). Most prominently, the lack of a representation of information uncertainty is a limiting factor in the current representation of the model. In AI-enabled systems, using stochastic models, uncertainty is an inherent characteristic (Hüllermeier & Waegeman, 2021). For example, uncertainty may occur in the system’s information acquisition stage due to missing data, reliability of sensors, noisy data, and incongruent data. It may occur in the system’s transformation of data because of interpolation, sampling, simplification, and errors in the algorithms. And finally, it may appear in the system’s output generation, due to approximations, mapping and classifications, and tolerances (Kunze et al., 2019). These uncertainties may affect the user’s trust in the system, SA, and decisions to take over control. However, as the PSW model, and its repurposed transparency version, does not include uncertainty, it was not part of the research focus of this dissertation. This means that the results, including the graphical depictions in the form of symbology, should be revisited in future research endeavours and include uncertainty elements. For example, a way forward would be to investigate how to integrate uncertainty information in the processing steps of the model (as eluded

above). Furthermore, representation of uncertainty information onto the various levels of transparency should be investigated further.

The results from the rankings indicate that participants preferred the medium- and high transparency levels compared to the low- and medium levels. The low transparency level, where only decisions and future actions were depicted was least preferred. There was no difference between the medium- and high transparency levels in terms of ranking, indicating that these transparency levels were equally preferred. This seems to imply that participants prefer more information rather than less information when it comes to making sense of an agent's decisions and actions, or at a minimum prefer access to the agent's underlying analysis processes. Nevertheless, there is no clear result pointing towards the optimal level of transparency across the dependent variables. This means that, when designing for transparency, it may be challenging to decide on which level of to implement. In Article five, a more demand-driven transparency approach was alluded to where users adjust the level of transparency depending on the task and context. This approach may be used to provide control to the supervisor over the amount of system information presented. A demonstration of such an approach was provided in a study by Vered et al. (2020) that found that the downsides of presenting transparency information may be avoided whilst maintaining its benefits. For example, when applied to autonomous shipping, supervisors may only depict a low level of transparency in situations with little to no traffic whilst "dialling up" the level of transparency for situations that require closer supervision. This way, future work should investigate the effectiveness of this approach in improving comprehension times compared to the sequential transparency approach as used in this study. In addition, future work should anticipate and mitigate the potential risks associated with a flexible approach to transparency by addressing information overload and the potential for confusion between information levels.

## 6 Conclusions

This chapter presents the main research findings and contributions from this dissertation. The research questions are revisited, the contributions of the research are discussed, and an outlook for recommended future work is provided.

### 6.1 Revisiting the research questions

As presented in the introduction section of this dissertation, the maritime industry is looking for ways to reduce its environmental footprint, there is an ongoing trend in the maritime industry to become more attractive to personnel, improve its safety record, and enhance its resilience against adverse conditions, whilst maintaining profitability. To this end, it is anticipated that advanced technologies, including AI-enabled systems, will play an important role in achieving these aims by enabling ships to sail without direct human involvement, introducing new designs, implementing novel propulsion technologies, and exploiting alternative means of operation. However, considering the safety-critical nature of the industry, these systems will need to demonstrate considerable reliability across a wide range of situations. Therefore, given the inherent limitations of such systems to handle novel and complex situations effectively and reliably, humans are foreseen to take a supervisory role to ensure performance and safety standards are met. However, there are well-known challenges related to assigning humans this function, as they are typically less involved in the system's information and decision-making loop. This potentially leads to biases in decision making, passive information processing, complacent behaviour, over- and underreliance on the system, and high workload when taking over manual control. Nevertheless, research has suggested that by disclosing the system's decisions, planned actions and internal reasoning to the supervisor, some of these challenges may be alleviated. However, considering the novelty of the application of AI-enabled systems in safety-critical domains, there is limited experience with the effect of transparency in these settings. Therefore, this dissertation purposed to explore new knowledge, methods, and tools on the role of transparency in supporting humans in supervisory control. Therefore, the overarching research question in this dissertation was as follows:

*How does agent transparency support human performance in supervisory control?*

This question was decomposed into five sub-questions:

*RQ1: What is the relationship between agent transparency and Situation Awareness, mental workload, and task performance?*

This dissertation started by performing a broad and systematic analysis of relevant scientific publications regarding agent transparency and human performance variables. Based on 17 scientific studies, this dissertation found a promising effect of transparency on SA and task performance, without affecting mental workload. The results were especially clear for studies where participants were responding to proposals or supervising automation. It was suggested that strategies to improve human performance when interacting with intelligent agents should focus on allowing humans to see into its information processing stages, considering the integration of information in existing HMI solutions.

*RQ2: How is human performance achieved in conventional- and supervised maritime collision avoidance?*

A GDTA mapped and analysed the goals, decisions, and cognitive tasks associated with conventional and supervised collision avoidance. This activity explored the shift in cognitive activities when the navigator's task changes from performing collision avoidance to supervising a system performing collision avoidance. It was suggested that to allow a supervisor to assess the decisions and planned actions of the system, the system needs to provide insight into its information processing. To support operators in this, explicit information requirements were identified that should allow for insight into the agent's decisions, planned actions, and underlying reasoning.

*RQ3: How does a model for human information processing form the basis for agent transparency in the ship autonomy context?*

An information processing model was adapted and repurposed to function as a model for transparency, describing the system's information processing steps as condition detection, condition analysis, action planning, and action control. The model was contextualized to the maritime collision avoidance setting such that the information from the GDTA could be structured into unique and distinct layers. This dissertation suggested that this model may serve as a means to structure the agent's information processing steps to create layers of transparency and to be used as a framework for transparent design.

*RQ4. How should a maritime collision avoidance system be made transparent to a human supervisor?*

Traffic situation situations and symbology were developed to operationalise transparency for a collision avoidance system. The symbology was integrated into the primary task display for collision avoidance and the information processing model was used to create distinct levels of transparency. This activity provided the groundwork for the empirical evaluation of transparency in a maritime collision avoidance context. In addition, the results demonstrated the value of the model as a design framework for creating levels of transparency for autonomous agents.

*RQ5. What is the relationship between agent transparency and Situation Awareness, mental workload, and task performance in maritime autonomous collision avoidance?*

Through a controlled experiment with licensed navigators, the effect of transparency on SA, mental workload, and task performance was evaluated. This dissertation demonstrated a promising effect of transparency on SA without affecting mental workload. However, the time to comprehend the situation increased with increased levels of transparency. These results indicate the benefits of applying transparency principles to autonomous collision avoidance systems, but that care should be taken in time-critical conditions where the added transparency information may affect timely decision making. Furthermore, considering the absence of the effect of transparency on mental workload, these results also indicate the value of applying a structured and systematic design process, as applied in this dissertation.

## 6.2 Contributions

This dissertation has provided several theoretical and practical contributions for transparency research and its applications (see Table 10).

In terms of theoretical contributions, this dissertation advanced the body of knowledge by systematically mapping and assessing transparency research (Article 1; van de Merwe, Mallam, & Nazir, 2024). Although earlier attempts had been made to provide overviews (Bhaskara et al., 2020; Rajabiyazdi & Jamieson, 2020), the systematic focus of Article 1, based on the PRISMA approach, reduced the review's potential for bias, and enhanced clarity, auditability, replicability, and transparency. Furthermore, this dissertation provided a detailed analysis of the change in information requirements from conventional to supervised collision avoidance (Article 2; van de Merwe, Mallam, Nazir, et al., 2024a). By mapping this information, based on a variety of quality sources, a deeper understanding was created of the potential future role of the operator, and the information needed to support this. In addition, this dissertation contributed by exploring the application of an adapted and repurposed model for information processing (Article 3; van de Merwe et al., 2023b). By using the well-known PSW model as a transparency model, a pragmatic approach to development of transparent design concepts was explored. Finally, this dissertation contributed by generating knowledge about the relationship between transparency, SA, mental workload, and task performance by performing a controlled experiment with experienced and licensed SMEs (Article 5; van de Merwe, Mallam, Nazir, et al., 2024b). The results of the experiment highlighted the effects of transparency on these variables and expanded the knowledge regarding transparency as a design principle for agents applied in safety critical domains.

Practically, this dissertation contributed with providing evidence for transparency as a design principle for supporting supervision of autonomous agents (Article 1; van de Merwe, Mallam, & Nazir, 2024; Article 5; van de Merwe, Mallam, Nazir, et al., 2024b). Based on the evidence provided in these articles, developers have an incentive to create transparent designs for their users knowing that, by following a set of design principles and processes, their efforts will have an effect in terms of human performance. Furthermore, this dissertation has made explicit the potential role-change that may be anticipated when introducing collision avoidance systems (Article 2; van de Merwe, Mallam, Nazir, et al., 2024a). Based on this information, ship owners may better understand what may be expected, in terms of change in cognitive activities, when introducing autonomous collision avoidance systems. In addition, the repurposed information processing model provides a pragmatic framework for developing transparent agents (Article 3; van de Merwe et al., 2023b). Although the model may be an oversimplification of human information processing, for the purpose of structuring the information processing of an agent, this conceptualisation is as simple as it is useful. Moreover, this dissertation developed a set of realistic traffic situations and symbology, that may be useful for re-use in future research and testing activities (Article 4; van de Merwe et al., 2023a). Finally, the layers and levels of transparency that were created to represent the agent's internal information processing provide developers with concrete examples of how collision avoidance systems can be made transparent to their users (Article 4; van de Merwe et al., 2023a).

Table 10. Contributions of this dissertation.

RQ	Article title	Key points	Contributions
Main	N/A	<p>Planned, performed, and reported on applied research on a relevant and timely topic</p> <p>Published in relevant-, and recognised journals and conferences</p>	<p>Theoretical contributions:</p> <ul style="list-style-type: none"> <li>The publications and dissertation contribute to enhancing and progressing the knowledge, methods, and tools regarding agent transparency and human performance</li> <li>The dissertation provides clear directions for future research</li> </ul> <p>Practical contributions:</p> <ul style="list-style-type: none"> <li>The results show that transparency is a viable design principle for providing insight into an agent’s information processing, decisions, and planned actions</li> </ul>
1	<p>Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review</p> <p>Published in: Human Factors (2024)</p>	<p>Systematically gathered and assessed empirical evidence for the relationship between agent transparency and key human factors variables</p> <p>Described domains, models, operationalisations, and human-automation interaction types</p> <p>Found a promising effect of transparency on SA and task performance without the cost of added mental workload</p>	<p>Theoretical contributions:</p> <ul style="list-style-type: none"> <li>The results contribute to the knowledge regarding transparency as a design principle for effective human-automation interaction</li> </ul> <p>Practical contributions:</p> <ul style="list-style-type: none"> <li>The results provide incentives to designers for applying transparency principles, especially for when humans respond to proposals and perform supervisory control</li> </ul>
2	<p>Supporting human supervision in autonomous collision avoidance through agent transparency</p> <p>Published in: Safety Science (2024)</p>	<p>Goals, decisions, and cognitive tasks were identified for conventional- and supervised collision avoidance</p> <p>SA requirements for agent transparency were defined using a GDTA</p>	<p>Theoretical contributions:</p> <ul style="list-style-type: none"> <li>The results provide a detailed analysis of the change in information requirements from conventional to supervised collision avoidance</li> </ul> <p>Practical contributions:</p> <ul style="list-style-type: none"> <li>The results provide concrete insights into the SA requirements for supervised collision avoidance</li> </ul>

RQ	Article title	Key points	Contributions
3	Towards an approach to define transparency requirements for maritime collision avoidance  Published in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting (2023)	Argued for using the PSW model as a transparency model  Adapted the PSW model to the maritime collision avoidance domain  Used the model to organise the SA requirements from Article 2 into layers of transparency	Theoretical contributions: <ul style="list-style-type: none"> <li>The results expand the applicability of the PSW model to represent a model for agent transparency</li> </ul> Practical contributions: <ul style="list-style-type: none"> <li>The results provide a set of minimum SA requirements, organised per layer of transparency</li> </ul>
4	Operationalising Automation Transparency for Maritime Collision Avoidance  Published in: TransNav, International Journal on Marine Navigation and Safety of Sea Transportation (2023)	Developed 70 realistic traffic situations to support empirical evaluation  Applied transparency model and SA requirements to develop realistic HMIs	Theoretical contributions: <ul style="list-style-type: none"> <li>The results provide the groundwork for empirical evaluations of transparency layers</li> </ul> Practical contributions: <ul style="list-style-type: none"> <li>The results provide insight into the practical value of the model as a design framework for transparent agents</li> </ul>
5	The Influence of Agent Transparency and Complexity on Situation Awareness, Mental Workload, and Task Performance  Published in: Journal of Cognitive Engineering and Decision Making (2024)	Experimentally evaluated the transparency model in an autonomous collision avoidance context  Found effects of transparency on SA and task performance, but not on mental workload  Participants preferred HMIs where analytical information was depicted	Theoretical contributions: <ul style="list-style-type: none"> <li>The results add to the knowledge of the effects of transparency on key human factors variables</li> <li>The results empirically evaluate the proposed transparency model</li> </ul> Practical contributions: <ul style="list-style-type: none"> <li>The results provide insight into the anticipated human performance effects of transparency when applied to autonomous agents</li> </ul>

### 6.3 Concluding remarks

The rapid development and deployment of advanced technologies across society will affect the way work is organised and has the potential to significantly alter the role of the human herein. In the maritime domain, AI-enabled systems may perform tasks previously performed by navigators, allowing for ships to sail autonomously and without direct human involvement. However, considering the current limitations of such systems to manage novel and complex situations, human operators are foreseen to play a critical role in overseeing their functioning and ensuring they perform according to requirements. Therefore, the operator's ability to understand, predict, and evaluate system behaviour

becomes a critical aspect of the human's supervisory task repertoire and lays the foundation for effective human-agent teams.

This dissertation investigated the role of agent transparency in supporting operators in this new role and contributed with knowledge, methods, and tools regarding transparency in general and its application to the maritime domain specifically. The aim of this dissertation was to generate and advance the knowledge on how supervisory control can be supported through agent transparency. This dissertation has contributed to this aim by recognising the importance of transparency in safety critical domains in terms of human performance, exploring and understanding the impact of autonomy on the operator's cognitive tasks, constructing and contextualising a model for transparency, operationalising transparency for the maritime context, and assessing its effects in an experimental setting. The results have implications for scientific research and for the application of transparency as a design principle for autonomous agents. In addition, this dissertation has made explicit the role-change that may be anticipated when introducing autonomous systems. In understanding the cognitive underpinnings, the results from this dissertation may be expanded towards competence and learning programs addressing supervisory control, the deployment of AI-enabled systems in operational settings, and the exploration of effective human-autonomy collaboration strategies. With these new insights, meaningful human work may be created where the combined capabilities of human-agent teams can be optimised. Ultimately, this dissertation advocates the relevance of affording human operators with insight into the reasoning of autonomous systems and established transparency as an important prerequisite on the path towards safe and effective human-supervisory control.



## 7 References

- Akdağ, M., Solnør, P., & Johansen, T. A. (2022). Collaborative collision avoidance for Maritime Autonomous Surface Ships: A review. *Ocean Engineering*, 250, 110920. <https://doi.org/10.1016/j.oceaneng.2022.110920>
- Allianz. (2023). *Safety and Shipping Review 2023* (p. 44). Allianz Global Corporate & Specialty. <https://commercial.allianz.com/content/dam/onemarketing/commercial/commercial/reports/A-GCS-Safety-Shipping-Review-2023.pdf>
- Allianz. (2024). *Allianz Risk Barometer* (p. 51). Allianz Commercial. <https://commercial.allianz.com/news-and-insights/reports/allianz-risk-barometer.html>
- Alsos, O., Veitch, E., Pantelatos, L., Vasstein, K., Eide, E., Petermann, F.-M., & Breivik, M. (2022). NTNU Shore Control Lab: Designing shore control centres in the age of autonomous ships. *Journal of Physics: Conference Series*, 2311, 012030. <https://doi.org/10.1088/1742-6596/2311/1/012030>
- American Bureau of Shipping. (2022). *Requirements for Autonomous and Remote Control Functions*. American Bureau of Shipping. [https://ww2.eagle.org/content/dam/eagle/rules-and-guides/current/other/323\\_gn\\_autonomous\\_2022/323-autonomous-reqts-aug22.pdf](https://ww2.eagle.org/content/dam/eagle/rules-and-guides/current/other/323_gn_autonomous_2022/323-autonomous-reqts-aug22.pdf)
- Anguera, M. T., Blanco-Villaseñor, A., Losada, J. L., Sánchez-Algarra, P., & Onwuegbuzie, A. J. (2018). Revisiting the difference between mixed methods and multimethods: Is it all in the name? *Quality & Quantity*, 52(6), 2757–2770. <https://doi.org/10.1007/s11135-018-0700-2>
- Aylward, K., Weber, R., Lundh, M., MacKinnon, S. N., & Dahlman, J. (2022). Navigators' views of a collision avoidance decision support system for maritime navigation. *Journal of Navigation*, 1–14. <https://doi.org/10.1017/S0373463322000510>
- Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6), 775–779. [https://doi.org/10.1016/0005-1098\(83\)90046-8](https://doi.org/10.1016/0005-1098(83)90046-8)
- Beck, H. P., Dzindolet, M. T., & Pierce, L. G. (2007). Automation Usage Decisions: Controlling Intent and Appraisal Errors in a Target Detection Task. *Human Factors*, 49(3), 429–437. <https://doi.org/10.1518/001872007X200076>
- Bhaskara, A., Duong, L., Brooks, J., Li, R., McInerney, R., Skinner, M., Pongracic, H., & Loft, S. (2021). Effect of automation transparency in the management of multiple unmanned vehicles. *Applied Ergonomics*, 90. <https://doi.org/10.1016/j.apergo.2020.103243>
- Bhaskara, A., Skinner, M., & Loft, S. (2020). Agent Transparency: A Review of Current Theory and Evidence. *IEEE Transactions on Human-Machine Systems*, 50(3), 215–224. IEEE Transactions on Human-Machine Systems. <https://doi.org/10.1109/THMS.2020.2965529>
- BIMCO. (2021, July 28). *New BIMCO/ICS seafarer shortage workforce report warns of serious potential officer shortage*. <https://www.bimco.org/news/priority-news/20210728---bimco-ics-seafarer-workforce-report>
- Booth, A., Sutton, A., & Papaioannou, D. (2016). *Systematic approaches to a successful literature review* (Second edition). Sage.
- Boruch, R. (1997). *Randomized Experiments for Planning and Evaluation*. SAGE Publications, Inc. <https://doi.org/10.4135/9781412985574>
- Bureau Veritas. (2019). *Guidelines for autonomous shipping* (Guidance Note NI 641 DT R01 E; pp. 1–38). Bureau Veritas. [https://erules.veristar.com/dy/data/bv/pdf/641-NI\\_2019-10.pdf](https://erules.veristar.com/dy/data/bv/pdf/641-NI_2019-10.pdf)
- Chen, J. Y. C., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. J. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, 19(3), 259–282. <https://doi.org/10.1080/1463922X.2017.1315750>
- Chen, J. Y. C., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. J. (2014). *Situation Awareness-Based Agent Transparency* (ARL-TR-6905). U.S. Army Research Laboratory. <https://doi.org/10.21236/ADA600351>
- Chen, T., Campbell, D. A., Gonzalez, F., & Coppin, G. (2014, December). The effect of autonomy transparency in human-robot interactions: A preliminary study on operator cognitive workload and situation awareness in multiple heterogeneous UAV management. *Proceedings of*

- Australasian Conference on Robotics and Automation 2014*.  
<https://www.araa.asn.au/acra/acra2014/papers/pap166.pdf>
- Chen, T., Campbell, D. A., Gonzalez, L. F., & Coppin, G. (2015). Increasing Autonomy Transparency through capability communication in multiple heterogeneous UAV management. *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2434–2439. <https://doi.org/10.1109/IROS.2015.7353707>
- Christoffersen, K., & Woods, D. D. (2002). 1. How to make automated systems team players. In *Advances in Human Performance and Cognitive Engineering Research* (Vol. 2, pp. 1–12). Emerald Group Publishing Limited. [https://doi.org/10.1016/S1479-3601\(02\)02003-9](https://doi.org/10.1016/S1479-3601(02)02003-9)
- Coleman, R. (2019). *Designing Experiments for the Social Sciences: How to Plan, Create, and Execute Research Using Experiments*. SAGE Publications, Inc. <https://doi.org/10.4135/9781071878958>
- Cowan, N. (2010). The Magical Mystery Four: How Is Working Memory Capacity Limited, and Why? *Current Directions in Psychological Science*, *19*(1), 51–57. <https://doi.org/10.1177/0963721409359277>
- Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319–340. <https://doi.org/10.2307/249008>
- de Vos, J., Hekkenberg, R. G., & Valdez Banda, O. A. (2021). The Impact of Autonomous Ships on Safety at Sea – A Statistical Analysis. *Reliability Engineering & System Safety*, *210*, 107558. <https://doi.org/10.1016/j.ress.2021.107558>
- DNV. (2021). *Autonomous and remotely operated ships* (DNV-CG-0264). DNV. <https://www.dnv.com/maritime/autonomous-remotely-operated-ships/class-guideline>
- DNV. (2023). *Assurance of AI-enabled systems* (DNV-RP-0671). <https://www.dnv.com/digital-trust/recommended-practices/assurance-of-ai-enabled-systems-dnv-rp-0671/>
- Du, N., Haspiel, J., Zhang, Q., Tilbury, D., Pradhan, A. K., Yang, X. J., & Robert, L. P. (2019). Look who’s talking now: Implications of AV’s explanations on driver’s trust, AV preference, anxiety and mental workload. *Transportation Research Part C: Emerging Technologies*, *104*, 428–442. <https://doi.org/10.1016/j.trc.2019.05.025>
- Eccles, D. W., & Aarsal, G. (2017). The think aloud method: What is it and how do I use it? *Qualitative Research in Sport, Exercise and Health*, *9*(4), 514–531. <https://doi.org/10.1080/2159676X.2017.1331501>
- Endsley, M. R. (1988). Design and Evaluation for Situation Awareness Enhancement. *Proceedings of the Human Factors Society Annual Meeting*, *32*(2), 97–101. <https://doi.org/10.1177/154193128803200221>
- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 217–249. <https://doi.org/10.4324/9781315092898-13>
- Endsley, M. R. (2000). Direct Measurement of Situation Awareness: Validity and Use of SAGAT. In E. Salas (Ed.), *Situational Awareness Analysis and Measurement* (1st ed., pp. 147–174). CRC Press. <https://doi.org/10.4324/9781315087924-9>
- Endsley, M. R. (2017). From Here to Autonomy: Lessons Learned from Human-Automation Research. *Human Factors*, *59*(1), 5–27. <https://doi.org/10.1177/0018720816681350>
- Endsley, M. R. (2021). A Systematic Review and Meta-Analysis of Direct Objective Measures of Situation Awareness: A Comparison of SAGAT and SPAM. *Human Factors*, *63*(1), 124–150. <https://doi.org/10.1177/0018720819875376>
- Endsley, M. R. (2023a). Ironies of artificial intelligence. *Ergonomics*, *0*(0), 1–13. <https://doi.org/10.1080/00140139.2023.2243404>
- Endsley, M. R. (2023b). Supporting Human-AI Teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior*, *140*, 107574. <https://doi.org/10.1016/j.chb.2022.107574>
- Endsley, M. R., Bolté, B., & Jones, D. G. (2003). *Designing for situation awareness: An approach to user-centered design*. Taylor & Francis.
- Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors*, *37*(2), 381–394. <https://doi.org/10.1518/001872095779064555>

- Endsley, M. R., Selcon, S. J., Hardiman, T. D., & Croft, D. G. (1998). Comparative analysis of SAGAT and SART for evaluations of situation awareness. *Proceedings of the Human Factors and Ergonomics Society, 1*, 82–86.
- Finomore, V. S., Shaw, T. H., Warm, J. S., Matthews, G., & Boles, D. B. (2013). Viewing the workload of vigilance through the lenses of the NASA-TLX and the MRQ. *Human Factors, 55*(6), 1044–1063. <https://doi.org/10.1177/0018720813484498>
- Fortune Business Insights. (2021, July). *The global marine vessel market is projected to grow from USD 170.75 billion in 2021 to USD 188.57 billion in 2028 at a CAGR of 1.43% in forecast period... Read More at:-* <https://www.fortunebusinessinsights.com/marine-vessel-market-102699>. <https://www.fortunebusinessinsights.com/marine-vessel-market-102699>
- Galy, E., Cariou, M., & Mélan, C. (2012). What is the relationship between mental workload factors and cognitive load types? *International Journal of Psychophysiology, 83*(3), 269–275. <https://doi.org/10.1016/j.ijpsycho.2011.09.023>
- Gawron, V. J. (2019a). *Human performance and situation awareness measures* (Third edition). CRC Press/Taylor & Francis Group.
- Gawron, V. J. (2019b). *Workload measures* (Third edition). CRC Press/Taylor & Francis Group.
- Göriztlehner, R., Borst, C., Ellerbroek, J., Westin, C., van Paassen, M. M., & Mulder, M. (2014). Effects of transparency on the acceptance of automated resolution advisories. *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2965–2970. <https://doi.org/10.1109/SMC.2014.6974381>
- Guznov, S., Lyons, J. B., Pfahler, M., Heironimus, A., Woolley, M., Friedman, J., & Neimeier, A. (2020). Robot Transparency and Team Orientation Effects on Human–Robot Teaming. *International Journal of Human–Computer Interaction, 36*(7), 650–660. <https://doi.org/10.1080/10447318.2019.1676519>
- Haberlandt, K. (1999). *Human memory: Exploration and application*. Allyn and Bacon.
- Hancock, G. M., Longo, L., Young, M. S., & Hancock, P. A. (2021). Mental workload. In G. Salvendy & W. Karwowski (Eds.), *Handbook of Human Factors and Ergonomics* (5th ed., pp. 203–226). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781119636113.ch7>
- Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 50*(9), 904–908. <https://doi.org/10.1177/154193120605000909>
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in Psychology* (Vol. 52, pp. 139–183). North-Holland. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Hollnagel, E. (2012). Coping with complexity: Past, present and future. *Cognition, Technology & Work, 14*(3), 199–205. <https://doi.org/10.1007/s10111-011-0202-7>
- Howell, K. E. (2013). *An introduction to the philosophy of methodology*. Sage.
- Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning, 110*(3), 457–506. <https://doi.org/10.1007/s10994-021-05946-3>
- Human Factors and Ergonomics Society. (2023). *What is Human Factors and Ergonomics?* <https://www.hfes.org/About-HFES/What-is-Human-Factors-and-Ergonomics>
- IEC. (2019). *IEC 60964:2019 Nuclear power plants control rooms design*. International Electrotechnical Commission.
- IEC. (2022). *IEC 62288:2022 Maritime navigation and radiocommunication equipment and systems (62288)*. International Electrotechnical Commission.
- IMO. (1974). *International Convention for the Safety of Life at Sea*. <https://www.imo.org/en/KnowledgeCentre/ConferencesMeetings/Pages/SOLAS.aspx>
- IMO. (1977). *Convention of the international regulations for preventing collisions at sea (COLREGS)*. International Maritime Organisation.
- IMO. (1978). *International Convention on Standards of Training, Certification and Watchkeeping for Seafarers*. <https://www.imo.org/en/OurWork/HumanElement/Pages/STCW-Conv-LINK.aspx>

- IMO. (1979). *Resolution A.422(XI) Performance standards for automatic radar plotting aids (ARPA)*. [https://www.wcdn.imo.org/localresources/en/KnowledgeCentre/IndexofIMOResolutions/AssemblyDocuments/A.422\(11\).pdf](https://www.wcdn.imo.org/localresources/en/KnowledgeCentre/IndexofIMOResolutions/AssemblyDocuments/A.422(11).pdf)
- IMO. (2013). *Guidelines for the approval of alternatives and equivalents as provided for in various IMO instruments (MSC.1/Circ.1455)*. International Maritime Organisation.
- IMO. (2019). *Frequently Asked Questions*. <https://www.imo.org/en/About/Pages/FAQs.aspx>
- IMO. (2021, May 25). *Autonomous ships: Regulatory scoping exercise completed*. <https://www.imo.org/en/MediaCentre/PressBriefings/pages/MASSRSE2021.aspx>
- IMO. (2022). *Maritime Safety Committee (MSC 105), 20-29 April 2022 (MSC.1/Circ.1638)*. International Maritime Organisation. <https://www.imo.org/en/MediaCentre/MeetingSummaries/Pages/MSC-105th-session.aspx>
- IMO. (2023). *2023 IMO strategy on reduction of GHG emissions from ships (Resolution MEPC.377(80))*. International Maritime Organisation. <https://www.imo.org/en/OurWork/Environment/Pages/2023-IMO-Strategy-on-Reduction-of-GHG-Emissions-from-Ships.aspx?ref=marineregulations.news>
- International Committee of Medical Journal Editors. (2023). *Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals*. International Committee of Medical Journal Editors. <https://www.icmje.org/icmje-recommendations.pdf>
- International Ergonomics Association. (2000). *What Is Ergonomics (HFE)?* <https://iea.cc/about/what-is-ergonomics/>
- ISO. (1998). *ISO 9241-11:1998 Ergonomic requirements for office work with visual display terminals (VDTs)—Part 11: Guidance on usability*. International Organization for Standardization.
- ISO. (2000). *ISO 11064-2:2000 Ergonomic design of control centres—Part 2: Principles for the arrangement of control suites*. International Organization for Standardization.
- ISO. (2011). *ISO 26800:2011 Ergonomics—General approach, principles and concepts*. International Organization for Standardization.
- ISO. (2015). *ISO 9000: 2015 Quality management systems—Fundamentals and vocabulary*. International Organization for Standardization.
- ISO. (2017). *ISO/IEC 38505-1:2017 Information technology—Governance of IT - governance of data—Part 1: Application of ISO/IEC 38500 to the governance of data*. International Organization for Standardization.
- ISO. (2019). *ISO 9241-210:2019 Ergonomics of human-system interaction—Part 210: Human-centred design for interactive systems*. International Organization for Standardization.
- ISO. (2020a). *ISO 9241-110:2020 Ergonomics of human-system interaction—Part 110: Interaction principles*. International Organization for Standardization.
- ISO. (2020b). *ISO/IEC TR 29119-11:2020 Software and systems engineering—Software testing—Part 11: Guidelines on the testing of AI-based systems*. International Organization for Standardization.
- ISO. (2021). *ISO/IEC TR 24029-1:2021 Artificial Intelligence (AI)—Assessment of the robustness of neural networks—Part 1: Overview*. International Organization for Standardization.
- Jarvie, I. C., & Zamora Bonilla, J. (2011). *The SAGE handbook of the philosophy of social sciences*. SAGE.
- Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., van Riemsdijk, M. B., & Sierhuis, M. (2014). Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, 3(1), 43–69. <https://doi.org/10.5898/JHRI.3.1.Johnson>
- Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a Definition of Mixed Methods Research. *Journal of Mixed Methods Research*, 1(2), 112–133. <https://doi.org/10.1177/1558689806298224>
- Kahneman, D. (1973). *Attention and effort*. Prentice-Hall.
- Kirk, R. E. (2013). *Experimental Design: Procedures for the Behavioral Sciences*. SAGE Publications, Inc. <https://doi.org/10.4135/9781483384733>
- Kirwan, B., & Ainsworth, L. K. (Eds.). (1992). *A Guide to task analysis*. Taylor & Francis.
- Kretschmann, L., Burmeister, H. C., & Jahn, C. (2017). Analyzing the economic benefit of unmanned autonomous ships: An exploratory cost-comparison between an autonomous and a

- conventional bulk carrier. *Research in Transportation Business and Management*, 25, 76–86. <https://doi.org/10.1016/j.rtbm.2017.06.002>
- Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019). Automation transparency: Implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics*, 62(3), 345–360. <https://doi.org/10.1080/00140139.2018.1547842>
- Kurt, I., & Aymelek, M. (2022). Operational and economic advantages of autonomous ships and their perceived impacts on port operations. *Maritime Economics & Logistics*, 24(2), 302–326. <https://doi.org/10.1057/s41278-022-00213-1>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. [https://doi.org/10.1518/hfes.46.1.50\\_30392](https://doi.org/10.1518/hfes.46.1.50_30392)
- Lee, J. D., Wickens, C. D., Liu, Y., & Boyle, L. N. (2017). *Designing for people: An introduction to human factors engineering* (3rd edition, revision 1). CreateSpace.
- Littman, M., Ajunwa, I., Berger, G., Boutilier, C., Currie, M., Doshi velez, F., Hadfield, G., Horowitz, M., Isbell, C., Kitano, H., Levy, K., Lyons, T., Mitchell, M., Shah, J., Sloman, S., Vallor, S., & Walsh, T. (2021). *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report*. Stanford University. <http://ai100.stanford.edu/2021-report>
- Loft, S., Bhaskara, A., Lock, B. A., Skinner, M., Brooks, J., Li, R., & Bell, J. (2021). The Impact of Transparency and Decision Risk on Human–Automation Teaming Outcomes. *Human Factors*, 00187208211033445. <https://doi.org/10.1177/00187208211033445>
- Lyons, J. B. (2013). Being transparent about transparency: A model for human-robot interaction. *2013 AAAI Spring Symposium Series*.
- Malterud, K. (2001). Qualitative research: Standards, challenges, and guidelines. *The Lancet*, 358(9280), 483–488. [https://doi.org/10.1016/S0140-6736\(01\)05627-6](https://doi.org/10.1016/S0140-6736(01)05627-6)
- Malterud, K., Siersma, V., & Guassora, A. D. (2015). Sample Size in Qualitative Interview Studies: Guided by Information Power. *Qualitative Health Research*, 1. <https://doi.org/10.1177/1049732315617444>
- Meister, D. (1999). *The history of human factors and ergonomics*. Lawrence Erlbaum Associates.
- Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent Agent Transparency in Human–Agent Teaming for Multi-UxV Management. *Human Factors*, 58(3), 401–415. <https://doi.org/10.1177/0018720815621206>
- Metzger, U., & Parasuraman, R. (2001). The role of the air traffic controller in future air traffic management: An empirical study of active control versus passive monitoring. *Human Factors*, 43(4), 519–528. <https://doi.org/10.1518/001872001775870421>
- Miller, G. A. (1994). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 101(2), 343–352. <https://doi.org/10.1037/0033-295X.101.2.343>
- MITRE. (2018). *Human-Machine Teaming Systems Engineering Guide* (MP180941; p. 68). MITRE Corporation. <https://www.mitre.org/publications/technical-papers/human-machine-teaming-systems-engineering-guide>
- Moacdieh, N., & Sarter, N. (2017). The Effects of Data Density, Display Organization, and Stress on Search Performance: An Eye Tracking Study of Clutter. *IEEE Transactions on Human-Machine Systems*, 47(6), 886–895. *IEEE Transactions on Human-Machine Systems*. <https://doi.org/10.1109/THMS.2017.2717899>
- Mogford, R. H. (1997). Mental Models and Situation Awareness in Air Traffic Control. *The International Journal of Aviation Psychology*, 7(4), 331–341. [https://doi.org/10.1207/s15327108ijap0704\\_5](https://doi.org/10.1207/s15327108ijap0704_5)
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, T. P. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine*, 6(7), 6. <https://doi.org/10.1371/journal.pmed.1000097>
- Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., & PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, 4(1), 1. <https://doi.org/10.1186/2046-4053-4-1>

- Moon, M. D. (2019). Triangulation: A Method to Increase Validity, Reliability, and Legitimation in Clinical Research. *Journal of Emergency Nursing*, 45(1), 103–105. <https://doi.org/10.1016/j.jen.2018.11.004>
- Morse, J. M., Barrett, M., Mayan, M., Olson, K., & Spiers, J. (2002). Verification Strategies for Establishing Reliability and Validity in Qualitative Research. *International Journal of Qualitative Methods*, 1(2), 13–22. <https://doi.org/10.1177/160940690200100202>
- Mosier, K., & Skitka, L. (1996). Human Decision Makers and Automated Decision Aids: Made for Each Other? In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (Vol. 40, pp. 201–220). Lawrence Erlbaum Associates, Inc.
- Mulrow, C. D. (1994). Systematic Reviews: Rationale for systematic reviews. *BMJ*, 309(6954), 597–599. <https://doi.org/10.1136/bmj.309.6954.597>
- National Academies of Sciences, Engineering and Medicine. (2022). *Human-AI Teaming: State of the Art and Research Needs*. The National Academies Press. <https://doi.org/10.17226/26355>
- NMA. (2020). *Guidance in connection with the construction or installation of automated functionality aimed at performing unmanned or partially unmanned operations* (RSV 12-2020). Norwegian Maritime Authority.
- Norman, D. A. (1990). The “problem” with automation: Inappropriate feedback and interaction, not “over-automation.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 327(1241), 585–593. <https://doi.org/10.1098/rstb.1990.0101>
- Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 56(3), 476–488. <https://doi.org/10.1177/0018720813501549>
- Osofsky, S., Sanders, T., Jentsch, F., Hancock, P., & Chen, J. Y. C. (2014). Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In R. E. Karlson, D. W. Gage, C. M. Shoemaker, & G. R. Gerhart (Eds.), *Proceedings volume 9084: Unmanned Systems Technology XVI*. <https://doi.org/10.1117/12.2050622>
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 52(3), 381–410. <https://doi.org/10.1177/0018720810376055>
- Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39(2), 230–253. <https://doi.org/10.1518/001872097778543886>
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, 30(3), 286–297. <https://doi.org/10.1109/3468.844354>
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs. *Journal of Cognitive Engineering and Decision Making*, 2(2), 140–160. <https://doi.org/10.1518/155534308X284417>
- Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, 52(2), 183–199. <https://doi.org/10.1016/j.im.2014.08.008>
- Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect* (1st edition). Basic Books.
- Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Wiley-Blackwell.
- Pietrzykowski, Z., Wojejsza, P., & Borkowski, P. (2017). Decision Support in Collision Situations at Sea. *The Journal of Navigation*, 70(3), 447–464. <https://doi.org/10.1017/S0373463316000746>
- Plano Clark, V. L., & Ivankova, N. V. (2016). *Mixed Methods Research: A Guide to the Field*. SAGE Publications, Inc. <https://doi.org/10.4135/9781483398341>
- Pokam, R., Debernard, S., Chauvin, C., & Langlois, S. (2019). Principles of transparency for autonomous vehicles: First results of an experiment with an augmented reality human–machine interface. *Cognition, Technology & Work*, 21(4), 643–656. <https://doi.org/10.1007/s10111-019-00552-9>

- Porathe, T. (2019). Maritime Autonomous Surface Ships (MASS) and the COLREGS: Do We Need Quantified Rules Or Is “the Ordinary Practice of Seamen” Specific Enough? *TransNav, International Journal on Marine Navigation and Safety of Sea Transportation*, 13(3). <https://doi.org/10.12716/1001.13.03.04>
- Porathe, T., Fjortoft, K., & Bratbergsengen, I. L. (2020). Human Factors, autonomous ships and constrained coastal navigation. *IOP Conference Series: Materials Science and Engineering*, 929(1), 012007. <https://doi.org/10.1088/1757-899X/929/1/012007>
- Rajabiyazdi, F., & Jamieson, G. A. (2020). A Review of Transparency (seeing-into) Models. *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 302–308. <https://doi.org/10.1109/SMC42975.2020.9282970>
- Rao, A. S., & Georgeff, M. P. (1995). BDI Agents: From Theory to Practice. *Proceedings of the First International Conference on Multiagent Systems*, 312–319. <https://www.aaai.org/Papers/ICMAS/1995/ICMAS95-042.pdf>
- Ritter, F. E., Kim, J. W., Morgan, J. H., & Carlson, R. A. (2013). *Running Behavioral Studies with Human Participants: A Practical Guide*. SAGE Publications, Inc. <https://doi.org/10.4135/9781452270067>
- Rødseth, Ø. J., Lien Wennersberg, L. A., & Nordahl, H. (2021). Towards approval of autonomous ship systems by their operational envelope. *Journal of Marine Science and Technology*. <https://doi.org/10.1007/s00773-021-00815-z>
- Rolls Royce. (2016). *Rolls-Royce unveils a vision of the future of remote and autonomous shipping*. <https://www.rolls-royce.com/media/press-releases/2016/pr-12-04-2016-rr-unveils-a-vision-of-future-of-remote-and-autonomus-shipping.aspx>
- Roth, G., Schulte, A., Schmitt, F., & Brand, Y. (2020). Transparency for a Workload-Adaptive Cognitive Agent in a Manned–Unmanned Teaming Application. *IEEE Transactions on Human-Machine Systems*, 50(3), 225–233. <https://doi.org/10.1109/THMS.2019.2914667>
- Russell, S. J., & Norvig, P. (with Chang, M., Devlin, J., Dragan, A., Forsyth, D., Goodfellow, I., Malik, J., Mansinghka, V., Pearl, J., & Woolridge, M.). (2022). *Artificial intelligence: A modern approach* (Fourth edition, global edition). Pearson.
- Sadler, G. G., Battiste, H., Ho, N. T., Hoffmann, L., Johnson, W. B., Shively, R., Lyons, J. B., & Smith, D. (2016). Effects of transparency on pilot trust and agreement in the autonomous constrained flight planner. *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, 1–9. <https://doi.org/10.1109/DASC.2016.7777998>
- Saunders, M., Lewis, P., Thornhill, A., & Bristow, A. (2019). “Research Methods for Business Students” Chapter 4: Understanding research philosophy and approaches to theory development (pp. 128–171).
- Selkowitz, A. R., Lakhmani, S. G., & Chen, J. Y. C. (2017). Using agent transparency to support situation awareness of the Autonomous Squad Member. *Cognitive Systems Research*, 46, 13–25. <https://doi.org/10.1016/j.cogsys.2017.02.003>
- Selkowitz, A. R., Lakhmani, S. G., Chen, J. Y. C., & Boyce, M. (2015). The Effects of Agent Transparency on Human Interaction with an Autonomous Robotic Agent. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 59(1), 806–810. <https://doi.org/10.1177/1541931215591246>
- Sheridan, T. B. (1992). *Telerobotics, automation, and human supervisory control* (pp. xx, 393). The MIT Press.
- Sheridan, T. B. (2021). Human Supervisory Control of Automation. In G. Salvendy & W. Karwowski (Eds.), *Handbook of Human Factors and Ergonomics* (5th ed., pp. 736–760). Wiley. <https://doi.org/10.1002/9781119636113.ch28>
- Sheridan, T. B., & Verplank, W. L. (1978). *Human and Computer Control of Undersea Teleoperators*: Defense Technical Information Center. <https://doi.org/10.21236/ADA057655>
- Skraaning, G., & Jamieson, G. A. (2021). Human Performance Benefits of The Automation Transparency Design Principle: Validation and Variation. *Human Factors*, 63(3), 379–401. <https://doi.org/10.1177/0018720819887252>
- Stanton, N. A. (2006). Hierarchical task analysis: Developments, applications, and extensions. *Applied Ergonomics*, 37(1), 55–79. <https://doi.org/10.1016/j.apergo.2005.06.003>

- Stanton, N. A., Salmon, P. M., Rafferty, L. A., Walker, G. H., Baber, C., & Jenkins, D. P. (2013). *Human Factors Methods: A Practical Guide for Engineering and Design* (2nd ed.). CRC Press. <https://www.taylorfrancis.com/books/9781317120162>
- Stitt, I. P. A. (2002). The COLREGS – Time for a Rewrite? *The Journal of Navigation*, 55(3), 419–430. <https://doi.org/10.1017/S0373463302001893>
- Stowers, K., Kasdaglis, N., Rupp, M. A., Newton, O. B., Chen, J. Y. C., & Barnes, M. J. (2020). The IMPACT of Agent Transparency on Human Performance. *IEEE Transactions on Human-Machine Systems*, 50(3), 245–253. <https://doi.org/10.1109/THMS.2020.2978041>
- Styles, E. A. (1997). *The psychology of attention*. Psychology Press.
- Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2019). *Using multivariate statistics* (Seventh edition). Pearson.
- UNCTAD. (2023a). *Review of Maritime Transport 2023* (UNCTAD/RMT/2023). United Nations Conference on Trade and Development. [https://unctad.org/system/files/official-document/rmt2023\\_en.pdf](https://unctad.org/system/files/official-document/rmt2023_en.pdf)
- UNCTAD. (2023b). *Seafarer supply, quinquennial, 2015 and 2021* [Dataset]. <https://unctadstat.unctad.org/datacentre/dataviewer/US.Seafarers>
- van de Merwe, K., Mallam, S., Engelhardt, Ø., & Nazir, S. (2023a). Operationalising Automation Transparency for Maritime Collision Avoidance. *TransNav, International Journal on Marine Navigation and Safety of Sea Transportation*, 17(2). <https://doi.org/10.12716/1001.17.02.09>
- van de Merwe, K., Mallam, S., Engelhardt, Ø., & Nazir, S. (2023b). Towards an approach to define transparency requirements for maritime collision avoidance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 67(1), 483–488. <https://doi.org/10.1177/21695067231192862>
- van de Merwe, K., Mallam, S., & Nazir, S. (2024). Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review. *Human Factors*, 66(1), 180–208. <https://doi.org/10.1177/00187208221077804>
- van de Merwe, K., Mallam, S., Nazir, S., & Engelhardt, Ø. (2024a). Supporting human supervision in autonomous collision avoidance through agent transparency. *Safety Science*, 169, 13. <https://doi.org/10.1016/j.ssci.2023.106329>
- van de Merwe, K., Mallam, S., Nazir, S., & Engelhardt, Ø. (2024b). The Influence of Agent Transparency and Complexity on Situation Awareness, Mental Workload, and Task Performance. *Journal of Cognitive Engineering and Decision Making*, 18(2), 156–184. <https://doi.org/10.1177/15553434241240553>
- van de Merwe, K., Oprins, E., Eriksson, F., & van der Plaats, A. (2012). The Influence of Automation Support on Performance, Workload, and Situation Awareness of Air Traffic Controllers. *The International Journal of Aviation Psychology*, 22(2), 120–143. <https://doi.org/10.1080/10508414.2012.663241>
- van Doorn, E., Horváth, I., & Rusák, Z. (2021). Effects of coherent, integrated, and context-dependent adaptable user interfaces on operators' situation awareness, performance, and workload. *Cognition, Technology & Work*, 23(3), 403–418. <https://doi.org/10.1007/s10111-020-00642-z>
- van Doorn, E., Rusák, Z., & Horváth, I. (2017). A situation awareness analysis scheme to identify deficiencies of complex man-machine interactions. *International Journal of Information Technology and Management*, 16(1), 53–72. <https://doi.org/10.1504/IJITM.2017.080958>
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, 27(3), 425–478. <https://doi.org/10.2307/30036540>
- Vered, M., Howe, P., Miller, T., Sonenberg, L., & Velloso, E. (2020). Demand-Driven Transparency for Monitoring Intelligent Agents. *IEEE Transactions on Human-Machine Systems*, 50(3), 264–275. <https://doi.org/10.1109/THMS.2020.2988859>
- Vidulich, M. A. (2000). The Relationship between Mental Workload and Situation Awareness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 44(21), 3-460-3–463. <https://doi.org/10.1177/154193120004402122>



- Vidulich, M. A., & Tsang, P. S. (2014). The Confluence of Situation Awareness and Mental Workload for Adaptable Human–Machine Systems: *Journal of Cognitive Engineering and Decision Making*. <https://doi.org/10.1177/1555343414554805>
- Vogt, P. W. (2005). *Dictionary of Statistics & Methodology*. SAGE Publications, Inc. <https://doi.org/10.4135/9781412983907>
- Wang, H., Liu, Z., Wang, X., Graham, T., & Wang, J. (2021). An analysis of factors affecting the severity of marine accidents. *Reliability Engineering & System Safety*, 210, 107513. <https://doi.org/10.1016/j.ress.2021.107513>
- Warden, T., Carayon, P., Roth, E. M., Chen, J., Clancey, W. J., Hoffman, R., & Steinberg, M. L. (2019). The National Academies Board on Human System Integration (BOHSI) Panel: Explainable AI, System Transparency, and Human Machine Teaming. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63(1), 631–635. <https://doi.org/10.1177/1071181319631100>
- Warm, J. S., Dember, W. N., & Hancock, P. A. (1996). Vigilance and workload in automated systems. In *Automation and human performance: Theory and applications*. (pp. 183–200). Lawrence Erlbaum Associates, Inc.
- Weber, H. (1995). Clarification of the Steering and Sailing Rules of the COLREGS. *The Journal of Navigation*, 48(2), 289–292. <https://doi.org/10.1017/S0373463300012753>
- Wickens, C. D. (2008). Multiple Resources and Mental Workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(3), 449–455. <https://doi.org/10.1518/001872008X288394>
- Wickens, C. D., & Carswell, C. M. (2021). Information processing. In G. Salvendy & W. Karwowski (Eds.), *Handbook of Human Factors and Ergonomics* (5th ed., p. 1603). John Wiley & Sons, Incorporated.
- Wickens, C. D., Clegg, B. A., Vieane, A. Z., & Sebok, A. L. (2015). Complacency and Automation Bias in the Use of Imperfect Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 57(5), 728–739. <https://doi.org/10.1177/0018720815581940>
- Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2013). *Engineering psychology and human performance* (Fourth edition). Pearson.
- Wright, J. L., Chen, J. Y. C., & Lakhmani, S. G. (2020). Agent Transparency and Reliability in Human–Robot Interaction: The Influence on User Confidence and Perceived Reliability. *IEEE Transactions on Human-Machine Systems*, 50(3), 254–263. <https://doi.org/10.1109/THMS.2019.2925717>
- Wróbel, K., Gil, M., Huang, Y., & Wawruch, R. (2022). The Vagueness of COLREG versus Collision Avoidance Techniques—A Discussion on the Current State and Future Challenges Concerning the Operation of Autonomous Ships. *Sustainability*, 14(24), Article 24. <https://doi.org/10.3390/su142416516>
- Wróbel, K., Montewka, J., & Kujala, P. (2017). Towards the assessment of potential impact of unmanned vessels on maritime transportation safety. *Reliability Engineering and System Safety*, 165, 155–169. <https://doi.org/10.1016/j.ress.2017.03.029>
- Xiao, Y., & Watson, M. (2019). Guidance on Conducting a Systematic Literature Review. *Journal of Planning Education and Research*, 39(1), 93–112. <https://doi.org/10.1177/0739456X17723971>
- Zhang, X., Wang, C., Jiang, L., An, L., & Yang, R. (2021). Collision-avoidance navigation systems for Maritime Autonomous Surface Ships: A state of the art survey. *Ocean Engineering*, 235, 109380. <https://doi.org/10.1016/j.oceaneng.2021.109380>
- Zhou, X.-Y., Huang, J.-J., Wang, F.-W., Wu, Z.-L., & Liu, Z.-J. (2020). A Study of the Application Barriers to the Use of Autonomous Ships Posed by the Good Seamanship Requirement of COLREGs. *The Journal of Navigation*, 73(3), 710–725. <https://doi.org/10.1017/S0373463319000924>



## PART II



## Appendix A – Coupling the Goal-Directed Task Analysis, PSW model, and HMI

Table 11. The (condensed) results from the Goal-Directed Task Analysis (WP2), structured using the PSW model (WP3), and linked to HMI symbology (WP4).

COLREGs		GDTA			PSW model	CAGA system HMI
No.	Title	Goals	Decisions	Information requirements	Step	Symbology
2	Responsibility	To clarify that deviation from the rules may be necessary to avoid immediate danger	How can immediate danger be avoided?	Location of detected constraints Location of detected vessels Estimated collision risk Location of detected vessels The CAGA system's intended trajectory and speed The CAGA system's intended trajectory and speed	1. Information acquisition 2. Information analysis 2. Information analysis 2. Information analysis 3. Decision selection 3. Decision selection	Target identifier Target risk classifier Risk compass Target risk classifier Own ship future track Action table
5	Look-out	To determine the presence of vessels, terrain and other navigational constraints	How can a full appraisal of the situation be achieved?	Location of detected constraints/objects Location of detected vessels Estimated collision risk	1. Information acquisition 1. Information acquisition 1. Information acquisition	Target identifier Target identifier Target conflict type classifier
		To determine risk of collision	How can it be determined that a collision risk exists?	Estimated collision risk Estimated collision risk Estimated collision risk	2. Information analysis 2. Information analysis 2. Information analysis	Risk compass Target risk classifier Target information table
6a	Safe speed	To determine safe speed	How can safe speed be determined?	The parameters the chosen speed is based on, i.e., (i) Visibility estimates (ii) Observed targets and constraints (iii) Meteorological estimates (iii) Load and ballasting values (iv) Observed targets and constraints (v) Observed meteorological conditions (v) Radar echoes from navigational hazards (v) Charted navigational hazards (vi) Lowest water level contours (vi) Free space between vessel and bottom (vi) Vessel draught (based on load and ballasting)  The effect of the parameters on safe speed, i.e., (i) The effect of visibility on ability to detect other vessels (ii) The number of vessels in the vicinity, concentration and type	2. Information analysis	Safe speed table

## Agent Transparency and Human Performance in Supervisory Control

COLREGs		GDTA			PSW model	CAGA system HMI
No.	Title	Goals	Decisions	Information requirements	Step	Symbology
				(iii) The effect of meteorological conditions on manoeuvrability (iii) The effect of load and ballasting on manoeuvrability (iv) Origin of light sources (v) The ability of vessel to stay on its course (vi) Vessel listing (one side of vessel lies deeper than other side) (vi) The effects of squat in relation to actual speed		
6b	Safe speed	To take into account limitations or radar equipment in determining safe speed	What is the effect of radar limitations on determining safe speed?	Limitations to the radar/ sensors Uncertainties in the radar/ sensor data	1. Information acquisition 1. Information acquisition	Sensor status table Sensor status table
7a	Risk of Collision	To determine collision risk	What is the collision risk?	Which vessels/objects are detected Which vessels/objects form a collision risk Status of relevant sensors that are used by the CAGA system Which vessels/objects form a collision risk Which vessels/objects form a collision risk Which vessels/objects form a collision risk	1. Information acquisition 1. Information acquisition 1. Information acquisition 2. Information analysis 2. Information analysis	Target identifier Target conflict type classifier Sensor status table Target risk classifier Target predicted track
7b	Risk of Collision	To obtain early warning of collision risk	How can collision risk be determined early?	Detected objects in the short to long range Plotting of targets Targets that form a potential collision risk Status of sensors Plotting of targets Targets that form a potential collision risk Plotting of targets Targets that form a potential collision risk Plotting of targets Targets that form a potential collision risk Plotting of targets Targets that form a potential collision risk Plotting of targets	1. Information acquisition 1. Information acquisition 1. Information acquisition 1. Information acquisition 2. Information analysis 2. Information analysis 2. Information analysis 2. Information analysis 2. Information analysis 2. Information analysis 2. Information analysis 2. Information analysis	Target identifier Target conflict type classifier Target conflict type classifier Sensor status table Target risk classifier Target risk classifier Target predicted track Target predicted track Target risk classifier Target risk classifier Target information table
7c	Risk of Collision	To obtain adequate level of information for risk estimation	How can sufficient and reliable information be obtained?	The number of sensors detecting targets The reliability of the sensors	1. Information acquisition 1. Information acquisition	Target information table Sensor status table
7d	Risk of Collision	To use compass bearings	How can collision risk be determined?	Target vessel type and size Changes to bearing of target vessel over time Changes to bearing of target vessel over time Target vessel type and size Target vessel type and size	1. Information acquisition 1. Information acquisition 2. Information analysis 2. Information analysis	Target type identifier Target conflict type classifier Target risk classifier Target information table Target type identifier

COLREGs		GDTA			PSW model	CAGA system HMI
No.	Title	Goals	Decisions	Information requirements	Step	Symbology
				Changes to bearing of target vessel over time	2. Information analysis	Target predicted track
				Changes to bearing of target vessel over time	2. Information analysis	Risk compass
				Changes to bearing of target vessel over time	2. Information analysis	Target risk classifier
				CPA sufficiently large to take into account size, tow and distance to target	2. Information analysis	CPA indicator
				CPA sufficiently large to take into account size, tow and distance to target	3. Decision selection	Own ship future track
8a	Action to avoid collision	To execute collision avoidance manoeuvres by taking positive, timely and actions with good seamanship, in accordance with the rules	What is a positive and timely collision avoidance manoeuvre with the observance of good seamanship?	What actions it plans to perform (course and/or speed changes) upon detection of collision risk When it will perform these planned actions What actions it plans to perform (course and/or speed changes) upon detection of collision risk When it will perform these planned actions	3. Decision selection 3. Decision selection 3. Decision selection 3. Decision selection	Own ship future track Own ship future track Action table Action table
8b	Action to avoid collision	To execute changes in course and/ or speed that are clear to the other vessel	Which ship manoeuvres are readily observable by another vessel?	The CAGA system's changes to current course and/or speed	3. Decision selection 3. Decision selection	Own ship future track Action table
8c	Action to avoid collision	To execute early and substantial changes to the course of the vessel	Can a conflict be resolved with course corrections alone?	The CAGA system's intended course The CAGA system's intended course	3. Decision selection 3. Decision selection	Own ship future track Action table
8d	Action to avoid collision	To pass target vessel at a safe distance	How can safe distance be determined?	CPA to target vessel during avoidance manoeuvre The vessel's intended trajectory	2. Information analysis 3. Decision selection	CPA indicator Own ship future track
8e	Action to avoid collision	To avoid a collision or allow for more time to assess the situation	Can a conflict be resolved with speed corrections alone?	The CAGA system's current speed The CAGA system's intention of engine reversing The CAGA system's current speed The CAGA system's intention of engine reversing The CAGA system's current speed	3. Decision selection 3. Decision selection 3. Decision selection 3. Decision selection 3. Decision selection	Own ship future track Own ship future track Action table Action table Safe speed table
8f	Action to avoid collision	Give-way vessel: To take early action to allow for safe passage of the stand-on vessel	How can vessel priority be determined, and collision avoided?	Vessel priorities (i.e., not to be impeded/ not to impede) Vessel priorities (i.e., not to be impeded/ not to impede) Vessel priorities (i.e., not to be impeded/ not to impede) Vessel priorities (i.e., not to be impeded/ not to impede) The CAGA system's intended trajectory and speed	2. Information analysis 2. Information analysis 3. Decision selection 3. Decision selection 3. Decision selection	Target information table Target manoeuvrability identifier Own ship action indicator Target ship action indicator Own ship future track

## Agent Transparency and Human Performance in Supervisory Control

COLREGs		GDTA			PSW model	CAGA system HMI
No.	Title	Goals	Decisions	Information requirements	Step	Symbology
		Stand-on vessel: To take appropriate action with consideration of the rules in case of collision risk	How can vessel priority be determined, and collision avoided?	The CAGA system's intended trajectory and speed	3. Decision selection	Action table
13a	Overtaking	To avoid collision risk in overtaking situations	How can overtaking be safely executed? How can overtaking be safely executed? How can overtaking be safely executed? How can overtaking be safely executed?	Identified target vessel as vessel to be overtaken  The CAGA system's intended trajectory (i.e., CPA during passing manoeuvre) and speed The CAGA system's intended trajectory (i.e., CPA during passing manoeuvre) and speed The CAGA system's intention to overtake	1. Information acquisition  2. Information analysis 3. Decision selection 3. Decision selection	Target conflict type classifier  CPA indicator Own ship future track Own ship future track
13b	Overtaking	To clarify if a vessel is overtaking another when coming up with another vessel	How can an overtaking situation be determined when coming up with another vessel?	Identified that own vessel is approaching target vessel (own vessel has higher speed than target vessel) Identified that own vessel is in sector of more than 22,5 degrees abaft beam Identified that own vessel is approaching target vessel (own vessel has higher speed than target vessel) Identified that own vessel is in sector of more than 22,5 degrees abaft beam	1. Information acquisition 1. Information acquisition 2. Information analysis 2. Information analysis	Target conflict type classifier Target conflict type classifier Target conflict type classifier Target conflict type classifier
13c	Overtaking	To clarify if a vessel is overtaking another when in doubt	How can overtaking situations be determined?	Identified target vessel as vessel to be overtaken Identified target vessel as vessel to be overtaken	1. Information acquisition 2. Information analysis	Target conflict type classifier Target conflict type classifier
13d	Overtaking	To avoid becoming a crossing vessel when overtaking	How can it be avoided to become a crossing vessel when overtaking?	The CAGA system's intended trajectory in relation to target vessel (CPA during passing manoeuvre)	2. Information analysis 3. Decision selection	CPA indicator Own ship future track
14a	Head-on situation	To avoid collision risk in head-on situations	How can head-on situations be safely resolved?	Identified target vessel as head-on vessel Identified target vessel as head-on vessel The CAGA system's intended collision avoidance trajectory and speed, i.e., that course change is to starboard The CAGA system's intended collision avoidance trajectory and speed, i.e., that course change is to starboard	1. Information acquisition 2. Information analysis 3. Decision selection 3. Decision selection	Target conflict type classifier Target conflict type classifier Own ship future track Action table
14b	Head-on situation	To define when two vessels are in a head-on collision situation	How can head-on situations be determined?	Identified target vessel as head-on vessel Target vessel's course Identified target vessel as head-on vessel	1. Information acquisition 2. Information analysis 2. Information analysis	Target conflict type classifier Target predicted track Target conflict type classifier



Appendix A – Coupling the Goal-Directed Task Analysis, PSW model, and HMI

COLREGs		GDTA			PSW model	CAGA system HMI
No.	Title	Goals	Decisions	Information requirements	Step	Symbology
14c	Head-on situation	To clarify if a vessel is in head-on collision situation when in doubt	How can head-on situations be determined?	Identified target vessel as head-on vessel	1. Information acquisition	Target conflict type classifier
				Identified target vessel as head-on vessel	2. Information analysis	Target conflict type classifier
15	Crossing situation	To avoid collision risk in crossing situations	How can priority be determined in crossing situations?  How can crossing situations be safely resolved?	Identified collision risk	1. Information acquisition	Target conflict type classifier
				Identified target vessel as crossing	1. Information acquisition	Target conflict type classifier
				Identified collision risk	2. Information analysis	Target risk classifier
				Identified collision risk	2. Information analysis	Risk compass
				Identified collision risk	2. Information analysis	Target information table
				Identified target vessel as crossing	2. Information analysis	Target conflict type classifier
				Assigned own vessel priority (i.e., give-way/stand-on)	2. Information analysis	Target information table
				Assigned own vessel priority (i.e., give-way/stand-on)	3. Decision selection	Own ship action indicator
				Assigned own vessel priority (i.e., give-way/stand-on)	3. Decision selection	Target ship action indicator
				Give-way: The CAGA system’s intended collision avoidance trajectory and speed, i.e., that course change avoids passing in front	3. Decision selection	Own ship future track
Stand-on: The CAGA system’s intended collision avoidance trajectory and speed, i.e., no change in course and speed	3. Decision selection	Own ship future track				
Stand-on: The CAGA system’s intended collision avoidance trajectory and speed, i.e., no change in course and speed	3. Decision selection	Action table				
16	Action by give-way vessel	To avoid collision risk as a vessel directed to keep out of the way	How can vessel priority be determined?  How can a vessel directed to keep out of the way keep clear?	Assigned own vessel priority (i.e., give-way/stand-on)	2. Information analysis	Target information table
				Assigned own vessel priority (i.e., give-way/stand-on)	3. Decision selection	Own ship action indicator
				Assigned own vessel priority (i.e., give-way/stand-on)	3. Decision selection	Target ship action indicator
				The CAGA system’s intended collision avoidance trajectory involves route changes	3. Decision selection	Own ship future track
				Identified target vessel at early stage	3. Decision selection	Own ship future track
				The CAGA system’s intended collision avoidance trajectory involves route changes	3. Decision selection	Action table
17a	Action by stand-on vessel	To avoid collision risk as a stand-on vessel	How can vessel priority be determined?  How can inappropriate action by the give-way vessel be determined?	Identified target vessel	1. Information acquisition	Target identifier
				Assigned own vessel priority (i.e., give-way/stand-on)	3. Decision selection	Own ship action indicator
				Assigned own vessel priority (i.e., give-way/stand-on)	3. Decision selection	Target ship action indicator
				The CAGA system’s intended trajectory and speed	3. Decision selection	Own ship future track
				The CAGA system’s intended trajectory and speed	3. Decision selection	Action table

## Agent Transparency and Human Performance in Supervisory Control

COLREGs		GDTA			PSW model	CAGA system HMI
No.	Title	Goals	Decisions	Information requirements	Step	Symbology
17b	Action by stand-on vessel	To avoid collision risk when the actions of the give-way vessel alone are not sufficient	How can insufficient actions by the give-way vessel be determined?	Target vessel course and speed The CAGA system's intended trajectory and speed Minimum critical CPA limit reached The CAGA system's intended trajectory and speed	2. Information analysis 3. Decision selection 3. Decision selection 3. Decision selection	Target predicted track Own ship future track Own ship future track Action table
17c	Action by stand-on vessel	To avoid taking actions to port in crossing situations	How can crossing situations be safely resolved?	The CAGA system's intended trajectory does not involve course changes to port when other vessel is on port side	3. Decision selection 3. Decision selection	Own ship future track Action table

## Appendix B – Guide to Human-Machine Interface and symbology

Figure 23 below shows a typical traffic situation used for this experiment, including additional overlaid information. Each information item is explained in the tables below.

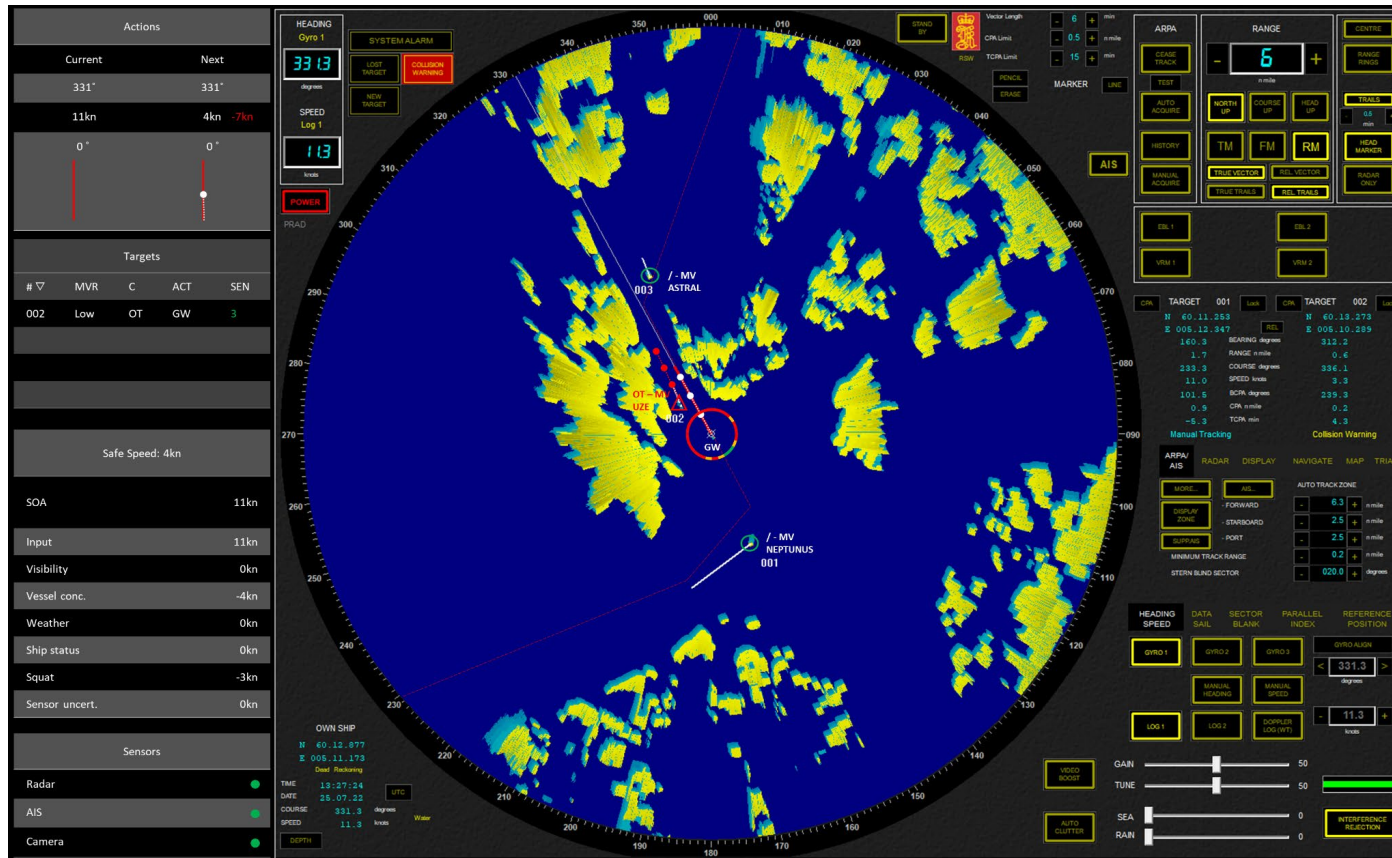



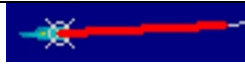

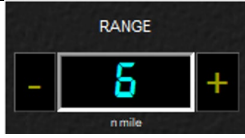
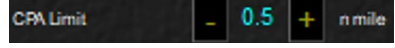
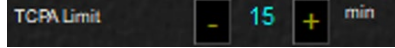

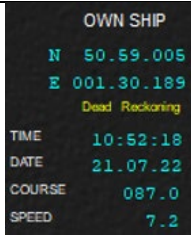


Figure 23. A typical traffic situation depicted on the radar screen.












Table 12. Common information elements depicted on the radar screen.

Item	Explanation	As shown on the interface
Radar echoes	Electromagnetic reflections detected by the radar	
Echo trails	Past residual image of radar echo	
Echo numbering	Target numbers assigned by the radar	
Own ship vector	6 mins for all situations, red	
Target ship Vectors	6 mins for all situations, white	
Radar range	6nm for all situations	
CPA limit	0.5nm is chosen as the CPA Limit for all situations	
TCPA limit	15 mins is chosen as the TCPA Limit for all situations	
Collision warning	Min. CPA exceeded for one or more targets, as plotted by ARPA	
Own ship	Only course, speed, and location are relevant for the situations	

Item	Explanation	As shown on the interface
Target details	Top two targets as assigned by ARPA	
Other	Other information elements part of ARPA are not used in the traffic situations.	

Table 13. Information specific to the collision avoidance system.

Item	Explanation	As shown on the interface
Own ship Current track	Current track for own ship The length of the red vector indicates the position and heading in 6 minutes The dotted red line is own ship's current route	
Own ship Future track	Future track by own ship depicted by up to three vectors in white. The angle of the vectors represents the course over ground The length of the vectors indicates its future position and heading per 6 minutes The dashes do not correlate with time (they are not to scale) and should not be used to determine precise future location	
Own ship Action indicator	GW: Own ship/ target ship will give-way SO: Own ship will stand-on	

Item	Explanation	As shown on the interface
Target ship action indicator	GW: Target ship will give-way Symbol is only shown when own ship is stand-on for a risk target	
Risk compass	A circle around the ship shows the risk associated with changing course for 15 degrees increments: Red: Risk of collision is high when changing to this direction Orange: Risk of collision is medium when changing to this direction Green: Risk of collision is medium when changing to this direction	
CPA indicator	Shown in red on own ship future track. CPA indicator indicates at what point own ship is closest to another ship. This may not be the original collision target.	
Target identifier	White circle: Target has been detected	
Target risk classifier	Green circle: Target does not pose a collision danger	
	Orange square: Target poses collision danger	
	Red triangle: Minimum CPA to be exceeded for this target	
Target conflict type classifier	OT: Overtaking Red: For targets violating CPA Orange: For other critical targets	
	CR: Crossing Red: For targets exceeding CPA Orange: For other critical targets	
	HO: Head-on Red: For targets exceeding CPA Orange: For other critical targets	
	/: Neither OT, CR, nor HO is applicable For example, a target is sailing away from own ship, or is a buoy	

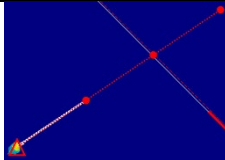
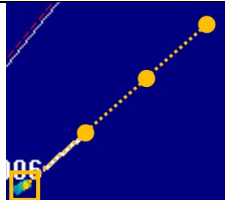

Item	Explanation	As shown on the interface
Target type identifier	Target name (example)	<b>CORALIUS</b>
	MV: Motor vessel	<b>MV</b>
	FV: Fishing vessel	<b>FV</b>
	Z: Stationary	<b>Z</b>
	BY: Buoy	<b>BY</b>
Target manoeuvrability identifier	RAM: Target vessel is restricted in its ability to manoeuvre	<b>RAM</b>
Target predicted track	Red: Predicted track, in 6 minutes increments, for target that is exceeding minimum CPA	
	Yellow = Predicted track, in 6 minutes increments, for target that is relevant during the evasion manoeuvre and may exceed minimum CPA	

Table 14. Information presented adjacent to the radar screen.

Tables	Explanation	As shown on the interface
Action	<p>A table showing own ship's current and next actions:                      Numerical representation of heading and speed                      6 minutes relative vectors indicating heading and speed</p> <p>Note that the vectors show actions relative to the current one.</p>	
Safe speed	<p>A table showing the chosen safe speed and the factors affecting it:                      SOA: Speed of Advance                      Input                      Visibility                      Vessel concentration                      Weather                      Ship status                      Squat                      Sensor uncertainty</p>	
Target information	<p>A table showing supplemental target information:                      Target number, prioritised                      Manoeuvrability: Low, Medium, High, RAM, FV                      Conflict type: OT, HO, CR                      Action of own ship against target if minimum CPA is exceeded: GW, SO                      Number of sensors detecting the target: 0-3</p>	



Tables	Explanation	As shown on the interface
Sensor status	<p>A table showing additional sensor status information for radar, AIS, and camera:</p> <p>Green: Status is OK</p> <p>Yellow: Status is degraded</p> <p>Red = Sensor is offline</p>	 <p>The screenshot shows a dark-themed interface titled 'Sensors'. It contains three rows of sensor status information:</p> <ul style="list-style-type: none"> <li><b>Radar:</b> A green dot indicates the status is OK.</li> <li><b>AIS:</b> A green dot indicates the status is OK.</li> <li><b>Camera:</b> A yellow dot indicates the status is degraded.</li> </ul>



## Appendix C – Examples of transparency levels

In Figure 24, based on the structured SA requirements from Table 7, own ship indicates its intended avoidance manoeuvre by drawing its planned track for the next three manoeuvring steps (each step corresponds to one vector length and equals six minutes). The system also depicts “GW” next to the own ship symbol which indicates it intends to give-way. This way, minimum transparency is provided to allow supervisors to understand that the system is about to initiate a 12-degree starboard turn and that it intends to give-way.

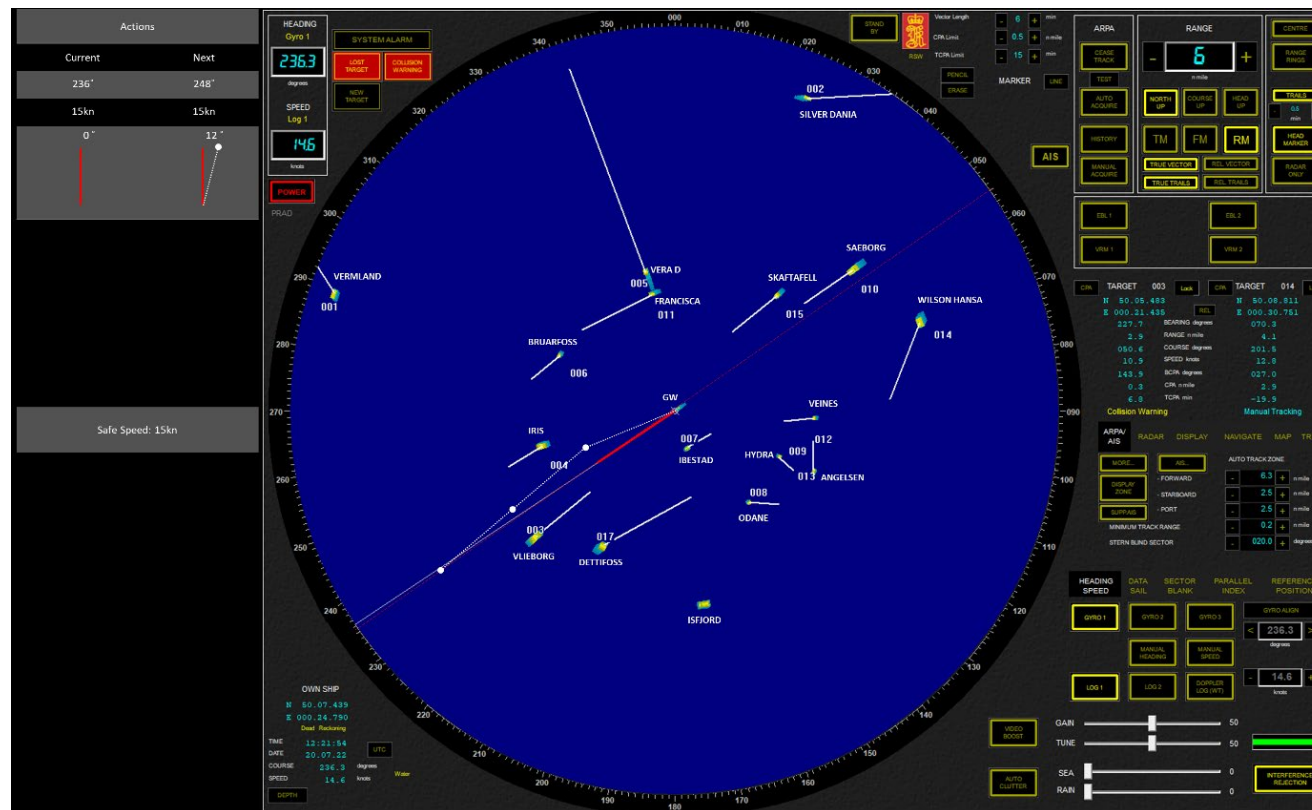


Figure 24. Traffic situation with a low level of transparency.

Figure 25 depicts that own ship considers two targets as especially relevant in this traffic situation. The target ship in red is depicted as the highest risk as for this ship, the minimum CPA limit is exceeded and is thus on collision course. The target in orange is also highlighted as own ship has considered this target to be of importance during the avoidance manoeuvre. Additional indicators next to the target symbols add information regarding the type of conflict and type of vessel, i.e., HO for head-on and MV for motor vessel. A manoeuvrability indicator is provided to indicate where own ship can manoeuvre within a one vector length. Finally, the factors influencing safe speed information is provided in table form on the left of the screen. Speed information can also be derived from the length of the vector.

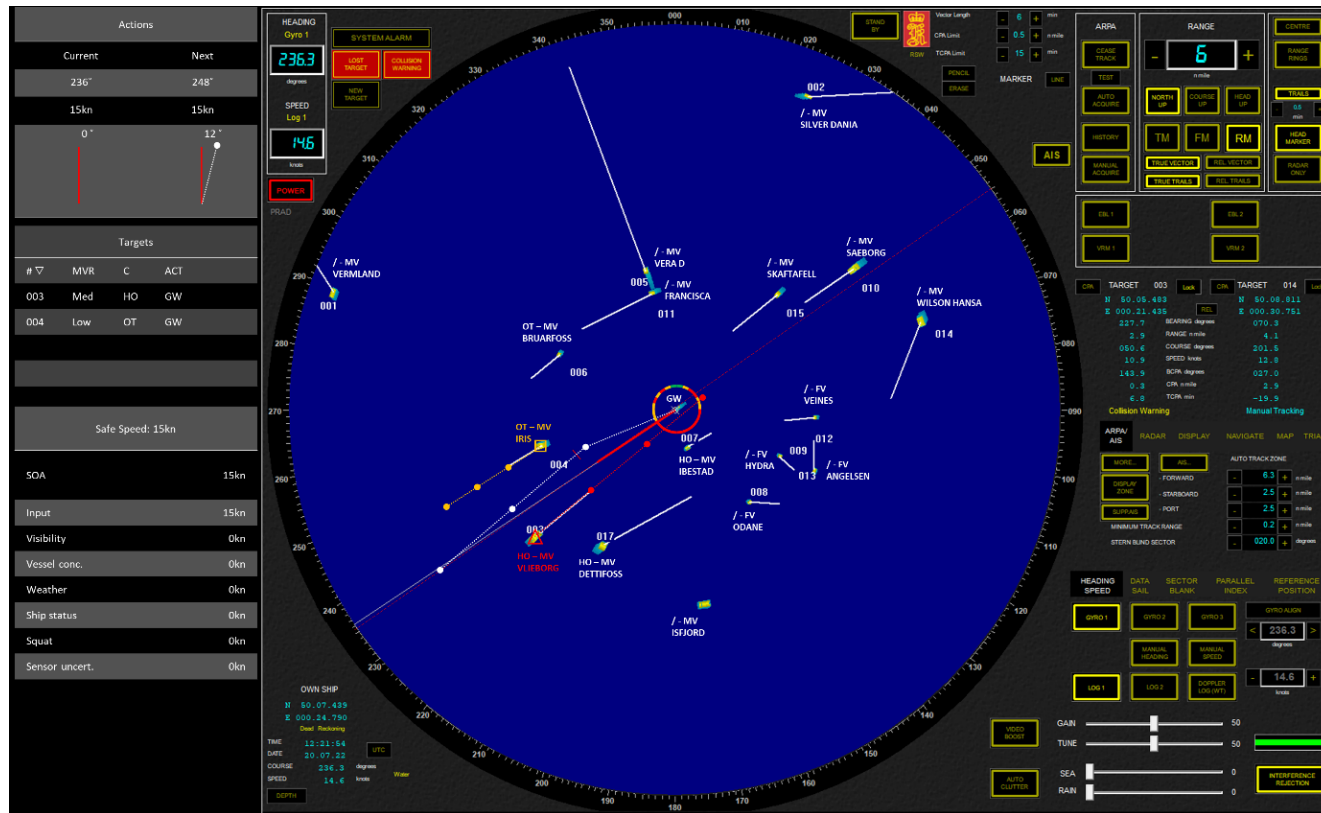


Figure 25. Traffic situation with a medium (A) level of transparency.

Figure 26 provides an example of how a transparent CAGA system could look like when all layers are depicted except the “information analysis” layer. In this transparency configuration, the supervisor is provided with information regarding the system’s decisions and actions, and which information is has acquired. However, it does not provide information about how it analyses this information, e.g., which risks it has determined. This level of transparency was included to provide an alternative to the cumulative approach discussed above where each level of transparency was added to the next one, i.e., low, medium, high.

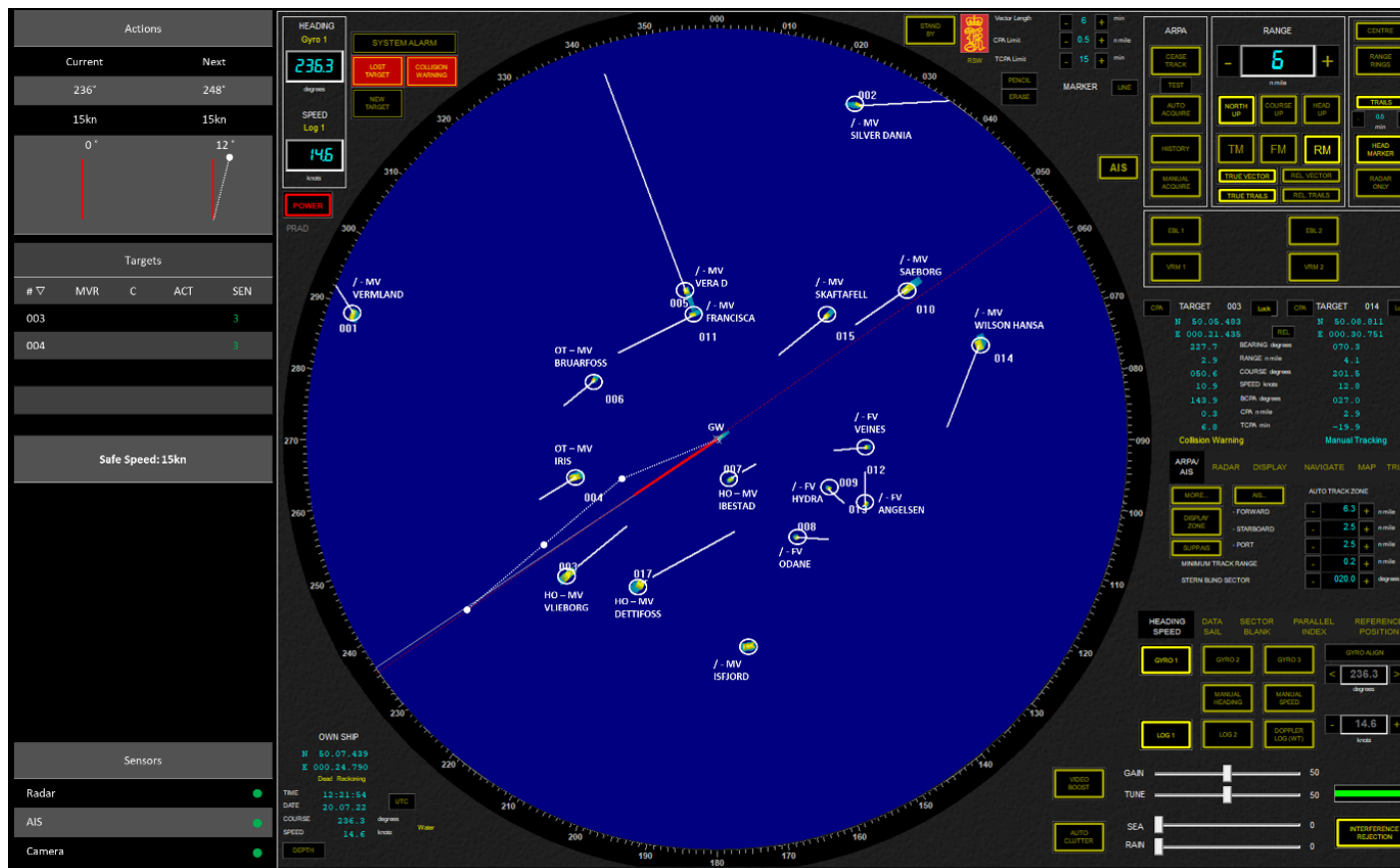


Figure 26. Traffic situation with a medium (B) level of transparency.

Finally, Figure 27 provides a depiction of what a transparent collision avoidance system could look like when all transparency information identified through the task analysis is provided on the HMI. Here, all targets have received identifiers (green circles), and initial classifications (ship types and relevant conflict type indicators). In addition, information regarding the status of the system's sensors is provided in the tables to the left of the radar screen.

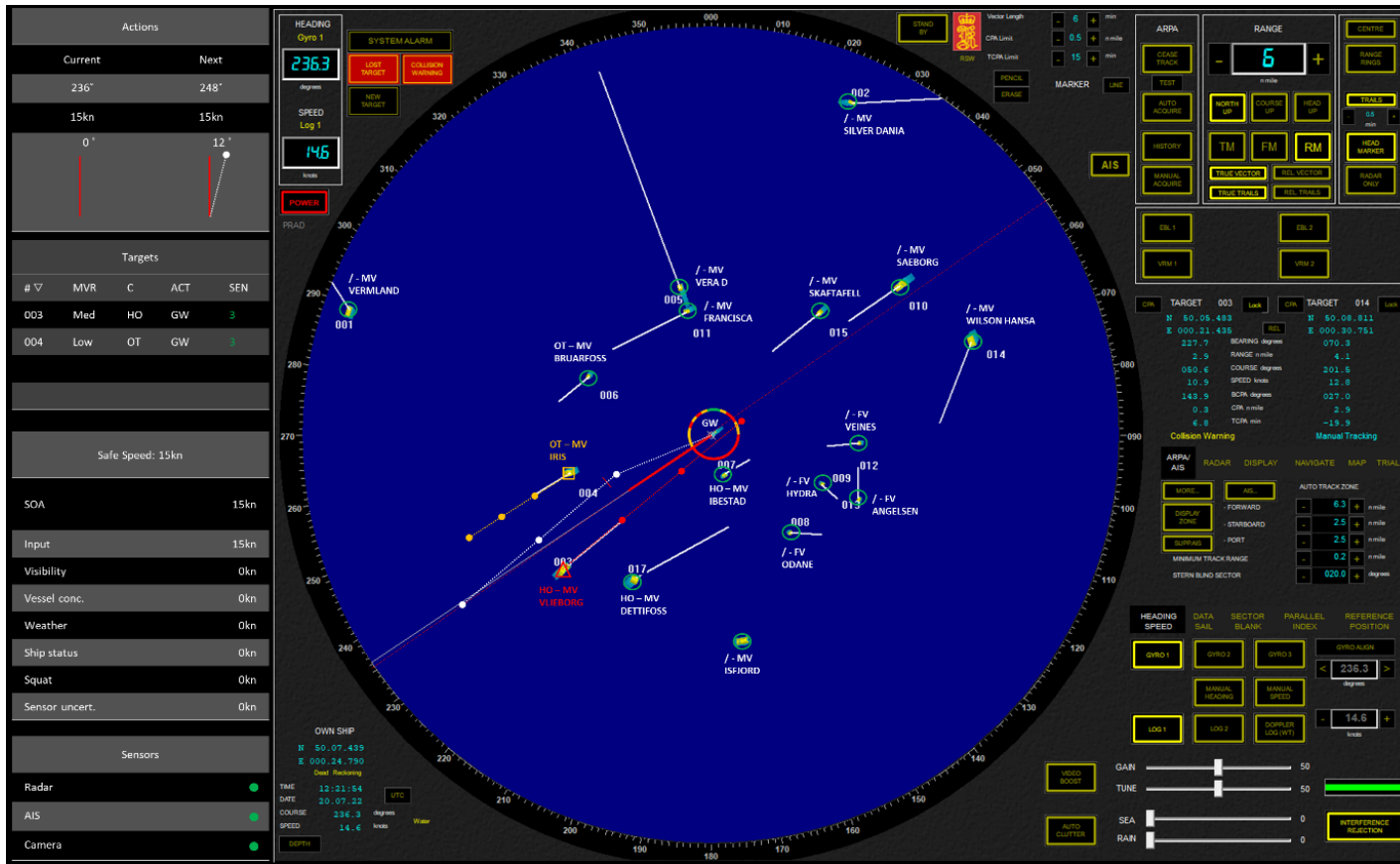


Figure 27. Traffic situation with a high level of transparency.

## Appendix D – Traffic situations used in the experiment

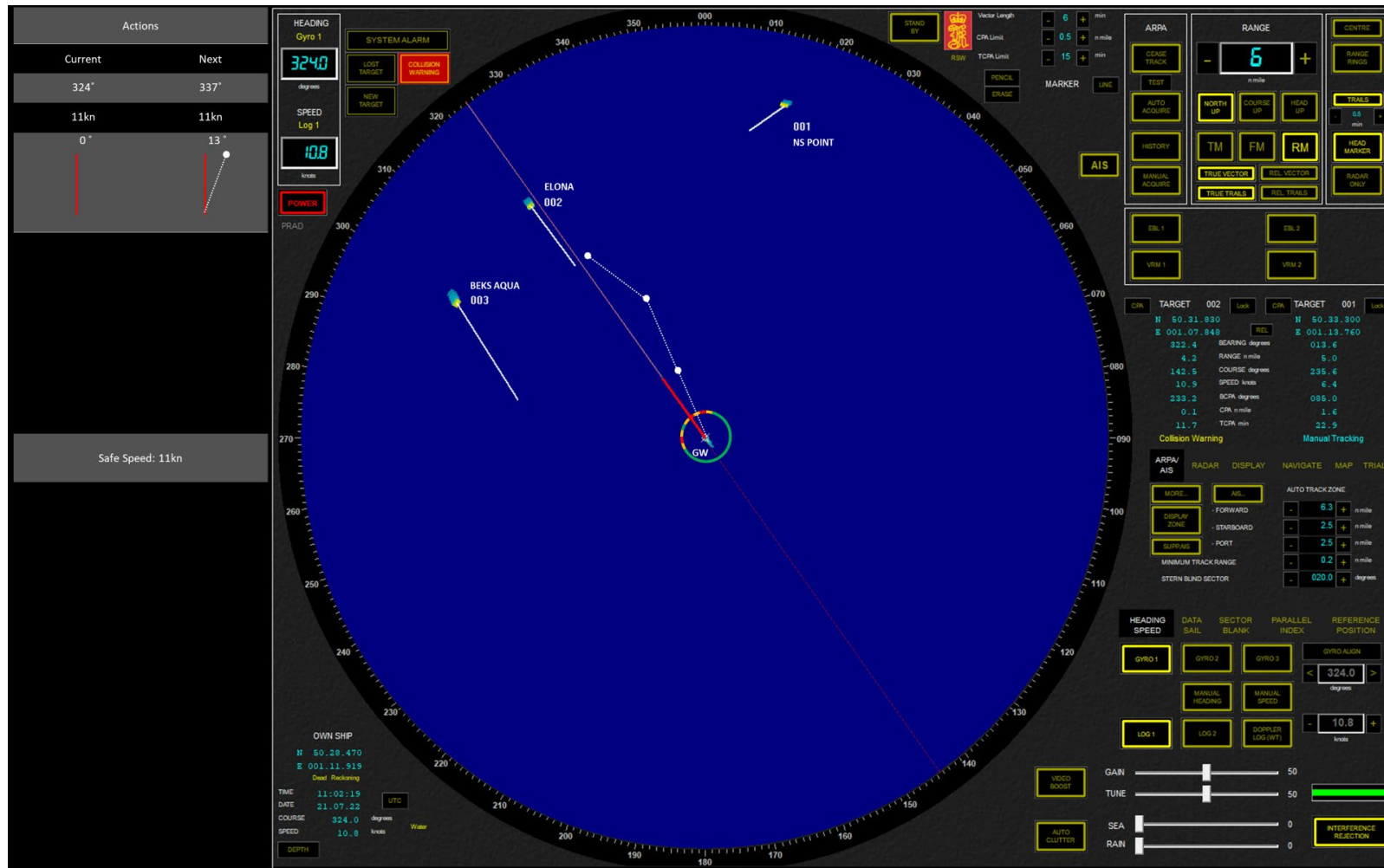


Figure 28. Trial 1 - complexity = low, transparency = low, head-on (7HOLLT3).

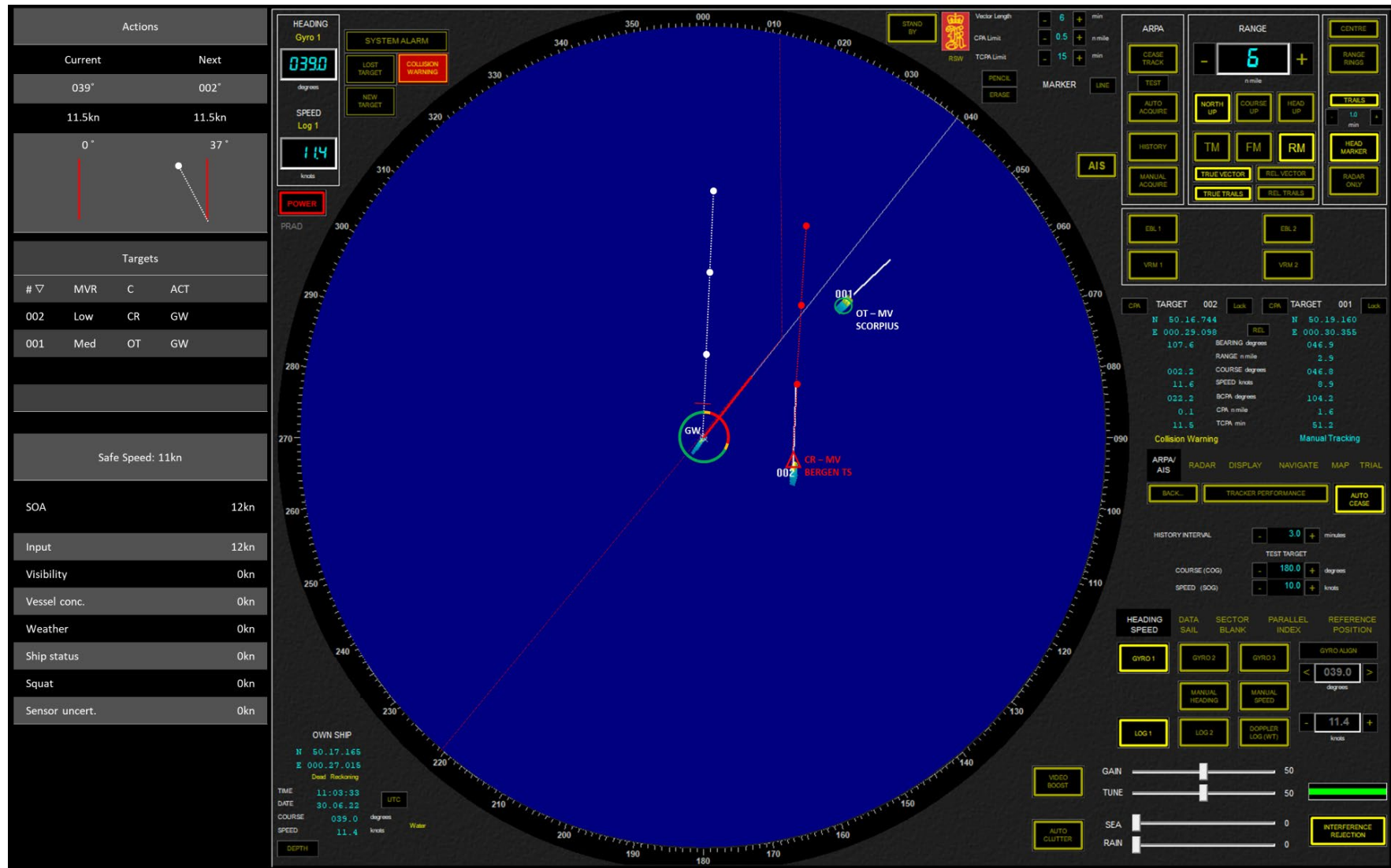


Figure 29. Trial 1 - complexity = low, transparency = medium (A), crossing (2CRLLT32).



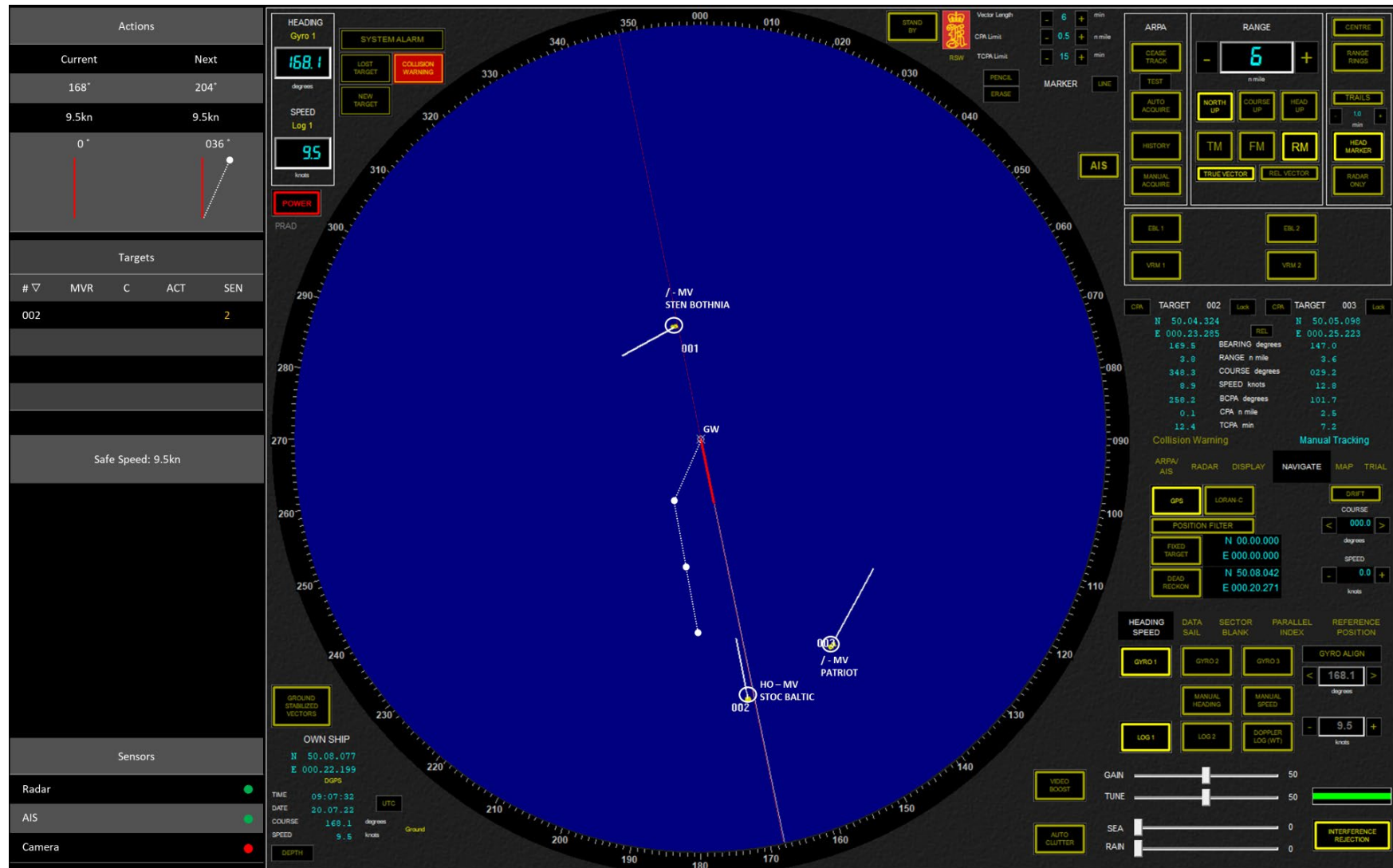


Figure 30. Trial 1 - complexity = low, transparency = medium (B), head-on (2HOLLT31).

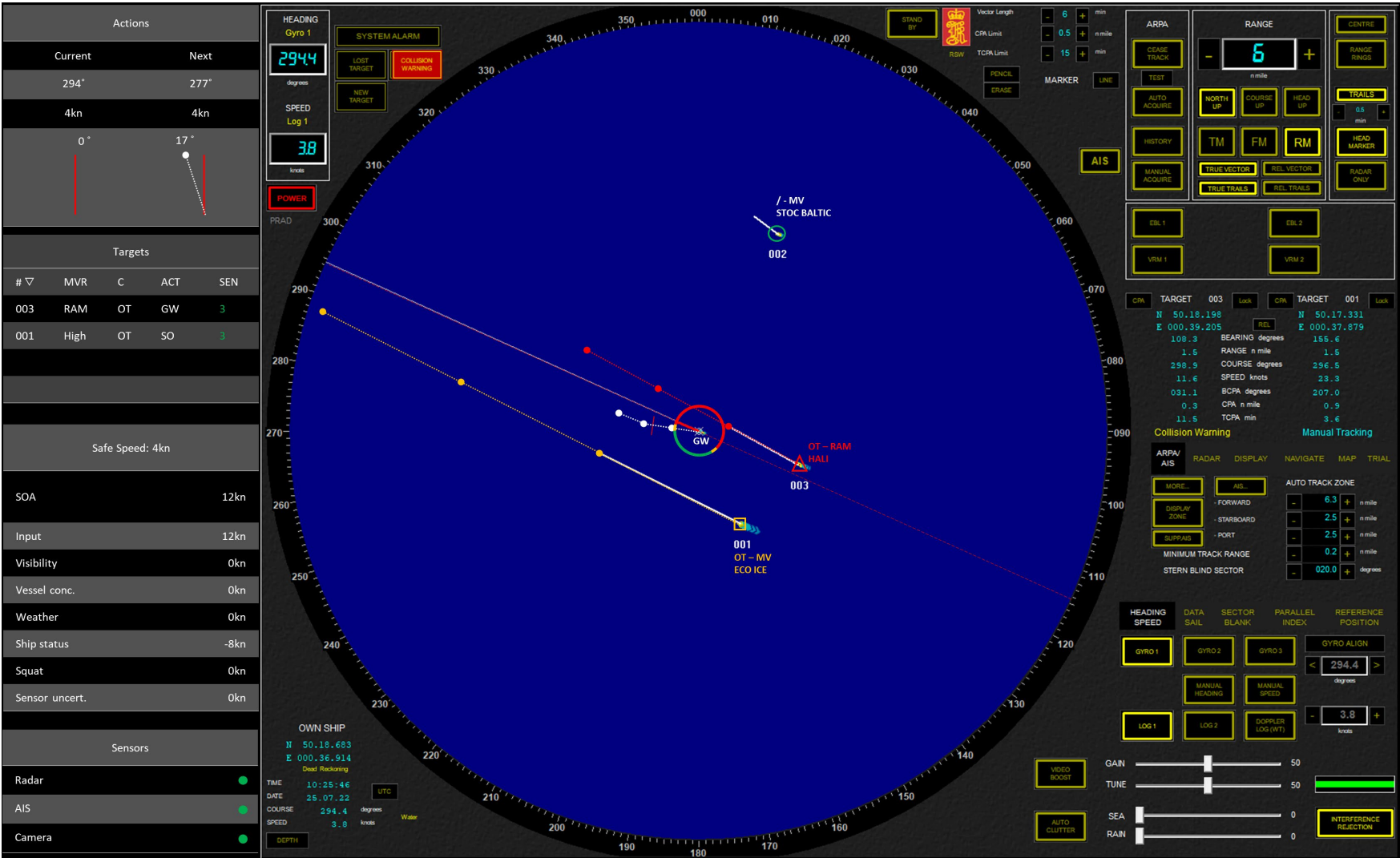


Figure 31. Trial 1 - complexity = low, transparency = high, overtaking (100TLLT321).

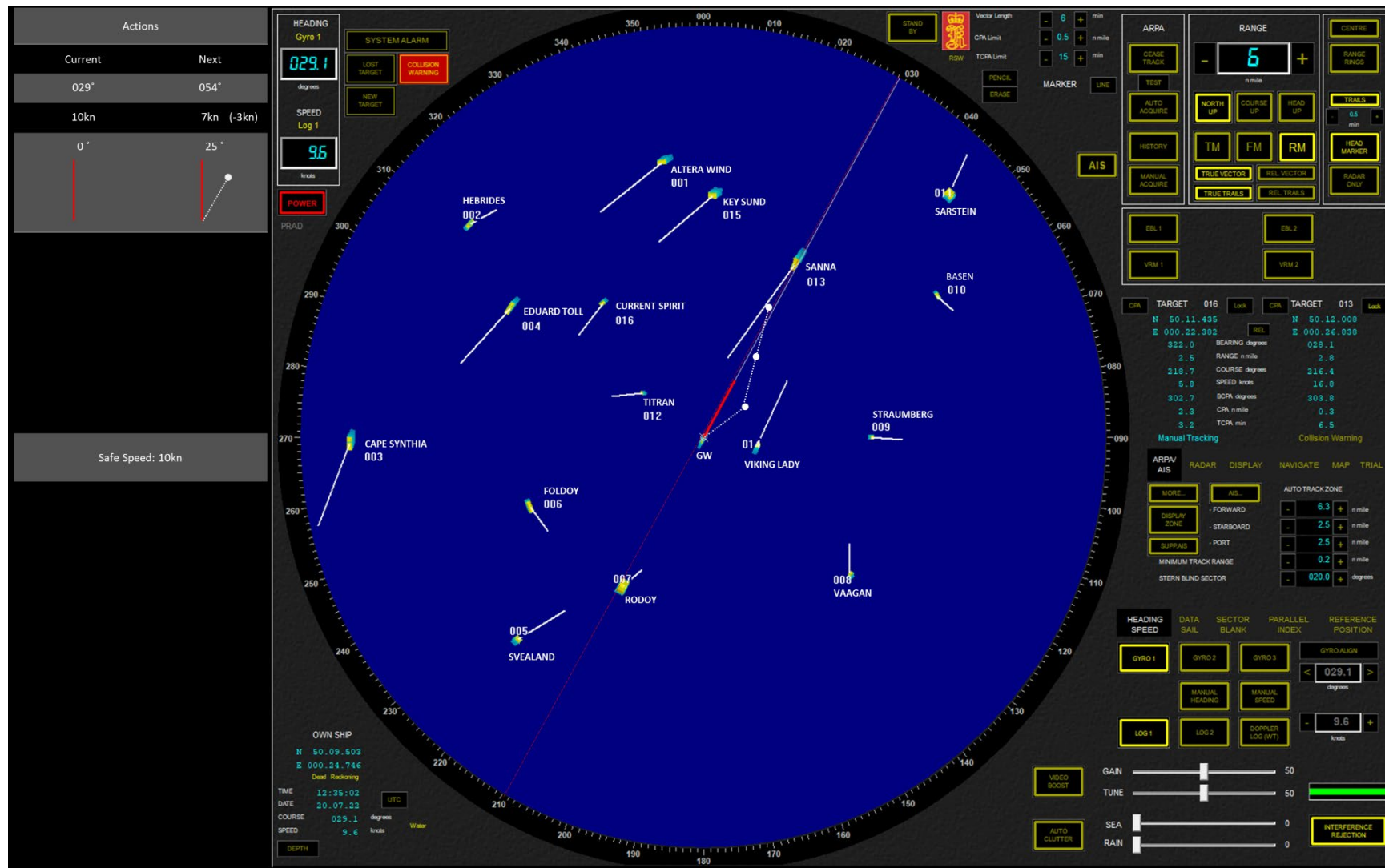


Figure 32. Trial 1 - complexity = high, transparency = low, head-on (15HOHLT3).

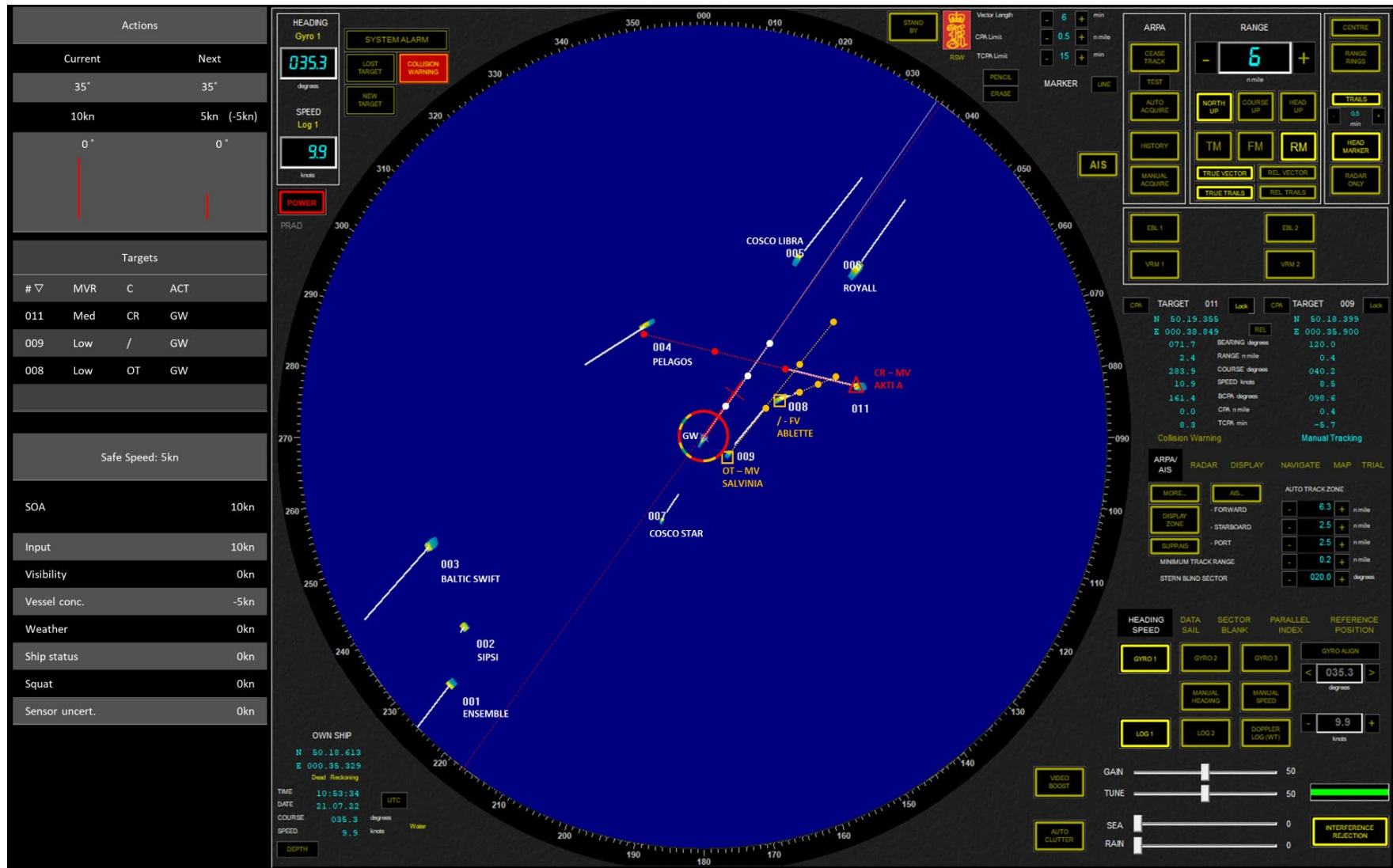


Figure 33. Trial 1 - complexity = high, transparency = medium (A), crossing (21CRHLT32).

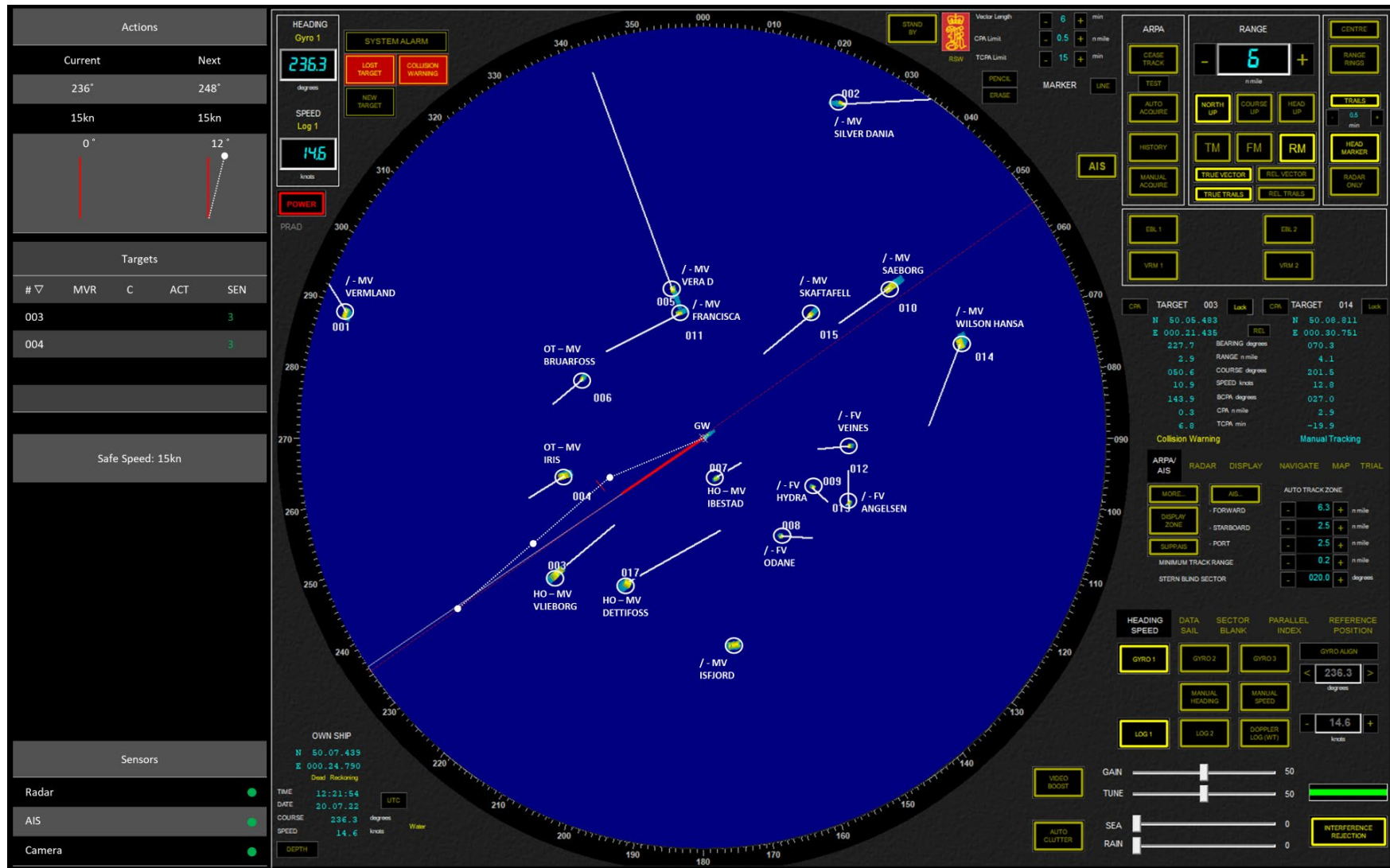


Figure 34. Trial 1 - complexity = high, transparency = medium (B), head-on (11HOHLT31).

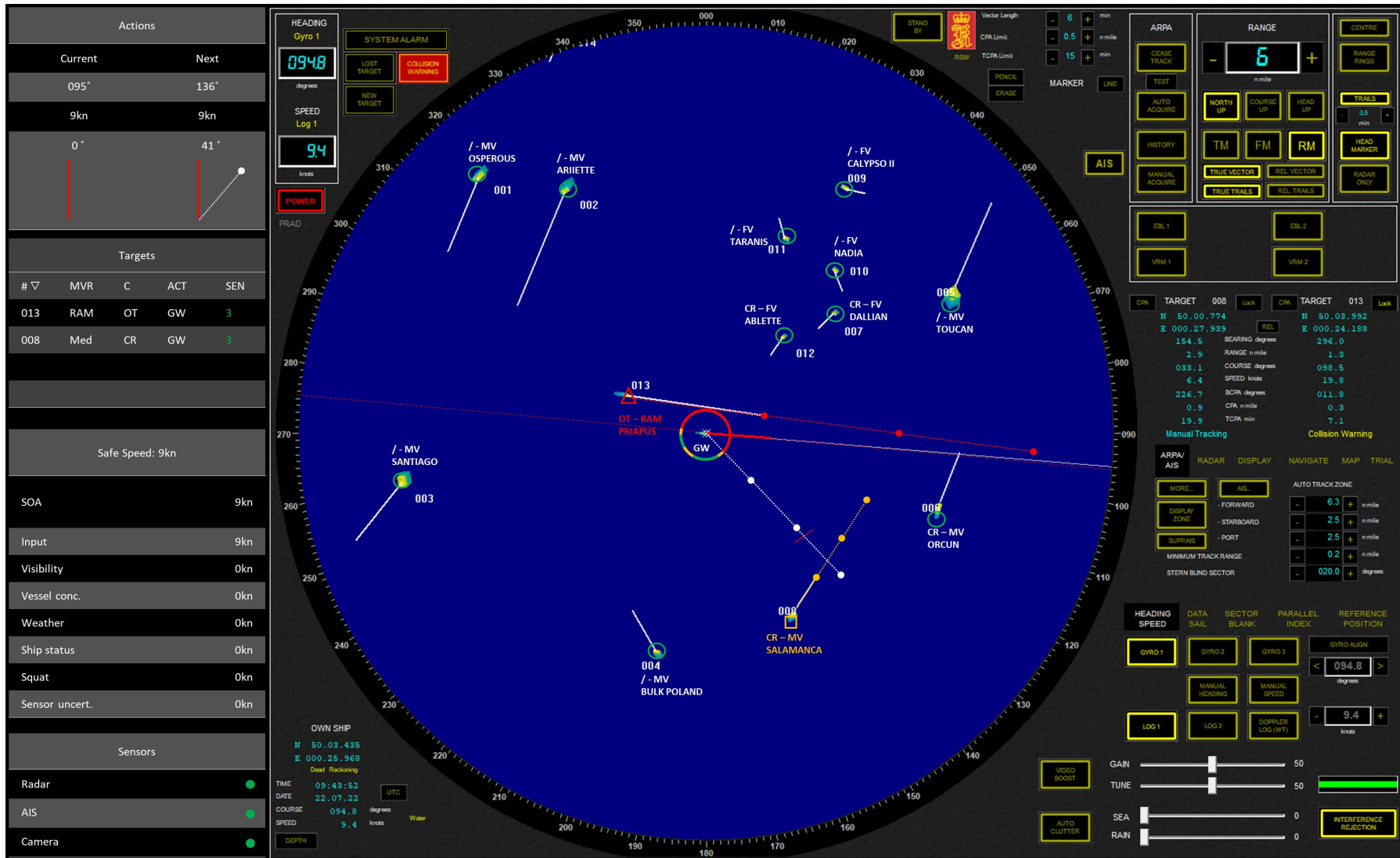


Figure 35. Trial 1 - complexity = high, transparency = high, overtaking (170THLT321).

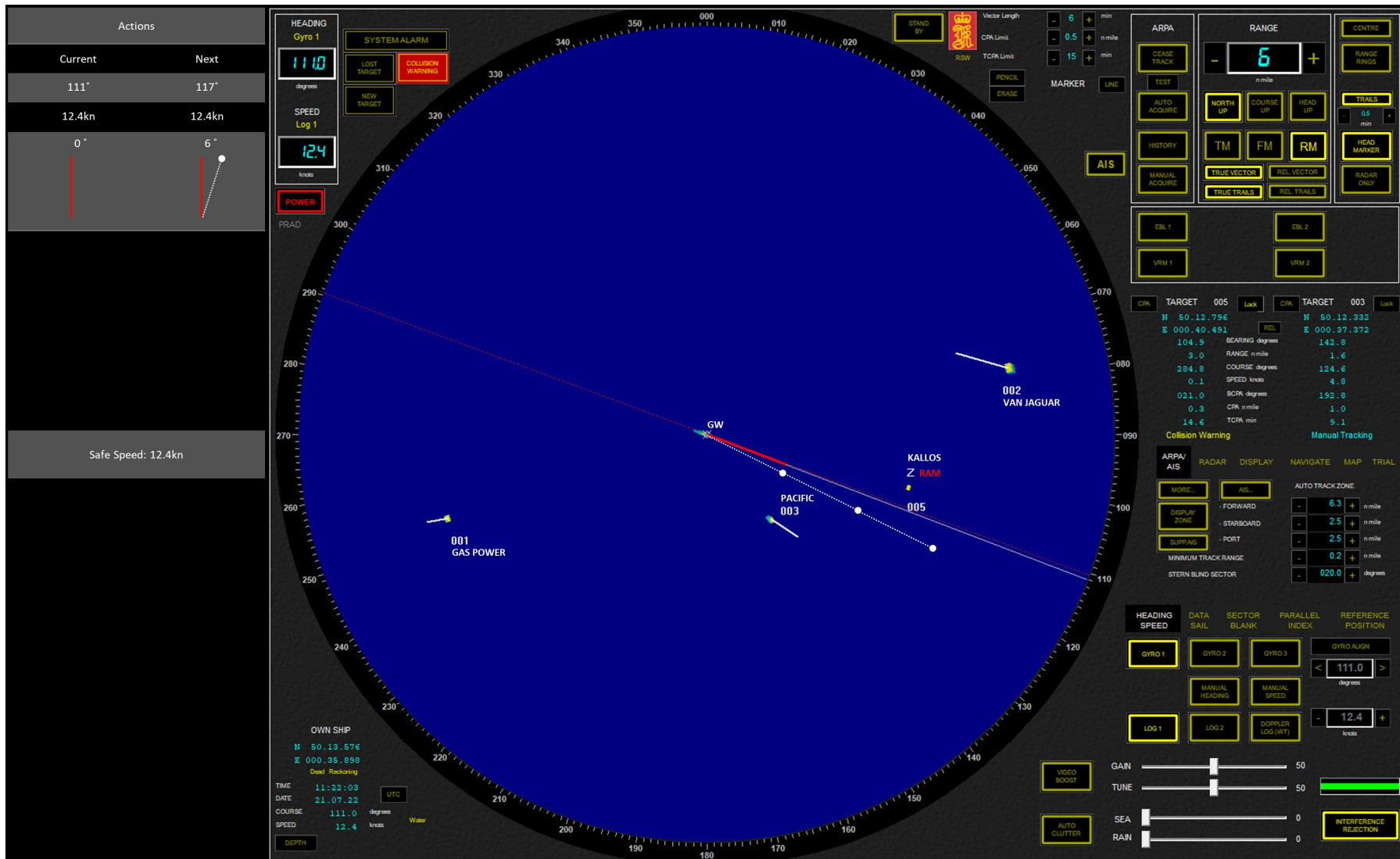


Figure 36. Trial 2 - complexity = low, transparency = low, head-on (9HOLTL3).

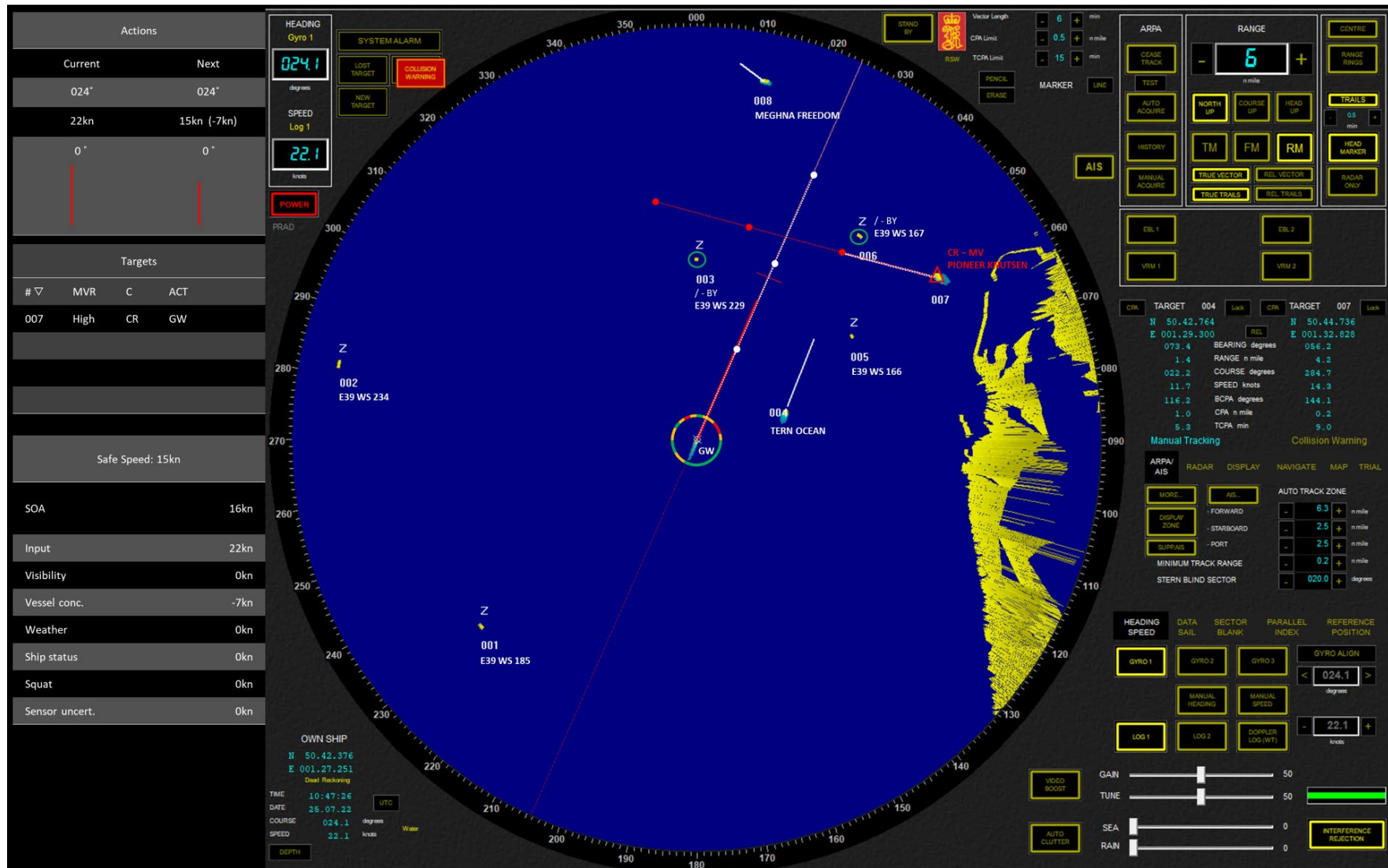


Figure 37. Trial 2 - complexity = low, transparency = medium (A), crossing (13CRLTL32).



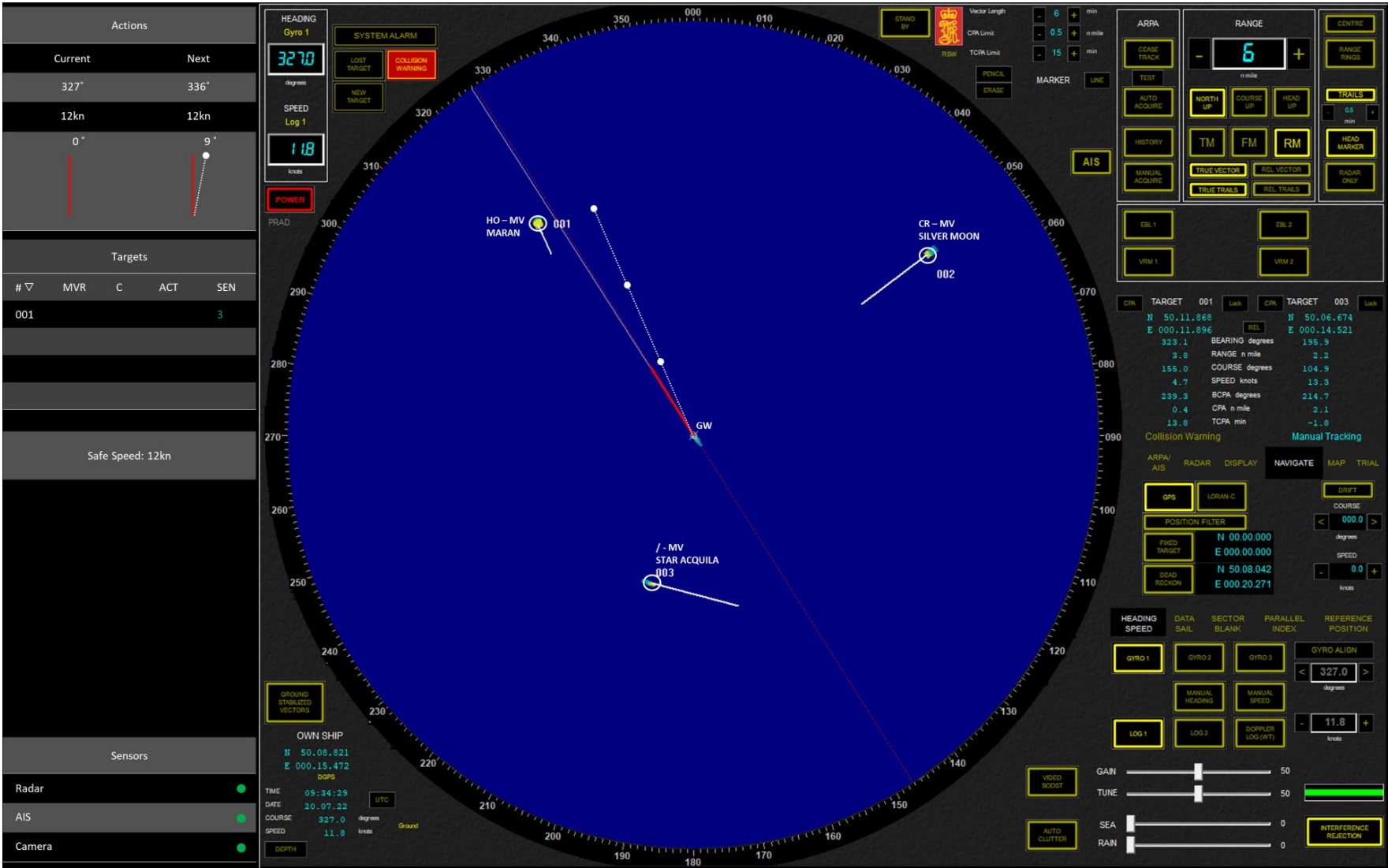


Figure 38. Trial 2 - complexity = low, transparency = medium (B), head-on (5HOLTL31).

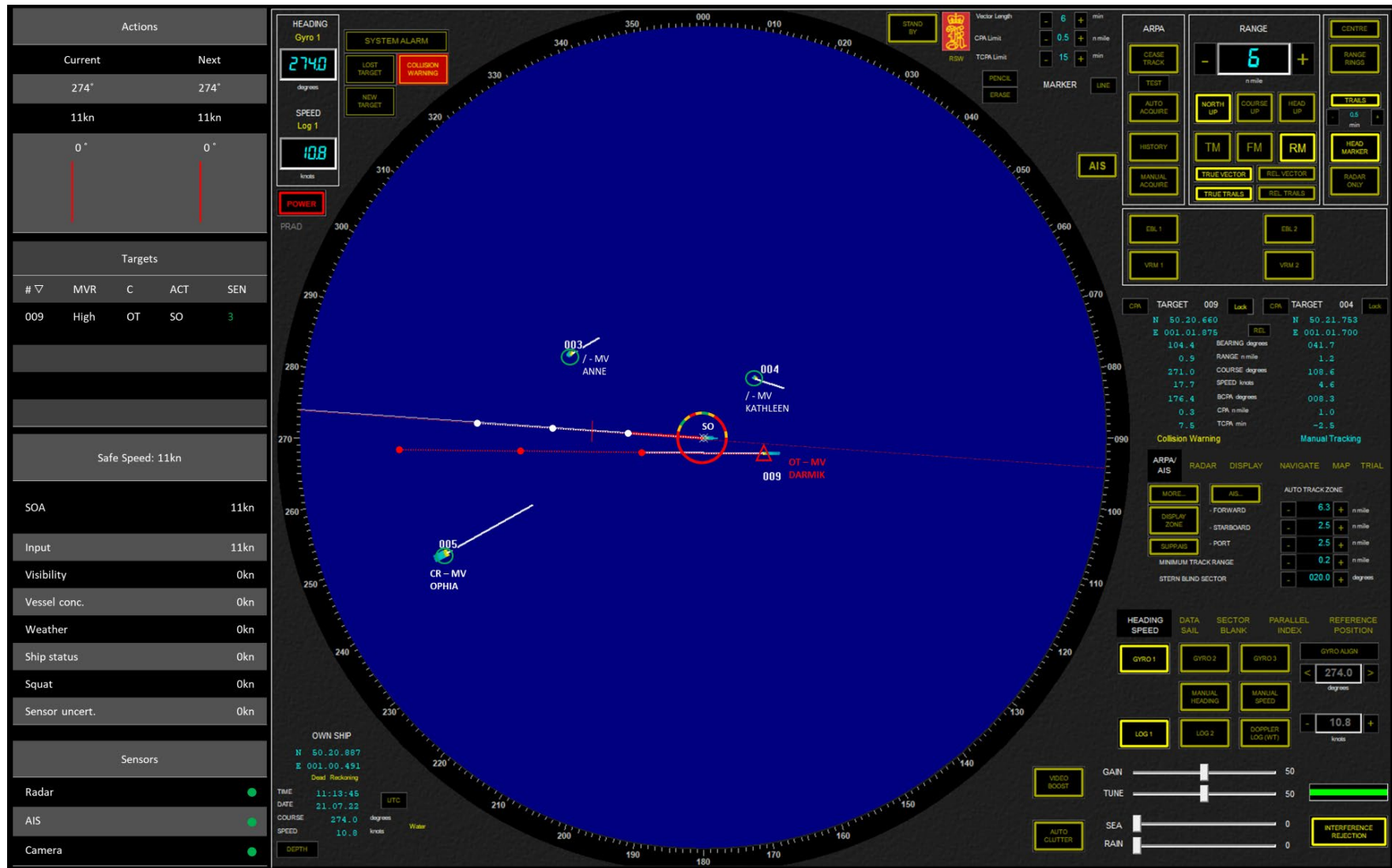


Figure 39. Trial 2 - complexity = low, transparency = high, overtaking (90TLTL321).

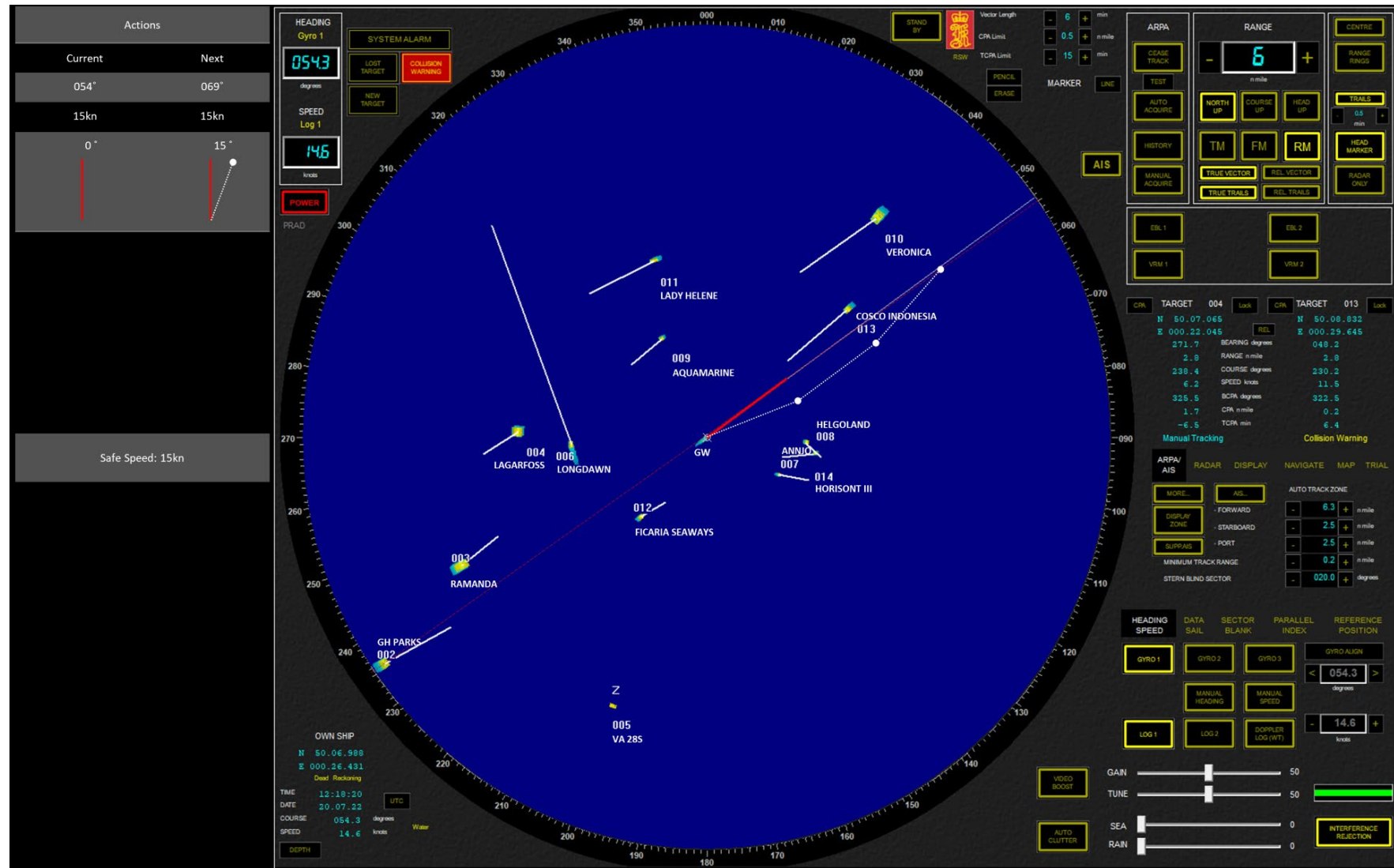


Figure 40. Trial 2 - complexity = high, transparency = low, head-on, (10HOHTL3).

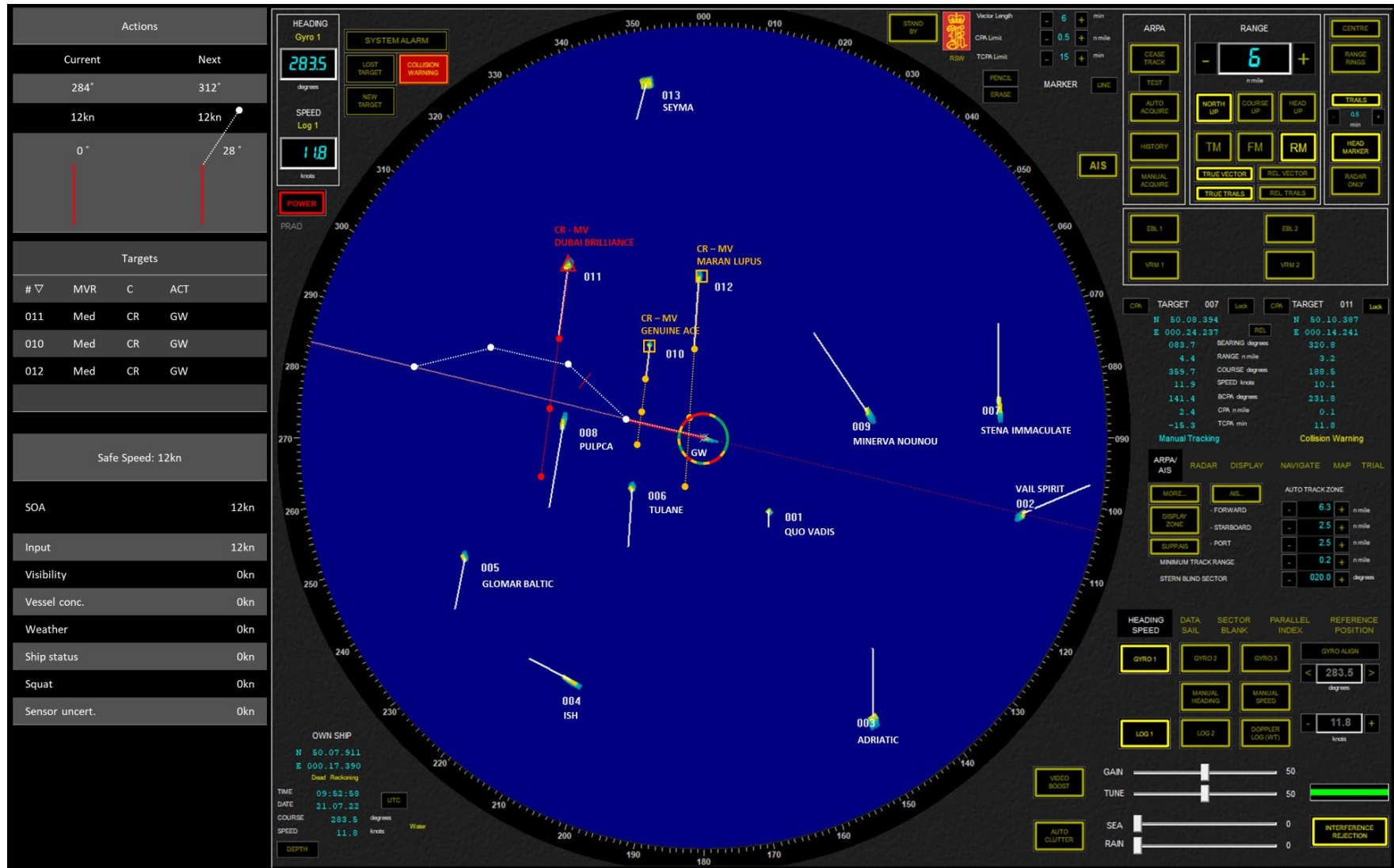


Figure 41. Trial 2 - complexity = high, transparency = medium (A), crossing (15CRHTL32).

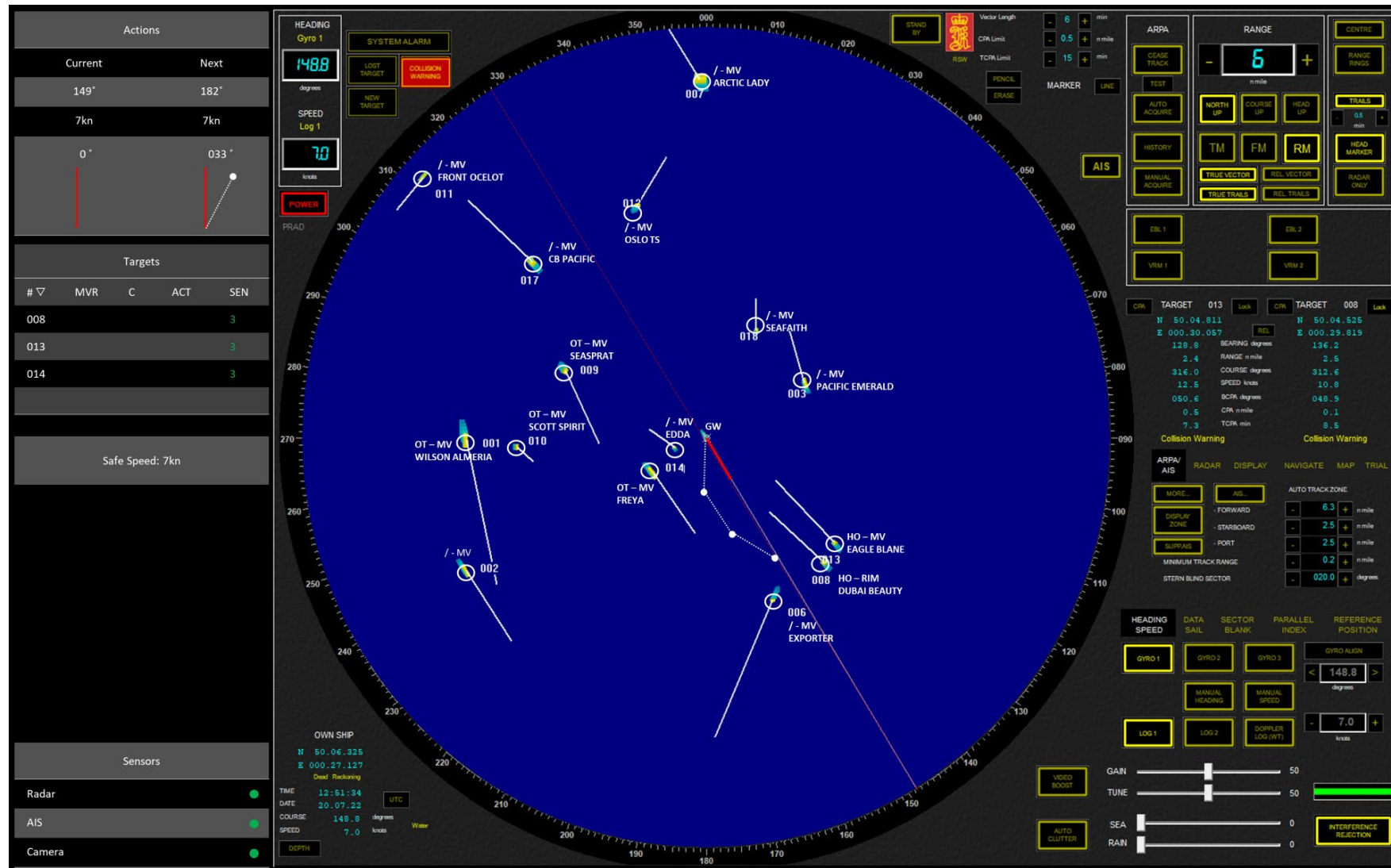


Figure 42. Trial 2 - complexity = high, transparency = medium (B) head-on (13HOHTL31).

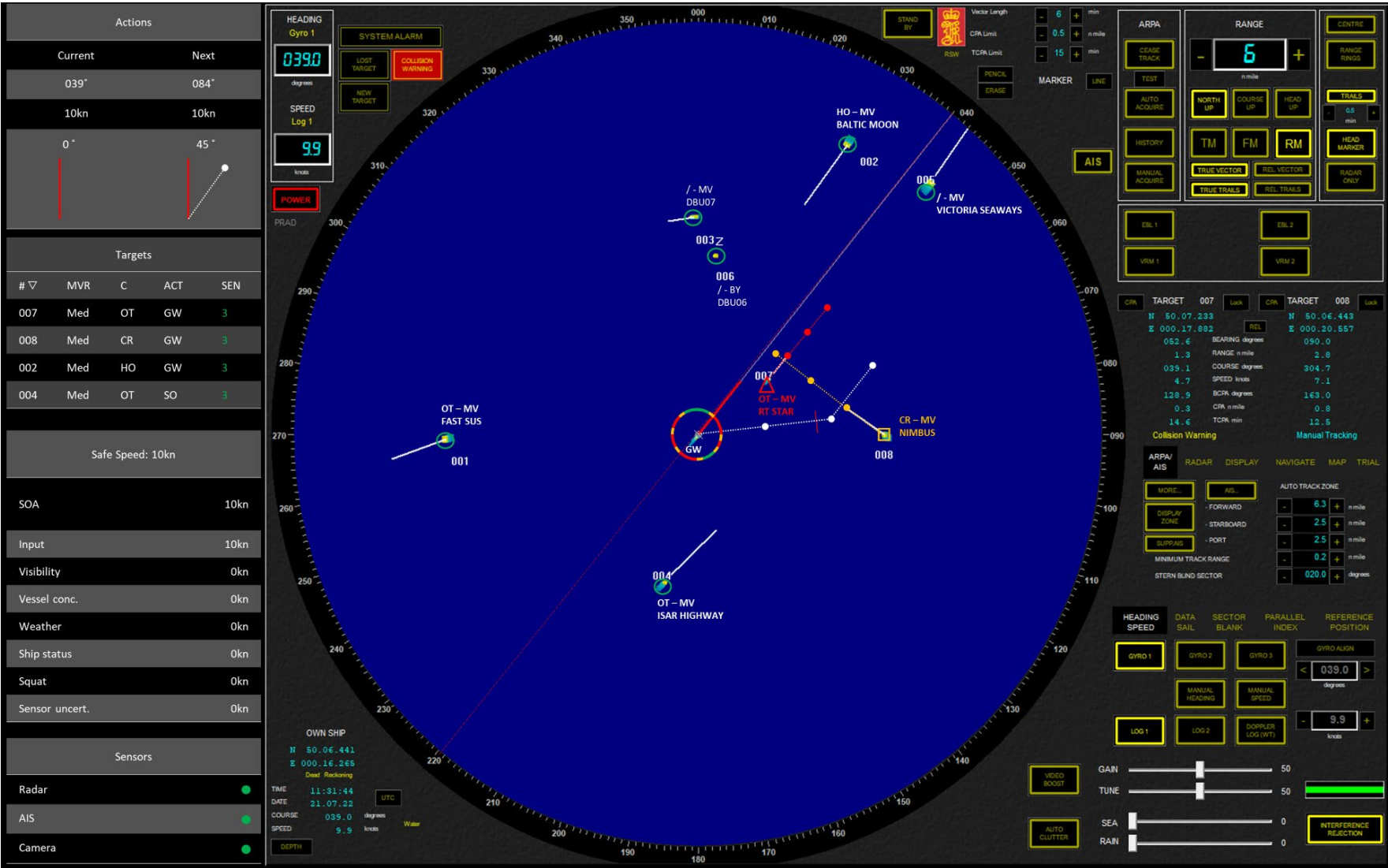


Figure 43. Trial 2 - complexity = high, transparency = high, crossing (130THTL321).

## Appendix E – Generic SAGAT queries

Requirements for the questions:

1. Questions should probe the participant’s understanding of how the autonomous collision avoidance system is managing the traffic situation
2. Questions should not probe the participant’s understanding of COLREGs directly
3. Questions should be answerable for all levels of transparency

In short: “What is your ship perceiving, thinking, and doing?”

*Table 15. Generic Level 1 SAGAT queries.*

*\*Three alternatives only in the situation-specific query.*

Condition detection - “How does the system perceive the elements of the current situation?”		Queries participant’s understanding of...	Can be answered by...	Trial 1	Trial 2
1.1	Which/ How many targets are forward of your ship’s beam? a. N-M b. O-P c. Q-R	Own ship’s detection of targets	Target identifier	10_OT_L_LT321 21_CR_H_LT32	15_CR_H_TL32
1.2	How many targets, within a +/- 20 degrees of arc forward of your ship’s beam, are currently sailing approximately in the same direction as you? a. N-M b. O-P c. Q-R	Own ship’s detection of targets Own ship’s risk determination	Target conflict type classifier	15_HO_H_LT3	10_HO_H_TL3
1.3	How many targets, within +/- 20 degrees of arc forward of your ship’s beam, are sailing approximately in the opposite direction of you? a. N-M b. O-P c. Q-R	Own ship’s detection of targets Own ship’s risk determination	Target conflict type classifier	2_HO_L_LT31	13_OT_H_TL321

	<b>Condition detection - “How does the system perceive the elements of the current situation?”</b>	<b>Queries participant’s understanding of...</b>	<b>Can be answered by...</b>	<b>Trial 1</b>	<b>Trial 2</b>
1.4	How many targets, within a +/- 20 degrees of arc forward of your ship’s beam, are currently sailing approximately away from you? a. N-M b. O-P c. Q-R	Own ship’s risk determination	Target conflict type classifier	2_CR_L_LT32	13_CR_L_TL32
1.5	How many targets, within a +/- 20 degrees of arc forward of your ship’s beam, are currently sailing approximately towards you? a. N-M b. O-P c. Q-R	Own ship’s risk determination	Target conflict type classifier	17_OT_H_LT321	9_OT_L_TL321
1.6	How many targets, within a +/- 20 degrees of arc forward of your ship’s beam, are crossing your bow? a. N-M b. O-P c. Q-R	Own ship’s risk determination	Target conflict type classifier	7_HO_L_LT3	5_HO_L_TL31
1.7	What is the location of target X?* a. Starboard b. Port c. Forward d. Aft	Own ship’s detection of targets	Target identifier Target conflict type classifier	11_HO_H_LT31	9_HO_L_TL3 13_HO_H_TL31



Table 16. Generic Level 2 SAGAT queries.

\*Three alternatives only in the situation-specific query.

Condition analysis - “How does the system comprehend the current situation?”		Queries participant’s understanding of...	Can be answered by...	Trial 1	Trial 2
2.1	In which collision situation is your ship in?*	Own ship’s understanding of collision situation type	Target conflict type classifier	2_HO_L_LT31	9_OT_L_TL321
	a. Head-on		Target information table		
	b. Overtaking				
	c. Crossing				
	d. No collision situation				
2.2	Which targets pose a collision risk to your ship?	Own ship’s risk determination	Target risk classifier	15_HO_H_LT3	15_CR_H_TL32
	a. Target X & Y		Target conflict type classifier		
	b. Target X & Z		Target predicted track		
	c. None of the above		Target information table		
2.3	Which target poses the highest collision risk to your ship?	Own ship’s risk determination	Target risk classifier	10_OT_L_LT321	13_CR_L_TL32
	a. Target X		Target conflict type classifier		
	b. Target Y		Target predicted track		
	c. No collision risk		Target information table		
2.4	What target poses a secondary risk, i.e., limiting your ability to perform an avoidance manoeuvre?	Own ship’s risk determination	Target predicted track		13_OT_H_TL321
	a. Target X				
	b. Target Y				
	c. No secondary risk				
2.5	What is the speed of target X relative to your own ship’s speed?	Own ship’s risk determination	Target predicted track	11_HO_H_LT31	10_HO_H_TL3
	a. Faster than own ship				
	b. Slower than own ship				
	c. Approximately the same speed				

<b>Condition analysis - "How does the system comprehend the current situation?"</b>		<b>Queries participant's understanding of...</b>	<b>Can be answered by...</b>	<b>Trial 1</b>	<b>Trial 2</b>
2.6	What is the course of the target X relative to your own ship's course? a. Crossing from starboard b. Crossing from port c. Not crossing	Own ship's risk determination	Target predicted track		
2.7	What is the direction of target X relative to your own ship's course? a. Towards own ship b. Away from own ship c. In the same direction as own ship	Own ship's risk determination	Target predicted track	7_HO_L_LT3	5_HO_L_TL31
2.8	What does your ship intend to do for target X? a. Give-way b. Stand-on c. Take no action	Own ship's understanding of target ship priority	Target conflict type classifier Target predicted track Target information table	17_OT_H_LT321	13_HO_H_TL31
2.9	Which target, within a +/- 20 degrees of arc forward of your ship's beam, is sailing in approximately the same direction as you? a. Target X b. Target Y c. None of the above	Own ship's risk determination	Target conflict type classifier Target predicted track Target information table		
2.10	Which target, within +/- 20 degrees of arc forward of your ship's beam, is sailing approximately in the opposite direction of you? a. Target X b. Target Y c. None of the above	Own ship's risk determination	Target conflict type classifier Target predicted track Target information table		

<b>Condition analysis - “How does the system comprehend the current situation?”</b>		<b>Queries participant’s understanding of...</b>	<b>Can be answered by...</b>	<b>Trial 1</b>	<b>Trial 2</b>
2.11	Which target, within +/- 20 degrees of arc forward of your ship’s beam, is sailing approximately away from you? a. Target X b. Target Y c. None of the above	Own ship’s risk determination	Target conflict type classifier Target predicted track Target information table	2_CR_L_LT32	
2.12	Which target, within +/- 20 degrees of arc forward of your ship’s beam, is currently sailing towards you? a. Target X b. Target Y c. None of the above	Own ship’s risk determination	Target conflict type classifier Target predicted track Target information table		9_HO_L_TL3
2.13	Which target, within +/- 20 degrees of arc forward of your ship’s beam, is currently crossing your bow? a. Target X b. Target Y c. None of the above	Own ship’s risk determination	Target conflict type classifier Target predicted track Target information table	21_CR_H_LT32	

Table 17. Generic Level 3 SAGAT queries.

*\*Three alternatives only in the situation-specific query.*

Action planning -	“What is the system’s projection of the situation’s future state?”	Queries participant’s understanding of...	Can be answered by...	Trial 1	Trial 2
3.1	What does your ship intend to do? a. Give-way b. Stand-on c. Take no action	Priority for own ship	Own ships future track Own ships action indicator Target ship action indicator Action table	21_CR_H_LT32	15_CR_H_TL32
3.2	Which action does your ship plan to take?*	Own ship’s intended manoeuvre (course and speed)	Own ship future track Action table		
3.3	In which direction does your ship plan to make a course change? a. Starboard b. Port c. No change in direction	Own ship’s intended course	Own ship future track Action table		13_CR_L_TL32
3.4	For target ship X, what is your ship’s intention? a. Keep on its starboard side b. Keep on its port side c. Pass its bow	Own ship’s intended course	Own ship future track	2_CR_L_LT32	13_OT_H_TL321
3.5	What is the <u>absolute</u> course your ship intends to sail in the next six minutes? a. A-B degrees b. B-C degrees c. C-D degrees	Own ship’s intended course	Own ship future track Action table	7_HO_L_LT3	13_HO_H_TL31

3.6	What is the <u>relative</u> course change your ships intends to implement in the next six minutes? a. A-B degrees b. B-C degrees c. C-D degrees	Own ship's intended course	Own ship future track Action table	17_OT_H_LT321	
3.7	What is the speed your ship intends to sail for the next six minutes? a. X kn b. Y kn c. Z kn	Own ship's intended speed	Safe speed table Own ship future track	2_HO_L_LT31	10_HO_H_TL3
3.8	What is the absolute speed change your ship intends to implement in the next six minutes? a. 0 kn (no change) b. X-Y kn c. Y-Z kn	Own ship's intended speed	Action table Safe speed table Own ship future track	15_HO_H_LT3	9_HO_L_TL3
3.9	What speed change does your ship plan to make? a. Increase b. Decrease c. No change	Own ship's intended speed	Own ship future track Action table Safe speed table	11_HO_H_LT31	9_OT_L_TL321
3.10	When does your ship intend to perform its avoidance manoeuvre? a. Immediately b. After X amount of minutes c. No manoeuvre	Own ship's intended manoeuvre	Own ship future track Action table	10_OT_L_LT321	5_HO_L_TL31



## Appendix F – NASA-TLX

Figure 44. The rating sheets used for the NASA-TLX (Hart & Staveland, 1988).

**Mental demand**

How much mental and perceptual activity was required?  
(For example, thinking, deciding, calculating, remembering, looking, searching, etc.)  
Was the task easy or demanding, simple or complex, forgiving or exacting?

0      1      2      3      4      5      6      7      8      9      10

Low High

**Physical demand**

How much physical activity was required?  
(For example, pushing, pulling, turning, controlling, activating, etc.)  
Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?

0      1      2      3      4      5      6      7      8      9      10

Low High

**Temporal demand**

How much time pressure did you feel due to the rate or pace at which the tasks or tasks elements occurred?  
Was the pace slow and leisurely or rapid and frantic?

0      1      2      3      4      5      6      7      8      9      10

Low High

**Performance**

How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)?  
How satisfied were you with your performance in accomplishing these goals?

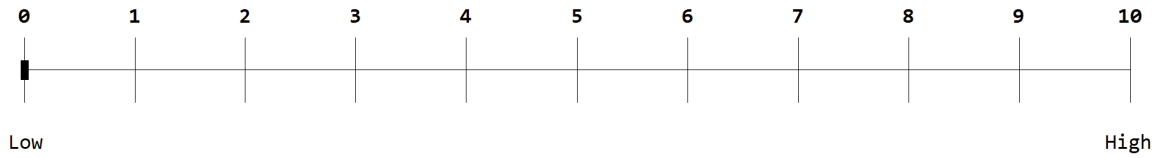
0      1      2      3      4      5      6      7      8      9      10

Good Poor

Note the location of the endpoints are different before answering.

**Effort**

How hard did you have to work (mentally and physically) to accomplish your level of performance?



**Frustration**

How insecure, discouraged, irritated, stressed, and annoyed versus, secure, gratified, content, relaxed, and complacent did you feel during the task?

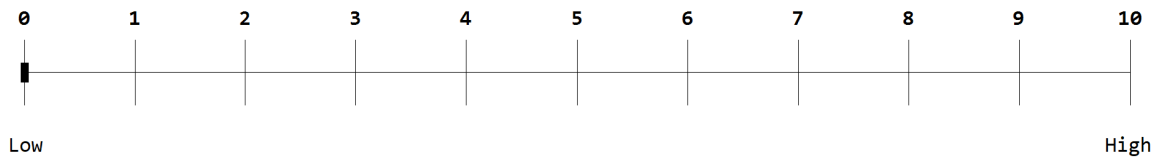


Figure 45. Weighting mental workload scores with pairwise comparisons (Hart & Staveland, 1988).

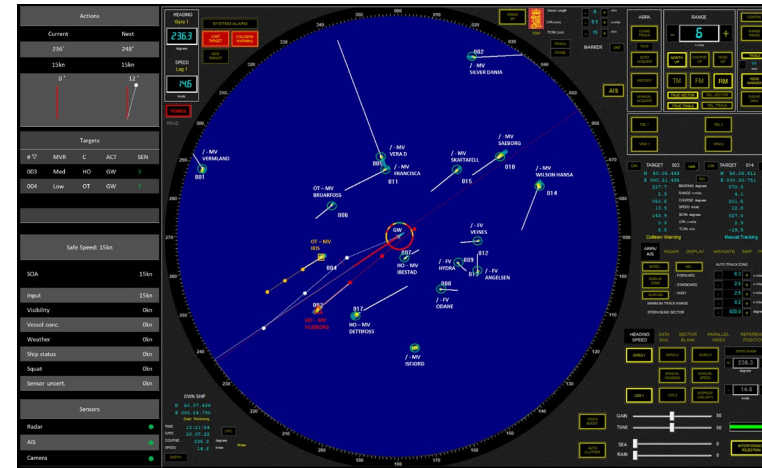
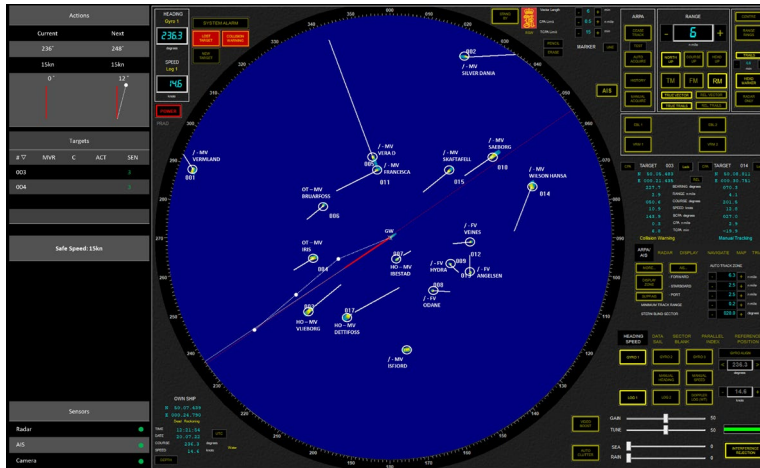
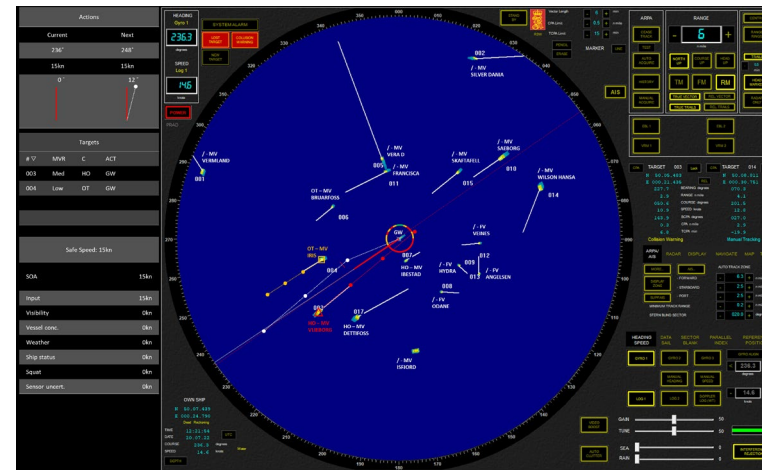
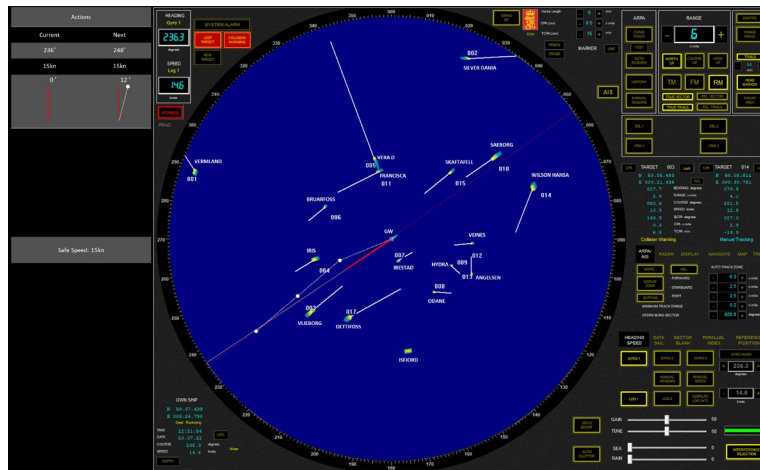
No.	Pair
1	Effort or Performance
2	Temporal demand or Effort
3	Performance or Frustration
4	Physical demand or Performance
5	Temporal demand or Frustration
6	Physical demand or Frustration
7	Physical demand or Temporal demand
8	Temporal demand or Mental demand
9	Frustration or Effort
10	Performance or Temporal demand
11	Mental demand or Physical demand
12	Frustration or Mental demand
13	Performance or Mental demand
14	Mental demand or Effort
15	Effort or Physical demand



# Appendix G – Ranking transparency levels

Table 18. The traffic situations used for ranking participants' preferences.

Top-left: Low, Top-right: Medium (A), Bottom-left: Medium (B), Bottom-right: High. For larger versions, see Appendix C – Examples of transparency levels.



*Table 19. Definitions and examples that were read verbatim to the participants prior to ranking.*

<b>Dimension</b>	<b>Definition</b>	<b>Example</b>
Observability	Observability means the system proactively communicates with you to let you know what it's thinking and doing and tells you how far along it is in accomplishing your joint work.	A collision avoidance system that tells you how it interprets the current traffic situation.
Predictability	Predictability means the system communicates with you about its intentions, goals, and future actions in various contexts.	A collision avoidance system that tells you how it predicts the current traffic situation will develop in the future.

*Table 20. Ranking preferences for transparency levels.*

<b>Dimension</b>	<b>Transparency level</b>	<b>Ranking (forced choice)</b>
Observability	Low	
	Medium (A)	
	Medium (B)	
	High	
Predictability	Low	
	Medium (A)	
	Medium (B)	
	High	

## Appendix I – Publications

### Article 1

van de Merwe, K., Mallam, S., & Nazir, S. (2024). Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review. *Human Factors*, 66(1), 180–208. <https://doi.org/10.1177/00187208221077804>

### Article 2

van de Merwe, K., Mallam, S., Nazir, S., & Engelhardtson, Ø. (2024). Supporting human supervision in autonomous collision avoidance through agent transparency. *Safety Science*, 169, 13. <https://doi.org/10.1016/j.ssci.2023.106329>

### Article 3

van de Merwe, K., Mallam, S., Engelhardtson, Ø., & Nazir, S. (2023). Towards an approach to define transparency requirements for maritime collision avoidance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 67(1), 483–488. <https://doi.org/10.1177/21695067231192862>

### Article 4

van de Merwe, K., Mallam, S., Engelhardtson, Ø., & Nazir, S. (2023). Operationalising Automation Transparency for Maritime Collision Avoidance. *TransNav, International Journal on Marine Navigation and Safety of Sea Transportation*, 17(2). <https://doi.org/10.12716/1001.17.02.09>

### Article 5

van de Merwe, K., Mallam, S., Nazir, S., & Engelhardtson, Ø. (2024). The Influence of Agent Transparency and Complexity on Situation Awareness, Mental Workload, and Task Performance. *Journal of Cognitive Engineering and Decision Making*, 18(2), 156–184. <https://doi.org/10.1177/15553434241240553>



**Article 1**

van de Merwe, K., Mallam, S., & Nazir, S. (2024). Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review. *Human Factors*, 66(1), 180–208. <https://doi.org/10.1177/00187208221077804>



# Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review

Koen van de Merwe , DNV, Høvik, Norway, University of South-Eastern Norway, Borre, Norway, Steven Mallam  and Salman Nazir , University of South-Eastern Norway, Borre, Norway

**Objective:** In this review, we investigate the relationship between agent transparency, Situation Awareness, mental workload, and operator performance for safety critical domains.

**Background:** The advancement of highly sophisticated automation across safety critical domains poses a challenge for effective human oversight. Automation transparency is a design principle that could support humans by making the automation's inner workings observable (i.e., "seeing-into"). However, experimental support for this has not been systematically documented to date.

**Method:** Based on the PRISMA method, a broad and systematic search of the literature was performed focusing on identifying empirical research investigating the effect of transparency on central Human Factors variables.

**Results:** Our final sample consisted of 17 experimental studies that investigated transparency in a controlled setting. The studies typically employed three human-automation interaction types: responding to agent-generated proposals, supervisory control of agents, and monitoring only. There is an overall trend in the data pointing towards a beneficial effect of transparency. However, the data reveals variations in Situation Awareness, mental workload, and operator performance for specific tasks, agent-types, and level of integration of transparency information in primary task displays.

**Conclusion:** Our data suggests a promising effect of automation transparency on Situation Awareness and operator performance, without the cost of added mental workload, for instances where humans respond to agent-generated proposals and where humans have a supervisory role.

**Application:** Strategies to improve human performance when interacting with intelligent agents should focus on allowing humans to see into its information processing stages, considering the integration of information in existing Human Machine Interface solutions.

**Keywords:** PRISMA, human-automation interaction, automation transparency, information disclosure, seeing into

---

Address correspondence to Koen van de Merwe, Group Research and Development, DNV, Veritasveien 1, Høvik, Oslo 1363, Norway; e-mail: [koen.van.de.merwe@dnv.com](mailto:koen.van.de.merwe@dnv.com)

## HUMAN FACTORS

2024, Vol. 66(1) 180–208

DOI:10.1177/00187208221077804

Article reuse guidelines: [sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)



Copyright © 2022, The Author(s).

## INTRODUCTION

The human factors community has long had an interest in understanding the interactions between humans and automation, that is, the tasks executed by a machine agent of a function previously performed by a human (Parasuraman & Riley, 1997; Rasmussen, 1983). Central topics of research include understanding the benefits and concerns of replacing humans with automation (e.g., Bainbridge, 1983; Strauch, 2018), the need for appropriate design of automation (Norman, 1990), the effect of automation failures on human take-over responses (Endsley & Kiris, 1995), factors pertaining to automation use, disuse, and misuse (Parasuraman & Riley, 1997), human performance in taking over from automation (Eriksson & Stanton, 2017; Hergeth et al., 2017; Weaver & DeLucia, 2020), and the consequences of levels of automation on Situation Awareness (SA), mental workload, and operator performance (Endsley & Kaber, 1999; Jamieson & Skraaning, 2020; Onnasch et al., 2014). Combined, these studies culminate to the notion of an automation conundrum (Endsley, 2017), which is the problem that the more reliable and robust automation becomes, the less likely it is that a human supervisor will notice critical information and will be able to effectively intervene when required. This problem may be exacerbated with advanced automation or intelligent agents able to function independently, but still require human supervision. Considering the rapidly developing and ubiquitous presence of technology in our society, there is an urgent and continuous need of research into understanding and enhancing interactions between humans and automation such that collaboration and performance can be supported (Hancock et al., 2013; O'Neill et al., 2020; Warden et al., 2019).

## Automation and Agents

The terms “automation” and “agent” are used interchangeably in the literature. For example, Lee and See define automation as “technology that actively selects data, transforms information, makes decisions and controls processes” (2004, p. 1). Parasuraman and Riley define automation as “the execution by a machine agent (usually a computer) of a function that was previously carried out by a human” (1997, p. 231). Rao and Georgeff (1995) describe a rational agent as one having certain “mental attitudes of Belief, Desires and Intention (BDI), representing, respectively, the information, motivational, and deliberative states of the agent” (1995, p. 1). In AI, the term “intelligent agent” refers to an autonomous entity having goal-directed behavior in an environment using observation through sensors and execution actions through actuators (Russell & Norvig, 2022). Examples of the application of agents can be seen in the automotive industry (Society of Automotive Engineers, 2021), healthcare (Coronato et al., 2020; Loftus et al., 2020), unmanned aerial vehicles (UAV) (Hocraffer & Nam, 2017), manufacturing (Elghoneimy & Gruver, 2012), and recent development towards maritime autonomous surface ships (IMO, 2018). Even though agents can be very sophisticated and can perform certain task with a high degree of independence, they often require some form of human supervision in case of failures or unforeseen situations. However, human supervision of such agents may pose challenges as AI behavior and reasoning can be difficult or even impossible to understand for humans (Doshi-Velez & Kim, 2017; Lipton, 2017). Still, to enable interaction between humans and agents, a system component capable of handling human-machine interactions is typically deployed, that is, the Human Machine Interface (HMI). The HMI supports human-machine interactions by providing relevant feedback to support SA and by allowing for appropriate input commands to support action execution.

Norman (1990) has previously advocated the use of appropriate feedback when interacting with automation, arguing that the problem with

keeping humans in the loop is not necessarily automation itself, but the lack of adequate information provided to them. Likewise, Christoffersen and Woods (Christoffersen & Woods, 2002) have discussed the need for systems to be observable to humans to enhance human-agent collaboration. That is, providing feedback to the operator in terms of its changes to the agent’s current state (events) allows for anticipatory reasoning (future states) and for quick detection of abnormalities through pattern recognition. Also, Lee and See (2004) argued for a number of elements that should be conveyed to the user, such as showing the automation’s purpose, past performance, and its processes and algorithms. In addition, intermediate internal process results should be shown that are understandable to the operator in a simplified way. Also, the purpose, design basis, and range of application should be conveyed that relate to the user’s goals. Supplying this information to the operator would result in appropriate reliance and trust in the automation. Hence, when humans interact with agents, the HMI can be used to convey the agent’s state, its modes, and limitations, and provide understandability and predictability regarding its current actions and future actions, that is, providing “transparency” to its user (Endsley, 2017).

## Transparency

There are two common interpretations of agent transparency found in the literature: “seeing-through” and “seeing-into” (Ososky et al., 2014; Sheridan & Verplank, 1978; Skraaning et al., 2020). The “seeing-through” interpretation states that automation should be designed in such a way as to appear invisible to its user. For example, in teleoperation using robots, transparent automation, for example, through means of low latency devices, effective feedback mechanisms, and immersive HMIs, allows an operator to perceive and manipulate the environment as if there was no automation in between. In this case, the automation is purposefully made invisible to the user allowing for enhanced awareness and “presence” of the remote environment. Conversely, the “seeing-into” interpretation aims to make the automation or agent visible to the human to allow for



enhanced understanding of the agent itself. In this case, the agent is made transparent, or better: “apparent” (Sheridan & Verplank, 1978; Skraaning et al., 2020), to its user by purposefully conveying what it is doing, why it is doing it, and what it will do next. In this perspective, transparency is an HMI design principle applied to the technology, based on the notion that information from and about the agent is directly observable to the user. In this paper, we will adopt the “seeing-into” perspective when referring to transparency.

Transparency information should allow for a user to “see into” the agent and better understand its inner processes, thereby enhancing the user’s ability to assess the agent’s performance and knowing when to manually take-over or not. Conversely, a lack of “transparency” (Endsley et al., 2003), “observability” (Christoffersen & Woods, 2002), “interpretability, explainability and predictability” (Hepworth et al., 2020), or “affordance” (Chen et al., 2014) of the agent may make it difficult for an operator to grasp what it is doing, why it is doing it, and what it is going to do next. This, in turn, may lead to poor decision making regarding when to use (and when not use) automation (Beck et al., 2007; Endsley & Kiris, 1995; Parasuraman & Riley, 1997). As such, exposing the inner workings of the automation to its human supervisor should, at least theoretically, enhance the operator’s performance.

### Transparency and Human Performance

Recent publications have explored evidence regarding automation transparency, that is, “seeing-into.” Bhaskara et al. (2020) identified and compared the dominant transparency models in the contemporary literature: Human-Robot Transparency Model (Lyons, 2013); Situation-Awareness Agent-based Transparency model (SAT; Chen et al., 2014). For these models, the authors reviewed five experimental studies that implemented transparency across a range of tasks and domains. Results from key human factors variables, including operator performance, SA, and mental workload indicated that there is emerging evidence regarding accurate use of automation with increased transparency, potential evidence for its

effect on SA and a potential cost in terms of mental workload, as measured through pupil diameter in one study (Wright et al., 2017). However, results were not consistent in terms of the correlation between the degree of transparency and performance variables. In other words, more transparency did not consistently produce improved operator performance outcomes. Hence, the effect of transparency may be dependent on other factors such as context and information type.

In a similar review, Rajabiyazdi and Jamieson (2020) reviewed the experimental evidence for four transparency models: Human-Robot Transparency Model (Lyons, 2013); (Dynamic) Situation-Awareness Agent-based Transparency model (SAT; Chen et al., 2014, DSAT; Chen et al., 2018); and the Coactive System Model based on Observability, Predictability, and Directability (Johnson et al., 2014). Five experimental studies were reviewed for their empirical evidence, of which two studies overlapped with Bhaskara et al. (2020). The authors concluded that the validation efforts for the transparency models have been largely incomplete or have provided inconclusive evidence. For example, there were differences among the studies in how the SAT model was interpreted and operationalized, that is, what level of transparency relates to which type of information, potentially leading to differences in outcomes. Also, even though some of the studies were based on the same theoretical model and applied in a similar context, they yielded inconsistent human performance outcomes in terms of SA, workload, and operator performance, amongst others. Nevertheless, considering the continuous development of advanced automation, the authors concluded that there is an ongoing and increasing need to further understand the means with which to convey its inner workings to the operators and assess its effect on human factors variables.

### This Study

This review aims to expand on the evidence base for automation transparency and operator performance by focusing on a broader body of literature beyond those studies discussed in the reviews mentioned earlier. This is to be achieved

by taking the original concept of transparency as the starting point for the review regardless of the transparency model. As the concept of “seeing-into” transparency is about conveying the inner workings of the automation to provide understandability and predictability about its actions, a broader scope may reveal additional insights not captured by model-specific studies (Bhaskara et al., 2020; Rajabiyazdi & Jamieson, 2020). This approach may uncover other studies not included in the abovementioned reviews that nevertheless provide evidence for the relationship between transparency and central human factors variables: SA, mental workload, and operator performance. These variables were chosen because information disclosure to reveal the inner workings of an agent is closely linked to the operator’s mental picture of the agent’s present and future state. As such, if the agent can convey to the user which information it is presently processing, how it is processing it, and what its future state will be, this would suggest that this information would have a positive effect on operator SA (Endsley, 1988, 1995). However, because transparent automation provides “understandability and predictability of actions” to a human operator (Endsley, 2017; Endsley et al., 2003), the HMI between the agent and the operator is often manipulated to allow for this. As mental workload concerns the allocation of limited internal resources in meeting external demands (Hancock et al., 2021), adding information increases the amount of information required to build and maintain SA, potentially requiring additional cognitive effort (Chen et al., 2014, 2018; Helldin et al., 2014). On the other hand, it may also be reasoned that assessing the performance of an agent is facilitated when information about the agent is made directly available to the user compared to when it is not (Chen et al., 2018). As such, the consequences of transparency information for mental workload may be mediated by other factors than amount of information only, for example, display design (Li et al., 2020; Vicente, 2002). Nevertheless, as transparent automation should allow an operator to better assess the agent’s performance, that is, its reliability, predictability, and ability (Lee & See, 2004), it should also improve the operator’s ability to perceive, comprehend and project the performance of the agent and thereby deciding whether to use the automation or not

(Beck et al., 2007; Parasuraman & Riley, 1997). This potential “free lunch” (Wickens, 2018), that is, the ability of transparency to alleviate some of the effect of the automation conundrum without reducing automation’s benefit, warrants a further and systematic focus.

## METHOD

This study uses the Preferred Reporting Items for Systematic review and Meta-Analysis protocol (PRISMA) as a basis for the systematic literature review (SLR; Moher et al., 2009, 2015). The PRISMA protocol provides a pre-defined and structured methodological approach to literature reviews including its data gathering, analysis, and reporting. Using a pre-defined approach reduces the potential for bias and enhances clarity, auditability, replicability, and transparency of the review (Booth et al., 2016). In brief, the PRISMA protocol uses a three-step approach starting with searching for relevant literature in relevant databases using a specified search string where the literature data is screened based on a pre-defined set of eligibility criteria. Second, an in-depth assessment is performed based on a review of the full texts generating a final dataset of literature. And finally, this dataset is analyzed as part of the qualitative data analysis.

### Database Search and Data Screening

The following inclusion criteria were established for the initial screening of the literature sample. First, only peer-reviewed studies published between the 1st of January 2000 and the 5th of January 2021 (the sample date) were considered. Second, studies that describe transparency effects on operator performance using experimental studies as a data source were considered.

The following exclusion criteria were established for the initial screening. First, non-English articles, articles from outside the time-period, non-peer reviewed, or gray literature (i.e., white papers, books, technical reports, book chapters, posters), and articles that not explicitly address automation transparency in experimental studies.

For screening the full-text literature, the following inclusion criteria were used. First, this

SLR was interested in studies presenting primary data that compared degrees of implementation of transparency in terms of SA and/or mental workload and/or operator performance metrics. Second, studies were considered if they met all the following characteristics based on the PICOC criteria (Booth et al., 2016; Petticrew & Roberts, 2006):

- Population: Users in the safety critical domain
- Intervention: Application of transparency in automation design
- Comparison: Comparing degrees of transparency
- Outcomes: The studies reported on SA, and/or mental workload, and/or operator performance metrics as dependent variables
- Context: The studies reported on findings obtain from a simulated- (experimental) and/or operational environment

To obtain the dataset, relevant databases were chosen based on their publication scope within the domains of psychology, technology, and engineering. The chosen databases were Scopus (with ScienceDirect for the full-text journals), IEEE Xplore, and Web of Science and were sampled using a search string.

The search string contains three components: the object of interest (e.g., automation), its characteristics (e.g., transparency), and its effect on operators (e.g., behavioral indicators and psychological constructs). The search aimed to balance breadth and depth of the field, and therefore the search was based on keywords only. The following search string was used in each of the chosen databases:

(Autom\* OR Autonom\* OR Robot OR Machine OR Agent)

AND

(Transparen\* OR Observab\* OR Explainab\* OR Afford\*)

AND

(“Operator performance” OR “Human performance” OR “Situation Awareness” OR Workload OR Effectiv\*)

Figure 1 provides the process and results of the database search. The search resulted in

a combined sample of 1714 articles of which there were 139 duplicates. Based on the sample of 1575 papers, the initial screening was performed based on the eligibility criteria described above. This consisted of a review of the titles and abstracts against the criteria. When in doubt, the paper was kept for full-text review. This resulted in a reduced sample of 59 articles for full-text review.

### Full-Text Review

The full-text review was performed by the first author based on the full-text eligibility criteria. A subset of 25 full-text papers out of the 59 papers were reviewed independently by the other authors. The results from this independent review of papers were cross verified with the results of the first author in a workshop. Any disagreements were resolved, and reasons for exclusion were noted. Of the full-text sample of 59 papers, 42 papers were excluded with reasons based on the pre-defined criteria (see Figure 1). As such, a final dataset of 17 full-text articles remained for inclusion in the qualitative analysis: 11 journal articles and six conference papers.

### Data Extraction and Analysis

Data from each individual study from the final dataset was extracted including the domain in which transparency was studied, the sample size, which (if any) transparency model was used, the Human-Automation Interaction type (HAI), how transparency was operationalized, and the comparisons that were made in the experimental study (see Table 1). For each of the studies the results were extracted, including SA effects of using the automation in the study, the effect on mental workload, and the behavioral/performance measures employed in the study (see Table 2).

## RESULTS

There are multiple ways in which the data in Tables 1 and 2 can be organized and interpreted depending on specific research needs. For our analysis, we have chosen to organize the data according to human-automation interaction type. For readers interested in looking into other

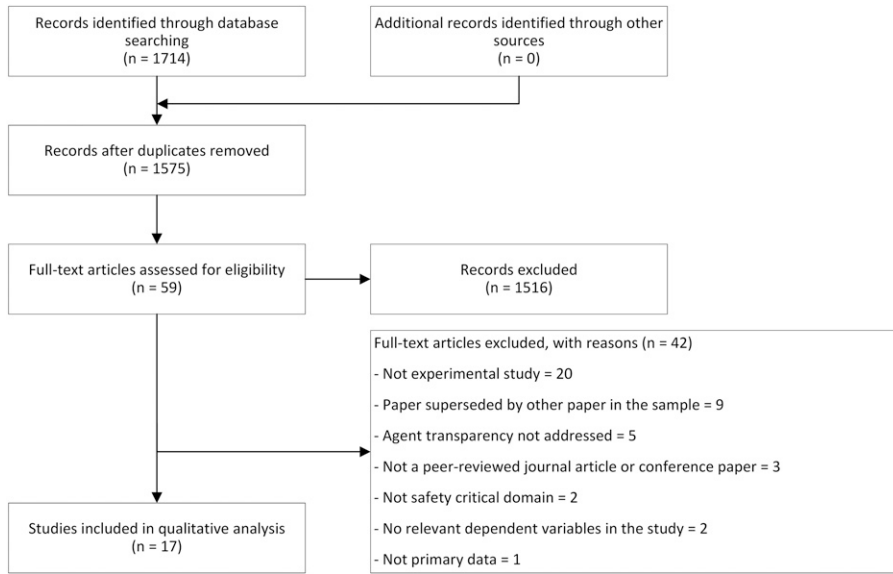


Figure 1. Flow diagram of the study selection based on the PRISMA protocol.

relations in the dataset, the tables are made available as [Supplemental Material](#) on the journal's Web site.

### Mapping Out How Transparency Has Been Studied

Table 1 describes the characteristics of the individual papers from the data sample. Each characteristic is discussed below.

**Research Domains.** The domain which had most focus on transparency research is the military (53%), with studies primarily focusing on UAV operations and ground troops support, and one study focused on the interactions with an automated pilot flying with a human in formation (wingman). Two (12%) studies were performed in the automotive domain in relation to autonomous vehicles. The other domains in which automation transparency was researched were civil defense (12%), civil aviation (12%), nuclear (6%), and robotics (6%).

**Transparency Models.** Eight studies (47%) used the SAT model (Chen et al., 2014) as a basis for the design of the automation. The studies that employed this model typically used

the various levels described by the model to develop user interfaces that provide users with relevant transparency information. For example, Selkowitz et al. (2017) developed a user interface showing an autonomous squad member's current resource levels (Level 1), prioritizations when following the squad (Level 2), consequences on future resource levels (Level 3), and the uncertainties related to this information. The other studies from the sample that used the SAT model have developed interfaces based on a similar approach (Bhaskara et al., 2021; Guznov et al., 2020; Mercado et al., 2016; Roth et al., 2020; Selkowitz et al., 2015; Stowers et al., 2020; Wright et al., 2020).

One study (6%) used Lyons' Human-Robot Transparency model (2013). Lyons describes the need for sharing information from the automation to the human (robot-to-human factors), as well as from the human to the automation (robot-of-human factors). Hence, Lyons' transparency model focuses on the requirements to the automation's information provision to the user, as well as the automation's capability to understand the human. Pokam et al. (2019) applied this model to develop the interface for an automated driving solution showing the conditions for when autonomous mode was available,

TABLE 1. Characteristics of the Studies

Reference	Domain, N	Model	HAI Type	Operationalization of Transparency and Comparisons
<a href="#">Mercado et al. (2016)</a>	Military UAV 30 Non-SME	SAT model: The visualization of the agent's current action and plans (Level 1), its reasoning and constraints (Level 2), and its projected outcomes and uncertainties (Level 3).	Respond to proposals: Monitor and control multiple unmanned vehicles (land, sea, air vehicles) and evaluate proposed plans (A or B) by an intelligent agent based on speed, coverage, and capability.	Level 1: Basic plan information provided by indicating which unmanned vehicles were in use and which paths they used. Level 1+2: Level 1 plus the agent's reasoning and rationale behind recommending the plans was provided via a text box and sprocket graphic. Level 1+2+3: Level 1+2 plus projection of uncertainty information related to a successful outcome. Uncertainty was presented through the opacity of vehicle icons, road colors, sprocket graphic wedges, and bullet points in the text box
<a href="#">Roth et al. (2020)</a>	Military UAV 10 Non-SME	SAT model: The visualization of the agent's current action and plans (Level 1), its reasoning and constraints (Level 2), and its projected outcomes and uncertainties (Level 3).	Respond to proposals: Perform mission planning and system management of a manned-unmanned teaming operation (manned helicopter + unmanned aerial vehicle). Execute a helicopter transport mission with take-off, transit, and landing. In mission planning, the participants had to evaluate the validity of the planning proposals performed by the agent and find violations. In system management, the participants performed the role as pilot-flying in the helicopter. Participants were tasked with monitoring and evaluating the agent's assistance. Flight control were not part of the tasks.	Low transparency: Level 1 information only. For mission planning the automation's goal, settings, and level of automation were displayed. For system management an "Adopted Tasks"-list was shown. High transparency: Level 1+2+3 information only. For mission planning the automation's goal, settings, and level of automation (Level 1), symbols representing the events that justified an intervention (Level 2), a timeline presenting the temporal outcomes projected by the agent (Level 3) were shown. For system management an "Adopted Tasks"-list (Level 1), "Critical Events"-list, "Neglected Tasks"-list, and "Current Load"-indicator (Level 2), and "To Do Tasks"-list and a timeline presenting the predicted future workload (Level 3) were shown.
<a href="#">Stowers et al. (2020)</a>	Military UxV 53 Non-SME	SAT model: The visualization of the agent's current action and plans (Level 1), its reasoning and constraints (Level 2), and its projected outcomes and uncertainties (Level 3).	Respond to proposals: Monitor and control multiple unmanned vehicles (land, sea, air vehicles) and evaluate proposed plans (A or B) by an intelligent agent based on speed, coverage, and capability. See also <a href="#">Mercado et al. (2016)</a> in this table.	Level 1+2: Level 1 and 2 information was displayed through the size of the unmanned vehicles' icons with larger icons depicting the faster unmanned vehicles. Level 1+2+3: Level 1+2 plus level 3 information displayed by an icon attached to the unmanned vehicles indicating the time it was from its goal location.

(Continued)

TABLE 1. (Continued)

Reference	Domain, N	Model	HAI Type	Operationalization of Transparency and Comparisons
Bhaskara et al. (2021)	Civil UxV 176 Non-SME	SAT model: The visualization of the agent's current action and plans (Level 1), its reasoning and constraints (Level 2), and its projected outcomes and uncertainties (Level 3).	Respond to proposals: Perform and complete unmanned vehicle control missions by selecting the most appropriate plan (A or B) against mission attributes. Participants were assisted by automation that based its decision on a formula taking into account time to search area, search time, and fuel consumption.	Level 1+2+3+U: Level 1+2+3 plus uncertainty information displayed through changes in opacity of the unmanned vehicles' icons. Level 1: The automation evaluated each unmanned vehicle's capabilities against the weighted mission attributes to determine and display the most suitable plan. No copy of the automation's formulae was provided. Level 1+2: As per Level 1. In addition, participants were informed of the automation's formulae and had a hard copy of these. Level 1+2+3: As per Level 1+2. In addition, participants received a visualization of the relative capability projection of the unmanned vehicles associated with Plans A and B (blue shaded bars presented on the interface). Transparency 1: No support provided. The Solution Space Diagram was turned off. Transparency 2: Support provided by showing heading bands, indicating unsafe heading regions. Transparency 3: Support provided by showing triangle-shaped conflict areas, indicating unsafe regions in speed, and heading.
Görtzlehner et al. (2014)	Air Traffic Control 12 Non-SME	Non-specific: Visual representation of a specific automated resolution advisory within the solution space for air traffic (i.e., go or no-go areas in speed and heading for an aircraft).	Respond to proposals: Ensure conflict free traffic in a free-flight Air Traffic Control scenario. Respond to resolution advisories by the automation by accepting or rejecting. Rate the agreement with the advisory.	Transparency 1: No support provided. The Solution Space Diagram was turned off. Transparency 2: Support provided by showing heading bands, indicating unsafe heading regions. Transparency 3: Support provided by showing triangle-shaped conflict areas, indicating unsafe regions in speed, and heading.
Sadler et al. (2016)	Flight planning 12 SME	Non-specific: The provision of the rationale behind automatically derived decision recommendation in three levels of transparency: Baseline, value, and logic.	Respond to proposals: Land aircraft on a landing site based on recommendations from an Autonomous Constrained Flight Planner.	Baseline: No explanation for how the automation arrived at its recommendation was provided. Value: Baseline plus the calculated success probability that drove the diversion recommendation was provided. Logic: Logic plus an additional explanation detailing the link between the probabilities and the information used to derive the recommendations was provided.
Guznov et al. (2020)	Robotics 88 Non-SME	SAT model: The visualization of the agent's current action and plans (Level 1), its reasoning	Supervise automation: Monitor and control a robot through an environment. Based on	Level 2: The robot's text message informed the participant of its actions and the reasoning

(Continued)

TABLE 1. (Continued)

Reference	Domain, N	Model	HAI Type	Operationalization of Transparency and Comparisons
Chen et al. (2014)	Military UAV 43 Non-SME	and constraints (Level 2), and its projected outcomes and uncertainties (Level 3).	a video feed and messages from the robot, participants could intervene if the robot was believed to be off the path.	processes behind them. For example, when the robot would approach a turn, it would report: "I see an obstacle on the right, so I'm turning left." Level 3: Level 2 plus its future states. For example, when the robot approached a turn, one of the multiple messages it reported was: "I see an obstacle on the left, so I will turn right in order to avoid collision." Non-transparent HMI: The unmanned aerial vehicle did not provide information regarding changes to its current and projected flight path. Transparent HMI: The unmanned aerial vehicle provided visual information regarding changes to its flight path. Minimal: The autonomous squad member provided a short, three-word description of the problem it encountered. Contextual: The autonomous squad member provided the decision it was requesting and a small amount of information about the situation to help the user make the decision. Constant: In addition to the above, the autonomous squad member provided the user with a constant stream of information. Limited transparency: No communication mechanism in place to allow the sharing of status between the unmanned aerial vehicles and participant. Increased transparency: A message dialogue box was in place to allow communication to be established between the unmanned aerial vehicles and the participant.
Sanders et al. (2014)	Ground troops support 73 Non-SME	Non-specific: Visualization of the implications of the level of automation on the unmanned aerial vehicle's autonomy capability to the operator through symbols and colors. Non-specific: The amount of information that the system provides to the user about its internal operations (i.e., explaining why it behaves as it does).	Supervise automation: Supervise a group of four unmanned aerial vehicles with various functional levels of automation in their search zones. Perform a search operation in the zones. Supervise automation: Control a soldier to find civilians and mark their location on a map. Also, assist an autonomous squad member (i.e., robot) by responding to questions (i.e., make navigational decisions).	Traditional: The human system interface did not provide explicit visual or verbal feedback about automation responsibilities, capabilities, goals, activities, and/or effects. Operator had to infer
Chen et al. (2015)	Military UAV 36 Non-SME	Non-specific: Visualization of autonomy and functional capabilities of the agent through a (textual) natural language dialogue.	Supervise automation: Monitor and control four unmanned aerial vehicles (in various levels of automation) in transit mode, respond to a hazardous event with an avoidance maneuver (hazard avoidance mode), and perform a search activity (search mode).	Increased transparency: A message dialogue box was in place to allow communication to be established between the unmanned aerial vehicles and the participant.
Skraaning & Jamieson (2021)	Nuclear 16 SME	Non-specific: The observability of responsibilities, capabilities, goals, activities, and/or effects of automation in the human system interface.	Supervise automation: Control a simulated nuclear power plant and deal with minor to major system upsets including taking corrective action.	Traditional: The human system interface did not provide explicit visual or verbal feedback about automation responsibilities, capabilities, goals, activities, and/or effects. Operator had to infer

(Continued)

TABLE 1. (Continued)

Reference	Domain, N	Model	HAI Type	Operationalization of Transparency and Comparisons
				<p>these attributes of automation from changes in the plant process as reflected in a conventional human system interface for supervisory control.</p> <p><i>Transparent:</i> The interface provided explicit verbal and visual information about automation activities to the operators. Key automatic devices on a large-screen overview display, dedicated displays for detailed monitoring, display of tracking of automation sequences, and verbal feedback about the activity of automatic systems from automation were used. Verbal feedback was provided each time automatic devices started or failed. The information was limited to behavioral feedback from automation (what happened) announced repeatedly.</p> <p>As above, and verbal feedback was provided only when the executive automatic programs started, or automatic devices failed. The feedback from automation was both behavioral (what happened) and diagnostic (why it happened) and was announced once.</p> <p><i>Traditional:</i> No explicit information regarding the information about the activities of the plant-wide procedure automation. Operators had to derive this from the process events and changes to system states.</p> <p><i>Transparent:</i> Explicit information about automation status and actions provided through a dedicated overview display showing the automation's progress. Color coding was used to depict status including a detailed list of procedural steps. The detailed automation display depicted historical and ongoing automation activities.</p> <p><i>Level 1:</i> The autonomous squad member provided its current location, its route, and its current resources.</p>
	Nuclear 18 SME	As above.	As above.	
	Nuclear 27 SME	As above.	As above and monitor automatic scripts responsible for performing a cold start of the plant to 50% reactor power. When the automatic procedure paused (e.g., due to a technical failure) intervene by assuming manual control, restarting automation or shutdown.	
Selkowitz et al. (2015)	Search & Rescue 45 Non-SME	SAT model: The visualization of the agent's current action and plans (Level 1), its reasoning	Monitor automation: Monitor an autonomous squad member as it moves through an urban area. The autonomous squad member moved	

(Continued)



TABLE 1. (Continued)

Reference	Domain, N	Model	HAI Type	Operationalization of Transparency and Comparisons
		and constraints (Level 2), and its projected outcomes and uncertainties (Level 3).	on its own accord taking into account obstacles and dangers. Its route was revealed from waypoint to waypoint by a navigation line.	Level 1+2: Level 1 plus the autonomous squad member provided its affordances and hazards it encounters during its task execution. Level 1+2+3: Level 1+2 plus the autonomous squad member provided its environmental constraints with their associated uncertainties and its predicted resources at the end of the mission.
Selkowitz et al. (2017)	Military UAV 60 Non-SME	SAT model: The visualization of the agent's current action and plans (Level 1), its reasoning and constraints (Level 2), its projected outcomes (Level 3), and its uncertainties (Level 3+U).	Monitor automation: Monitor a simulated environment for threats and mark these on the display. Monitor an autonomous squad member's display for its decisions.	Level 1: The autonomous squad member provided its current resource levels, its understanding of the squad's current status, its understanding of the environment around it, the current highest influence on its motivator (e.g., time), and its current action/position. Level 1+2: Level 1 plus the autonomous squad member's reasoning behind its current action (e.g., a clock icon). Level 1+2+3: Level 1+2 plus the projected outcomes of its current actions and reasoning (e.g., projected time displayed). Level 1+2+3+U: Level 1+2+3 plus the associated uncertainty for the information (e.g., time uncertainty and icon color).
Wright et al. (2020)	Ground troops support 56 Non-SME	SAT model: The visualization of the agent's current action and plans (Level 1), its reasoning and constraints (Level 2), and its projected outcomes and uncertainties (Level 3).	Monitor automation: Monitor a video feed from a soldier team and evaluate the autonomous squad member in correctly identifying and responding to events the soldier squad encountered. Detect threats in the surrounding environment and identify events the squad encountered.	Surface-level information: The interface contained at-a-glance information regarding the autonomous squad member's current actions (Level 1), the reasons for its action (Level 2), and the results of its actions (Level 3). In-depth information: Surface-level information plus additional information depicting the underlying factors that led to each specific information in the surface-level module.
Pokam et al. (2019)	Autonomous vehicles 45 SME	Human-Robot Transparency model: The information that a robot needs to convey to a human (its intentions, tasks, analysis, and environmental constraints) and vice versa (how tasks are distributed and awareness of the human state).	Monitor automation: Monitor the behavior of an autonomous vehicle from the driver's seat under different transparency conditions. No intervention was required.	HMI 1: The vehicle displayed no additional information about its autonomy. HMI 2: The vehicle displayed its acquired information and its action execution. HMI 3: The vehicle displayed its acquired information, its analysis, and its action execution.

(Continued)

TABLE 1. (Continued)

Reference	Domain, N	Model	HAI Type	Operationalization of Transparency and Comparisons
Du et al. (2019)	Autonomous vehicles 32 SME	Non-specific: The provision of explanations to justify why an action was or was not taken by the automation.	Monitor automation: Monitor the behavior of an autonomous vehicle from the driver's seat. There was no need to take over control of the vehicle.	HMI 4: The vehicle displayed its information acquired, its analysis, and its decision making. HMI 5: The vehicle displayed its information acquired, its analysis, its decision making, and its action execution. No explanation: The autonomous vehicle provided no explanation about its actions. After explanation: The autonomous vehicle presented an explanation within 1 s after actions had been taken. Before explanation: The autonomous vehicle provided an explanation 7 seconds prior to its action.
Panganiban et al. (2020)	Military aviation 40 Non-SME	Non-specific: The intentional design of a system to communicate its capabilities and current state to support human-machine teaming.	Monitor automation: Fly a fighter aircraft and attack a missile site whilst cloaking the plane from detection from a nearby defensive Surface-to-Air-Missile. The autonomous wingman supported the pilot through performing surveillance (neutral condition) or additional cloaking of the pilot against the Surface-to-Air-Missile (benevolent condition).	Neutral: The wingman provided only information on its immediate task activities, with no additional transparency into its intentions. Benevolent: The wingman communicated its intention to support the human and to correct its errors, signaling its awareness of the human partner's expectations.

understand the actions by the vehicle, why a given maneuver was carried out and showing what the automation perceived in order to understand its analyses and decisions.

Eight studies (47%) were not limited to a single transparency model but used various transparency sources as the basis for automation design. For example, [Skraaning and Jamieson \(2021\)](#) stated that the automation displays that were used in their nuclear control room study were designed “with the transparency principle in mind” (2021, p. 380). They define transparency as “the design principle that the responsibilities, capabilities, goals, activities and/or effects of automation should be directly observable in the [Human System Interface]” and refer to [Norman \(1990\)](#), [Christoffersen and Woods \(2002\)](#), [Johnson et al. \(2014\)](#), and others as their inspirational sources. Likewise, [Du et al. \(2019\)](#) focused on explanations provided by the automation as a means to expose users “to the inner workings or logic used by the automated system” (2019, p. 429). Also, [Chen et al. \(2014, 2015\)](#), [Sanders et al. \(2014\)](#), [Göritzlehner et al. \(2014\)](#), [Sadler et al. \(2016\)](#), and [Panganiban et al. \(2020\)](#) have used various transparency sources as inspiration for their automation design.

The dataset did not include experimental studies for the Coactive System Model based on Observability, Predictability, and Directability ([Johnson et al., 2014](#)).

### Human-Automation Interaction Type.

In six studies (35%) participants were tasked with *responding to proposals* provided by the automation. [Mercado et al. \(2016\)](#) and [Stowers et al. \(2020\)](#) performed similar experiments where participants were asked to monitor and control multiple unmanned vehicles (land, air, and sea; UxV) in a base-defense task. An intelligent agent generated proposals on how to best defend the base based on speed, coverage, and capabilities of the unmanned vehicles. The participants were required to choose the most optimal plan. Similarly, [Bhaskara et al. \(2021\)](#) required participants to select the best unmanned vehicle to perform a task. Participants were assisted by a system that provided two plans with regards to which unmanned vehicle was

most capable based on its time to reach a search area, search time needed and fuel consumption. The participants were asked to check the accuracy of the proposals against a set of criteria and choose the best one. [Roth et al. \(2020\)](#) also required participants to check the validity of the agent’s proposals and find violations to previously given constraints for an UAV mission. Participants in the experiment by [Göritzlehner et al. \(2014\)](#) took the role of an air traffic controller and were tasked with ensuring conflict-free traffic in a simulated airspace. The automation provided advisories to resolve conflict situations, and the participants were required to either accept or reject these based on their perception of the situation. Finally, [Sadler et al. \(2016\)](#) used airline pilots in the role of enhanced ground operators that were required to land aircraft at alternative landing sites when their primary destination was unavailable. An Autonomous Constrained Flight Planner was used to provide the operators with recommended diversions which they were asked to check for its validity.

Five studies (30%) required participants to *supervise the automation* (i.e., monitor, respond to, and manually operate) when required. In three separate experiments, [Skraaning and Jamieson \(2021\)](#) required licensed operators to monitor, control, and operate a nuclear plant under different levels of transparency and types of automation. For the condition where transparency was applied at the component level, the operators were required to operate the plant and respond to system upsets. For the condition with plant-wide automation, the operators were required to monitor an agent in operating the plant by itself but intervene in case of interrupts. On a much more limited scale, [Guznov et al. \(2020\)](#) asked their participants to monitor and operate a simple robot in navigating a track. Each time the robot went off-track, the participants were required to intervene and put the robot back on track. Similarly, [Sanders et al. \(2014\)](#) requested their participants to maneuver a soldier through an environment whilst looking for civilians and mark these on a map. In addition, they were asked to assist the soldier’s robotic teammate in responding to navigational requests (i.e., deciding where to go in ambiguous situations). Finally, [Chen et al. \(2014, 2015\)](#) tasked their

participants with monitoring UAV and perform manual avoidance maneuvers as a result of hazardous situations in the environment. In addition, a search task was performed where participants marked items of interest at the target area.

In six studies (35%) participants were tasked with only *monitoring the automation*. In [Selkowitz et al. \(2015\)](#), [Selkowitz et al. \(2017\)](#), and [Wright et al. \(2020\)](#), participants were required to monitor an autonomous squad member through a video feed where their primary task was to monitor the actions and information provided by the autonomous squad member. As a secondary task they were asked to monitor the environment for threats. No manual intervention was required for the autonomous squad member. [Du et al. \(2019\)](#) and [Pokam et al. \(2019\)](#) required participants to monitor the behavior of a self-driving vehicle. No intervention was required by the participants irrespective of the scenario or the level of transparency applied. In the study by [Panganiban et al. \(2020\)](#), participants were supported by an automated wingman that was tasked with countering threats by enemy Surface-to-Air missiles. The participants were required to monitor the automation only for the level of support it provided for the mission and the way it communicated its support to the participant.

**Operationalizations of Transparency and Comparisons.** The design of transparent automation depends on the task, the context, and the domain in which the automation is applied. As such, what and how information is displayed to the user is affected by the specific domain in which transparency is applied and what tasks the agent and user are expected to perform. [Table 1](#) provides an overview of the various operationalizations in our sample. As illustration, [Selkowitz et al. \(2017\)](#) and [Wright et al. \(2020\)](#), using the same simulator test-bed, operationalized transparency through displaying icons and colors representing the agent's status (e.g., a battery indicator), its goals (e.g., a number within an icon representing a way-point on a map), its reasoning (e.g., a time indicator show this as its priority), its projected outcomes (e.g., a red box next to a clock icon indicating a loss of

time), and its uncertainty (e.g., a light red border around an event icon). The level of transparency was manipulated by showing more or less of this information per experimental run.

In terms of experimental comparisons, all studies employed a cumulative approach where transparency followed a continuum, that is, from less to more transparent automation. Subsequently, the experiments compared designs with varying levels of transparency and measured their effect on relevant dependent variables. As illustration, [Mercado et al. \(2016\)](#) used the SAT model to develop the user interface for unmanned vehicle operations and designed an experiment to assess the effect of each of the levels of transparency described by the model: Level 1 transparency provided only basic plan information, Level 2 transparency provided the automation's reasoning and rationale behind the recommendations, and Level 3 provided the automation's projections and uncertainties. Based on this experimental design, comparisons were made between the levels of transparency in terms of their effect on their dependent variables.

## Describing the Empirical Evidence

[Table 2](#) describes the empirical evidence from each of the studies.

*Automation Transparency and SA.* Situation Awareness was measured in nine out of 17 studies. The instruments that were used to measure the construct were Situation Awareness Global Assessment Technique (SAGAT), Situation Awareness Rating Technique (SART), a confidence in own SA measure, and a Process overview Measure. The results of the studies fell in two categories, improved SA, and no effect.

[Selkowitz et al. \(2017\)](#) reported an effect of transparency on SA in terms of improved L2 and L3 SA when monitoring an autonomous squad member navigating through an urban area. Adding affordances, hazards, environmental constraints, and uncertainties seems to help the operators in obtaining a better picture of the situation. Likewise, [Roth et al. \(2020\)](#) also found improved SA for tasks relating to mission planning and system management for a manned-

TABLE 2. Study Results

Reference	HAI Type	Situation Awareness	Effect	Mental Workload	Effect	Operator Performance	Effect
Mercado et al. (2016)	Respond to proposals			Scores on NASA-TLX	↔	Correct use of proposals	↑
				Mean eye fixation duration	↔	Correct rejection of proposals	↑
				Pupil diameter	↔	Response time to proposals	↔
				Saccadic amplitude	↔		
				Saccade duration	↔		
Roth et al. (2020)	Respond to proposals	Scores on SAGAT	↑	Scores on Bedford Mental	↔	Response time to proposals	↓
Stowers et al. (2020)	Respond to proposals	Scores on SART	↔	Workload scale	↔	Accuracy of decisions	↔
				Scores on NASA-TLX	↔	No. of correct responses	↑
Bhaskara et al. (2021)	Respond to proposals			Scores on NASA-TLX	↔	Response time	↑
						Acceptance of correct proposals	↑
				Secondary task performance (auditory recognition task)	↔	Rejection of incorrect proposals	↑
Göritzlehner et al. (2014)	Respond to proposals			0-100 scale	↔	Accuracy of automation use	↑
						Decision time	↓
Sadler et al. (2016)	Respond to proposals					Advisory accept/reject	↔
						Agreement with advisory	↔
Guznov et al. (2020)	Supervise automation					Separation conflicts	↓
						Separation violations	↔
						Verifications of plans	↓
						Exploring for alternatives	↔
						Agreement with plans	↔
Chen et al. (2014)	Supervise automation	Scores on SART	↔	Scores on NASA-TLX	↑	No. of correct responses	↔
		Scores on SAGAT	↑	Scores on NASA-TLX	↓	No. of correct rejections	↔
Sanders et al. (2014)	Supervise automation			NASA-TLX results not reported	?		
				DSSQ results not reported	?		

(Continued)

TABLE 2. (Continued)

Reference	HAI Type	Situation Awareness	Effect	Mental Workload	Effect	Operator Performance	Effect
Chen et al. (2015)	Supervise automation	Scores on SAGAT	↑	Scores on NASA-TLX	↓	Initial response times for UAV to proceed on-course	↔
		Scores on SART	↔	Scores on Perceived Task Complexity scale	↓	Event response times for UAV to avoid hazards	↓
Skraaning & Jamieson (2021)	Supervise automation	Scores on SART	↑	Scores on Perceived Task Complexity scale	↓	Success rate in finding items	↑
		Scores on Process Overview Measure	↔	Scores on Perceived Task Complexity scale	↓	Response time to events	↓
		Scores on SAGAT (L1 SA)	↔	Scores on Perceived Task Complexity scale	↓	Detecting deviations and performing verifications	↑
		Scores on SAGAT (L2 SA)	↑	Scores on Perceived Task Complexity scale	↓	Achieving main goals	↑
Selkowitz et al. (2015)	Monitor automation	Scores on SAGAT (L3 SA)	↑	Scores on Perceived Task Complexity scale	↔	Response time to events	↓
		Confidence in own L1 SA	↔	Scores on NASA-TLX	↔	Detecting deviations and performing verifications	↑
		Confidence in own L2 SA	↑	Scores on NASA-TLX	↔	Achieving main goals	↓
Selkowitz et al. (2017)	Monitor automation	Confidence in own L3 SA	↑	Scores on NASA-TLX	↔	Detecting deviations and performing verifications	↔
		Confidence in own L1 SA	↔	Scores on NASA-TLX	↔	Achieving main goals	↔
		Confidence in own L2 SA	↑	Scores on NASA-TLX	↔	Self-rated task performance	↔

(Continued)

TABLE 2. (Continued)

Reference	HAI Type	Situation Awareness	Effect	Mental Workload	Effect	Operator Performance	Effect
Wright et al. (2020)	Monitor automation	Scores on SAGAT	↔	Scores on NASA-TLX	↔	Detecting targets Time to identify and assess events	↔
Pokam et al. (2019)	Monitor automation	Scores on SAGAT	↔				
Du et al. (2019)	Monitor automation			Scores on NASA-TLX	↔		
Panganiban et al. (2020)	Monitor automation			Scores on NASA-TLX	↓		

Key: ↑ indicates improvement/increase, ↔ indicates that no effect was found, ↓ indicates decline/reduction, ? indicates that the outcome is unspecified, and blank cells indicate the variable was not measured.

unmanned helicopter teaming operation. When adding symbols that represented the agent's reasoning (e.g., the events that justified an intervention), projected outcomes, and uncertainties, improved SA was reported by the participants when using the SAGAT method. Furthermore, [Skraaning and Jamieson \(2021\)](#) found improved SA in their second experiment using SART. Here, nuclear control room operators were given explicit verbal and visual information about automation activities. When verbal feedback was both behavioral and diagnostic (i.e., what equipment failed and why) in contrast to only behavioral or no verbal feedback, operators reported improved SA. Finally, [Chen et al. \(2014, 2015\)](#) reported evidence for the effect of transparency on SA when providing the UAV's capability to the user. When the UAV provided visual information regarding the changes to its flight path, that is, a presentation of the agent's previous, present, or projected flight path, SA improved. Likewise, when the operator was able to communicate with the agent using a natural language dialogue (e.g., a message reading "Please control my altitude and speed, I can follow my flight path") participants reported improved SA.

Some studies found that transparency did not positively affect SA. For example, in their first and third experiment, [Skraaning and Jamieson \(2021\)](#) did not find differences between a traditional and transparent HMI, as measured by SART, when the feedback by the system was limited to behavioral information only (i.e., what equipment failed and not why; first experiment). Furthermore, no effect was found, as measured by the Process Overview Measure, when plant-wide agent-like automation was introduced (third experiment), including detailed information regarding the agent's historical and ongoing activities. Likewise, [Wright et al. \(2020\)](#) did not find differences in SA between their transparency manipulations. For an autonomous squad member task, they provided in-depth information on the HMI indicating the underlying factors as to why specific surface-level information was presented. However, adding in-depth information did not lead to better SA amongst the participants. Also, [Guznov et al. \(2020\)](#) did not find evidence for

improved SA. Participants were tasked with monitoring and controlling a robot through an environment. When the robot communicated its perceptions and actions only (e.g., "I see an obstacle on the right, so I am turning left"), no differences for SA were found compared to when the robot also included its projected future outcomes (i.e., "I see an obstacle on the left, so I will turn right *in order to avoid a collision*"; emphasis added). Moreover, [Pokam et al. \(2019\)](#) found similar results when participants were asked to monitor the actions of an autonomous vehicle. Finally, [Selkowitz et al. \(2017, 2015\)](#) and [Roth et al. \(2020\)](#) did not find an effect of transparency on SA when monitoring an autonomous squad member or when evaluating proposals for an UAV mission (when using the SART method), respectively.

*Automation Transparency and Mental Workload.* Mental workload was measured in two ways: objectively (eye-movements, secondary task performance) and subjectively (NASA-Task Load Index (NASA-TLX), Perceived Task Complexity scale, Dundee Stress State Questionnaire (DSSQ), a 0-100 scale, and the Bedford Mental Workload scale).

First, [Selkowitz et al. \(2017\)](#) used eye-tracking and found that the duration of fixations on the displays increased as a function of transparency. This experiment introduced additional symbology on the display (e.g., motivators for the autonomous squad member, predicted outcomes, uncertainty information), and it appears that adding this information led to increased dwell time on the display. Second, [Guznov et al. \(2020\)](#) also found an increase in mental workload, measured by using the NASA-TLX, as a result of transparency. They found that the primary driver was a significant difference in the "physical workload" sub-scale of the NASA-TLX. The authors concluded that an increase in the amount of text led to additional reading load, which may have been interpreted by the participants as increased physical demand.

Some studies either did not record a difference in mental workload as a function of transparency or recorded a reduction. For experiment 3 in Skraaning and Jamieson's study



(2021), the authors developed two additional displays with which the plant-wide agent-oriented automation could be monitored. These displays showed for example, which part of an automated sequence was being executed, if there were any alerts, the list of actions to be taken, historical and ongoing activities. This information, presented on separate displays, was available in addition to the information in the non-transparent condition. Nevertheless, the operators reported no differences in terms of mental workload. Similarly, Mercado et al. (2016) developed a user interface for evaluating proposed plans for monitoring and controlling multiple unmanned vehicles. Transparency information consisted of text boxes, sprocket graphics, opacity of icons, colors, and bullet points. Mental workload was measured using the NASA-TLX and a range of eye-tracking measures. No differences were found between the transparency levels in terms of mental workload.

Skraaning and Jamieson (2021) measured mental workload using the Perceived Task Complexity scale on nuclear control room operators. In experiment 1 and 2, transparency was introduced at the component-level. That is, transparent automation in this experiment was operationalized in terms of visual presentation of automatic activity next to the components on the displays, dedicated displays for detailed monitoring of controllers and programs and verbal and visual information about the automation's activities. Providing this additional information resulted in lower perceived mental workload by the participants. For a different task and setting, Panganiban et al. (2020) also found reduced mental workload when an automated wingman communicated its intentions to support the human and to correct the human's errors. According to this result, knowing that there is an automated teammate present to support one's actions results in reduced mental effort on the participants' own tasks. Finally, Chen et al. (2014, 2015) found that providing UAV capability information to the participants resulted in lower workload, as measured by the NASA-TLX.

One study reported that two workload measures were used (NASA-TLX and DSSQ)

but did not report the results (Sanders et al., 2014).

*Automation Transparency and Operator Performance.* Operator performance was measured in two ways: objectively (task and response accuracy, response time, detection of events, goal achievement), and subjectively (self-rated task performance). In addition, some studies used more general measures of behavior: verification activities upon receiving advice by the automation, exploration of alternatives and agreement to proposals.

Participants in Mercado et al.'s (2016) study reported improvements in correct acceptances (i.e., an acceptance of a proposal when it was correct) and correct rejections (i.e., a rejection of a proposal when it was incorrect) with increased transparency. Stowers et al. (2020), in a similar study, replicated these results by showing higher percentages of correct responses on proposed plans. Bhaskara et al. (2021) also provided evidence that increased automation transparency leads to improved decision accuracy on proposals provided by an automated agent ("the Recommender"). In terms of response time, Skraaning and Jamieson (2021) found reduced response times for component-level transparency. Transparency focused display design led to faster responses to minor and major systems upsets. In addition, there is some supporting evidence of the positive effect of transparency in terms of faster initiation of evasive maneuvers of UAVs to hazardous events (Chen et al., 2015) and in the time needed to evaluate the validity of planning proposals in a joint helicopter and UAV mission (Roth et al., 2020). Finally, Skraaning and Jamieson (2021) found that increased transparency at the component-level increased detection of process deviations (e.g., alarms) and goal achievement (e.g., successfully executing all steps in a start-up sequence). This result was corroborated by Chen et al. (2015) who found improved goal achievement in terms of items of interest found when performing an UAV search task.

Wright et al. (2020) found little evidence for the effect of transparency on the accuracy of detecting targets in the surrounding environment when monitoring an autonomous squad member. Similarly, Skraaning and Jamieson (2021)

reported that when operators were tasked with monitoring plant-wide, agent-like automation performing a cold start-up of a nuclear power plant (experiment 3), no clear benefits were reported when responding to system upsets. In the low transparency condition, the operators had to derive the state of the plant based on process parameters only. In the high transparency condition, the operators had dedicated displays available to show the agent's plant-wide activities. Still, no differences were found in terms of goal achievement and self-rated task performance. Finally, Mercado et al. (2016) found little evidence for the effect of transparency on response time to planning proposals in an unmanned vehicle military perimeter defense task.

Stowers et al. (2020) reported slower response times to proposed plans made by an intelligent agent when monitoring and controlling multiple unmanned vehicles. Also, Skraaning and Jamieson (2021) reported a reduction in detecting process deviations and in performing verifications of system information when dedicated displays were used showing the activities of the agent-oriented plant-wide automation.

## DISCUSSION

Whilst performing the review, variations in terms of scientific rigor between the studies became apparent. As noted earlier by Bhaskara et al. (2020), experimental studies regarding automation transparency have primarily used non-subject matter experts as participants. It is important that research set in the context of applied-, and safety critical domains, translates to the actual domain it purports to be relevant for. Twelve studies (71%) in our review reported using non-subject matter experts as participants in their experiments. Typically, these studies used university students or laypeople from the local community who were compensated for their effort in terms of course credits or financial payment. Only four studies (23%) used subject matter experts. Skraaning and Jamieson (2021) used licensed nuclear control room operators, Sadler et al. (2016) used airline pilots, and Pokam et al. (2019) and Du et al. (2019) used automobile drivers. One study did not mention

what type of participants were used (Guznov et al., 2020). Furthermore, there were large differences in sample sizes between the studies, from 10 to 176 participants. Although more challenging to perform, especially with typically difficult to recruit subject matter experts, studies with larger sample sizes do provide more robust statistical results (Funder & Ozer, 2019; Schönbrodt & Perugini, 2013). This means that the results from some of the studies with relatively small sample sizes should be treated with some caution. Moreover, different studies used different techniques to measure the constructs of SA, mental workload, and operator performance. For example, Roth et al. (2020) measured SA using the SAGAT and the SART method. The SAGAT found a positive effect of automation transparency and the SART did not. Possibly, the SART is more an indicator of confidence in one's own SA than of SA itself (Endsley, 1988). Nevertheless, comparing results that were based on different measurement methods can be challenging because of differences in sensitivities and reliabilities of these methods. In this study, we have focused on the experimental outcomes, as opposed to the methodological analysis and discussion of the various measurement tools implemented across the reviewed studies.

## Transparency, SA, Mental Workload, and Operator Performance

In the introduction, we alluded to the relationship between SA, mental workload, and operator performance by stating that transparency might alleviate some of the negative effects of automation for SA and operator performance, albeit at the potential cost of mental effort. Increased mental workload arises in cases where multiple tasks are competing for the same resources and task requirements exceed mental capacity (Wickens et al., 2013). When the resources required to build and maintain SA overlap with resources required for task performance, mental capacity may be exceeded which may affect SA and subsequently performance (Endsley, 1995; Endsley & Garland, 2000).

For the relationship between transparency and SA, there are some indications for the increased

disclosure of information by an agent and improved SA. Notwithstanding information clutter due to poor interface design (Kim et al., 2011), transparency information may make it easier for an operator to assess what the agent is doing and why by making relevant information readily available to the operator (Endsley, 2017). The studies by Chen et al. (2014, 2015) show overall improvements in SA, the study by Selkowitz et al. (2017) found improved SA for Level 2 and 3 SA (but did not report overall results), and the results from Skraaning and Jamieson (2021) and Roth et al. (2020) show some mixed results depending on how transparency was implemented and which measurement instrument was used respectively. Still, having the information directly perceivable on the interface could reduce the burden on mental processing capacity by reducing the need for keeping multiple information elements in working memory (van Doorn et al., 2021).

For mental workload, only two studies in our sample showed an increase, the remaining studies found either no effect or found a reduction. Interestingly, one of these was measured using eye-tracking and showed an increase in fixation durations, indicating increased information processing with increased transparency (Selkowitz et al., 2017). However, the other study that also measured fixation duration using eye-tracking did not find any significant result (Mercado et al., 2016). Nevertheless, most of our studies seem to indicate that increasing transparency did not affect the participants to such an extent that it led to information overload. Conversely, adding transparency information did not consistently lead to reductions in workload either. In all the experiments in our sample, participants were required to assess the performance of an agent, either through evaluating decision options, intervening in an ongoing process, performing manual activities, or monitoring the agent. One may expect that assessing the performance of an agent, and its associated cognitive effort, would be facilitated when the information about the agent was made available to the user compared to when it was not. Only the studies by Chen et al. (2014, 2015), Skraaning and Jamieson (2021; experiment 1 and 2), and Panganiban et al. (2020) found this effect.

For operator performance, it was expected that performance would improve with increased transparency. There are some indications that transparent automation leads to better discrimination between correct use of proposals and correct rejections in those studies in which this was measured (Bhaskara et al., 2021; Mercado et al., 2016; Stowers et al., 2020). Although some studies did not report any differences in decision accuracy (Guznov et al., 2020; Roth et al., 2020; Wright et al., 2020), there were also no studies that reported a deterioration. This seems to indicate there is some merit in applying transparency principles for tasks where automation usage decisions need to be made. We also found a moderately positive relationship between transparency and response times to events, that is, system prompts or proposals (Bhaskara et al., 2021; Chen et al., 2015; Roth et al., 2020; Skraaning & Jamieson, 2021; experiment 1 and 2).

As good SA, without requiring excessive mental effort, increases the probability for good operator performance (Endsley, 1995; van de Merwe et al., 2012; van Doorn et al., 2021), we assessed those studies in which SA, mental workload, and performance were measured together. Five of the 17 studies measured these three variables in conjunction. For three of these studies, we see neutral or improved SA scores, neutral or reduced workload together with improved response times (Chen et al., 2015; Roth et al., 2020, for SAGAT only; Skraaning & Jamieson, 2021, experiment 1 and 2), goal achievement (Chen et al., 2015; Skraaning & Jamieson, 2021, experiment 1 and 2), and detecting process deviations and performing verifications (Skraaning & Jamieson, 2021, experiment 1 and 2). Guznov et al. (2020) found increased workload scores but no effects for SA and the number of correct responses and correct rejections. Wright et al. (2020) did not find any effect for SA, mental workload, and performance on detecting target and time to identify and assess events. Finally, Skraaning and Jamieson (2021, experiment 3) found no effects for SA, workload, and operator performance, and even reduced performance for detecting and verifying events, when participants were using plant-wide, agent-like automation

where transparency information was made available through dedicated displays. This indicates that the benefits of transparency may be affected by agent type, but also how transparency information is made available to operators. The absence of transparency benefits for this study may be attributed to operator capacity issues in simultaneously monitoring the process and the agent, in addition to the attention-grabbing effect of the (separate) transparency interface.

### Transparency and Human-Automation Interaction types

In the results section, we identified that the studies from the sample can be categorized in three distinct human-automation interaction types; that is, participants were tasked with responding to proposals, supervising automation, and monitoring automation. Knowing that the automation interaction paradigm influences system oversight and intervention (Endsley, 2017), a better understanding for which types of tasks transparent automation would provide the most benefit may provide valuable insights to engineers developing transparent designs. The allocation of roles between humans and automation, as well as the automation's level of sophistication, is important determinants in this relationship (Endsley & Kaber, 1999). For example, automation may provide decision support to a human in direct control (Manzey et al., 2012; Metzger & Parasuraman, 2005; Rieger & Manzey, 2020), or automation may take the form of an intelligent agent that works largely independent, but with the human in a supervisory role, ready to intervene when needed (Borst et al., 2017). As the function distribution between agents and humans dictate the distribution of tasks, this in turn dictates the human information needs to perform these tasks. Different function distributions therefore lead to different operator tasks, which lead to different information (i.e., transparency) needs (van Doorn et al., 2017). Hence, how functions and tasks are distributed between humans and agents is therefore an important element in understanding the relationship between transparency and human performance. As designing collaborative human-agent systems entails making choices with regards to “who does

what with what information,” it is important to understand how the purported transparency benefits translate across different human-agent interaction types.

For the studies where participants responded to proposals, the data in Table 2 suggests a relation between transparency and improved correct evaluation of proposals without affecting workload. None of the studies found changes to workload as measured through rating scales, secondary task performance and eye-tracking. For operator performance, the studies by Mercado et al. (2016), Stowers et al. (2020), and Bhaskara et al. (2021) found improved use of correct proposals and improved correct rejection of incorrect proposals. Only Roth et al. (2020) did not find an effect. In terms of response times to proposals however, the picture is less clear. Stowers et al. (2020) found an increase, Mercado et al. (2016) and Bhaskara et al. (2021) found a reduction, and Roth et al. (2020) found no differences. Furthermore, the study by Göritzlehner et al. (2014) showed a reduction in number of separation conflicts, and Sadler et al. (2016) found a reduction in the pilots' verification of the proposed plans. Unfortunately, there is insufficient data to conclude on SA, as for this interaction type, only one study measured the construct and it showed contrasting outcomes (Roth et al., 2020). Still, the results indicate that transparency has performance benefits for this interaction type without adding workload.

For supervising automation, a moderately positive relation was seen between transparency, improved SA, reduced workload, and improved operator performance. The studies by Chen et al. (2014, 2015) and Skraaning and Jamieson (2021) found no change to SA (experiment 1) or improved SA (experiment 2), reduced mental workload and improved response times, ability to detect events and goal achievement. Skraaning and Jamieson's third experiment did not replicate these findings. Here they found no differences for SA and workload and a decrease in operator performance. Only Guznov et al. (2020) found an increase in mental workload, when supervising a robot through a maze, with no differences for SA and operator performance reported. Nevertheless, also for this interaction

type, the results tend towards performance benefits with limited effect on mental workload.

For monitoring automation, the relationship between the HF variables and transparency is somewhat unclear, however. Only the study by [Wright et al. \(2020\)](#) measured the three constructs for this interaction type but found no differences. None of the other studies captured operator performance, so the data for this construct is rather limited for this interaction type. This is understandable as the participants were not required to do anything other than monitoring. For SA and mental workload, there are some indications for improved SA at the cost of visual processing in monitoring an autonomous squad member ([Selkowitz et al., 2017](#)). Reduced mental workload was found when collaborating with an automated wingman ([Panganiban et al., 2020](#)). However, the study by [Selkowitz et al. \(2015\)](#) did not find any relationship between transparency, SA, and mental workload. Also, the rest of the (individual) study results did not indicate a relationship with transparency for this interaction type.

### Practical Implications

The results from these studies are relevant for whenever systems are developed where humans are required to work with agents to achieve a common goal. However, the use of agents may provide challenges for human interaction as agents using neural networks are known to be opaque and difficult to interpret ([Sanneman & Shah, 2020](#)). As such, although these agents are powerful and flexible in their application, they may come at the cost of interpretability and understandability for a human operator ([Doshi-Velez & Kim, 2017](#)). For an agent to be transparent to a human, it would imply the system should provide understandability and predictability of its actions ([Endsley, 2017](#)); that is, see into the information processing stages of the agent such that its outcomes are understandable to its user ([Hepworth et al., 2020](#)).

Research into strategic conformance, that is, the extent of compatibility between human and agent information processing, seems to suggest improved automation acceptance rates and reduced response times to system proposals.

This suggests that systems that “make sense” to the human are easier to supervise as it alleviates some of the workload related to trying to understand what the system is doing and why ([Hilburn et al., 2014](#); [Westin et al., 2015](#)). To this end, the well-known human information processing model by [Parasuraman et al. \(2000\)](#) may be used as a basis for developing transparent displays to achieve increased compatibility between human and agent information processing. For example, an agent operating in a real-world setting, for example, an anti-collision tool for autonomous maritime navigation ([Statheros et al., 2008](#)), may be able to detect and integrate information based on a suite of sensors, perform object classification, create a representation of its environment, plan actions considering relevant constraints, and execute appropriate actions ([DNV, 2018](#)). Making these stages understandable to a user could imply graphically depicting relevant information it has detected (e.g., using bounding boxes around objects), classify this information (e.g., the type of objects and their characteristics), represent their relevance (e.g., in terms of potential collision risks), and indicate potential and highlight optimal solutions based on a cost function (e.g., fuel, time, safety), possibly including uncertainties. Finally, these solutions could be presented as a choice to the operator or automatically executed, depending on the agent’s capabilities.

Adding information to the HMI of an intelligent agent that is compatible with human information processing strategies, provided adequate display design, should imply improved human decision making without adding mental workload. Furthermore, when the human is required to monitor, respond to, and manually operate a function (i.e., supervise), improvements in operator performance, mental workload, and SA can be anticipated when the agent presents the underlying information for its decision making and (proposed) actions. However, careful consideration should be given to how transparency is practically implemented and integrated in existing HMI solutions (i.e., primary task displays) such that operator performance is sufficiently supported ([National Academies of Sciences, 2021](#)).

## Limitations

Performing a systematic literature review requires making choices regarding the specificity of the study and its replicability. This review appreciates that there may be research on transparency that is published in non-scientific channels (e.g., reports from research institutes), studies that have researched the construct without using the terms in our search string or have published in channels not captured in our databases. This means that, although this study has aimed to perform a broad review of the literature, it is likely there is research on transparency that is not covered by our SLR. However, for the sake of replicability, this paper has chosen to make the sampling and analysis of the data as objective and open as possible. This means that no additional research was added to the sample that was not found in the search results.

The search spanned a range of over 20 years of research on automation transparency. However, results revealed that experimental studies focusing specifically on automation transparency is a recent topic of interest, at least in terms of number of hits in our data sample. The oldest study in the sample that meets our eligibility criteria was published in 2014. A possible explanation for this are the strict eligibility criteria used. This SLR only includes experimental studies on the topic of transparency, in safety critical domains, for which a limited set of human factors variables were measured. As such, articles that discuss transparency conceptually (e.g., presenting models, frameworks, definitions), that were outside the safety critical domain (e.g., caregiving robots, explainability of algorithms for loan application decisions), that presented secondary data (e.g., reviews), or that did not measure SA, mental workload, or operator performance (e.g., only usability, acceptance, or trust), were not considered. A broader set of eligibility criteria could have resulted in additional data, albeit at the cost of specificity. As such, although transparency has been discussed in publications before (e.g., [Endsley et al., 2003](#); [Meister, 1999](#)), there seems to be a relationship between the time the construct was formalized into theoretical models ([Chen et al., 2014](#);

[Johnson et al., 2014](#); [Lyons, 2013](#)) and the experimental studies these generated.

Finally, differences in statistical reporting made comparison between the studies challenging. Some studies provided full statistical disclosure in terms of  $p$ -values, effect sizes, confidence intervals, sample sizes, and graphical representations of the data, whereas other studies provided very limited to no statistical information. As such, this made comparison across the studies challenging and prohibited a more rigorous quantitative comparison.

## Conclusions and Future Work

This review mapped the “seeing-into” transparency literature to address the relationships between transparency and central human factors variables. The data provided indications that human performance is enhanced when a function keeps the operator in the loop by presenting proposals and stating the reasons for them. Furthermore, when the human is required to monitor, respond to, and manually operate a function (i.e., supervise), improvements in operator performance, mental workload, and SA can be anticipated when the agent presents the underlying information for its decision making and (proposed) actions. Adding this information to the HMI of an intelligent agent, provided adequate display design, should imply improved human performance without adding mental workload. However, there are subtle variations in SA, mental workload, and operator performance for specific tasks, agent-types, levels of information disclosure, and level of integration of transparency information in primary task displays. Future work should focus on understanding which information types are valuable in conveying agent transparency information (see also [National Academies of Sciences, 2021](#)). As a starting point, the information processing model by [Parasuraman et al. \(2000\)](#) was suggested to allow increased compatibility between the agent’s and human’s information processing ([Hilburn et al., 2014](#); [Westin et al., 2015](#)). However, the degree to which this model is suitable as tool to set agent transparency requirements should be investigated further.

This study focused on the relationship between agent transparency and operator performance in

combination with two primary psychological constructs SA and mental workload. However, automation transparency is frequently researched in relation to other variables, such as trust in automation (Chen et al., 2018; Lee & See, 2004; Oliveira et al., 2020; Schmidt et al., 2020). Trust is the attitude that an agent (or automation) will help achieve a goal in uncertain and vulnerable circumstances (Lee & See, 2004) and is an important element in determining automation usage. Operators may not use automation when they don't trust it, even though it is reliable. Conversely, high trust in automation may lead to overreliance, that is, using automation when it should not be (Parasuraman & Riley, 1997). Transparent automation should help an operator to calibrate their trust in automation such that automation is only used when it should be (Lee & See, 2004). Although this study did not include trust as part of its inclusion criteria, the relevance of trust in relation to automation transparency is not disputed. Likewise, additional variables such as cognitive processes, system design features, environmental features, and emergent characteristics involved in automation oversight and interaction performance (Endsley, 2017) were similarly excluded. As such, this study focused on the key human variables SA and mental workload in addition to operator performance. Future studies could focus on establishing comprehensive evidence regarding additional key variables in agent transparency and assess their scientific consensus and practical merit.

### ACKNOWLEDGMENTS

The authors would like to thank Øystein Engelhardtson, DNV Group R&D, and the anonymous reviewers for their significant contributions and reflections to the work.

### KEY POINTS

- Automation transparency is a design principle aimed at enabling operators to understand what automation is doing, why it is doing it, and what it

will do next (i.e., “seeing-into” transparency). It is a means to address the challenges related to human performance in interacting with systems that have high degrees of automation.

- This study systematically gathered and assessed empirical evidence for the relationship between automation transparency, Situation Awareness, mental workload, and operator performance using the PRISMA method.
- There are three transparency models that dominate the transparency research, however, there is a significant body of research investigating transparency without conforming to any particular model. The human-automation interaction types employed in the research can be categorized into responding to agent-generated proposals, supervisory control, and monitoring only. All studies investigated the effect of the amount and type of transparency information on performance variables.
- The empirical results from the studies point towards a promising effect of automation transparency on operator performance, without the cost of added mental workload, for instances where humans respond to agent-generated proposals and where humans have a supervisory role.
- There are subtle variations in SA, mental workload, and operator performance for specific tasks, agent-types, levels of information disclosure, and level of integration of transparency information in primary task displays. There were limited findings for our variables when humans were monitoring automation only.
- The outcomes have practical implications for the design of systems where humans and automation work towards a common goal.

### ORCID iDs

Koen van de Merwe  <https://orcid.org/0000-0002-0168-872X>

Steven Mallam  <https://orcid.org/0000-0003-1713-2977>

Salman Nazir  <https://orcid.org/0000-0002-2058-6147>

## SUPPLEMENTAL MATERIAL

The online supplemental material is available with the manuscript on the *HF* website.

## REFERENCES

- Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6), 775–779. [https://doi.org/10.1016/0005-1098\(83\)90046-8](https://doi.org/10.1016/0005-1098(83)90046-8)
- Beck, H. P., Dzindolet, M. T., & Pierce, L. G. (2007). Automation usage decisions: controlling intent and appraisal errors in a target detection task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 49(3), 429–437. <https://doi.org/10.1518/001872007X200076>
- Bhaskara, A., Duong, L., Brooks, J., Li, R., McInerney, R., Skinner, M., Pongracic, H., & Loft, S. (2021). Effect of automation transparency in the management of multiple unmanned vehicles. *Applied Ergonomics*, 90, 103243. <https://doi.org/10.1016/j.apergo.2020.103243>.
- Bhaskara, A., Skinner, M., & Loft, S. (2020). Agent transparency: A review of current theory and evidence. *IEEE Transactions on Human-Machine Systems*, 50(3), 215–224. <https://doi.org/10.1109/THMS.2020.2965529>
- Booth, A., Sutton, A., & Papaioannou, D. (2016). *Systematic approaches to a successful literature review* (2nd ed.). Sage.
- Borst, C., Bijsterbosch, V. A., van Paassen, M. M., & Mulder, M. (2017). Ecological interface design: Supporting fault diagnosis of automated advice in a supervisory air traffic control task. *Cognition, Technology & Work*, 19(4), 545–560. <https://doi.org/10.1007/s10111-017-0438-y>
- Chen, J. Y. C., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, 19(3), 259–282. <https://doi.org/10.1080/1463922X.2017.1315750>
- Chen, J. Y. C., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. J. (2014). *Situation awareness-based agent transparency (ARL-TR-6905)*. U.S. Army Research Laboratory. <https://doi.org/10.21236/ADA600351>.
- Chen, T., Campbell, D., Gonzalez, L. F., & Coppin, G. (2015). Increasing Autonomy Transparency through capability communication in multiple heterogeneous UAV management. In 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October, 2015, pp. 2434–2439. <https://doi.org/10.1109/IROS.2015.7353707>
- Chen, T., Campbell, D. A., Gonzalez, F., & Coppin, G. (2014). The effect of autonomy transparency in human-robot interactions: A preliminary study on operator cognitive workload and situation awareness in multiple heterogeneous UAV management. In Proceedings of Australasian Conference on Robotics and Automation 2014. <https://www.araa.asn.au/acra/acra2014/papers/pap166.pdf>
- Christoffersen, K., & Woods, D. (2002). 1. How to make automated systems team players. In *Advances in human performance and cognitive engineering research* (Vol. 2, pp. 1–12). Emerald Group Publishing Limited. [https://doi.org/10.1016/S1479-3601\(02\)02003-9](https://doi.org/10.1016/S1479-3601(02)02003-9)
- Coronato, A., Naeem, M., De Pietro, G., & Paragliola, G. (2020). Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine*, 109, 101964. <https://doi.org/10.1016/j.artmed.2020.101964>.
- DNV (2018). *DNVGL-CG-0264: Autonomous and remotely operated ships*. <http://rules.dnvgl.com/docs/pdf/dnvgl/cg/2018-09/dnvgl-cg-0264.pdf>
- Doshi-Velez, F., & Kim, B. (2017). *Towards A rigorous science of interpretable machine learning*. ArXiv:1702.08608 [Cs, Stat] <http://arxiv.org/abs/1702.08608>
- Du, N., Haspiel, J., Zhang, Q., Tilbury, D., Pradhan, A. K., Yang, X. J., & Robert, L. P., Jr. (2019). Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transportation Research Part C: Emerging Technologies*, 104, 428–442. <https://doi.org/10.1016/j.trc.2019.05.025>.
- Elghoneimy, E., & Gruver, W. A. (2012). Agent-based decision support and simulation for wood products manufacturing. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6), 1656–1668. <https://doi.org/10.1109/TSMCC.2012.2213809>
- Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. *Proceedings of the Human Factors Society Annual Meeting*, 32(2), 97–101. <https://doi.org/10.1177/154193128803200221>
- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1), 217–249. <https://doi.org/10.4324/9781315092898-13>
- Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 59(1), 5–27. <https://doi.org/10.1177/0018720816681350>
- Endsley, M. R., Bolté, B., & Jones, D. G. (2003). *Designing for situation awareness: An approach to user-centered design*. Taylor & Francis.
- Endsley, M. R., & Garland, D. J. (Eds.). (2000). *Situation awareness: Analysis and measurement*. Lawrence Erlbaum Associates.
- Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, 42(3), 462–492. <https://doi.org/10.1080/001401399185595>
- Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(2), 381–394. <https://doi.org/10.1518/001872095779064555>
- Eriksson, A., & Stanton, N. A. (2017). Takeover time in highly automated vehicles: Noncritical transitions to and from manual control. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 59(4), 689–705. <https://doi.org/10.1177/0018720816685832>
- Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, 2(2), 156–168. <https://doi.org/10.1177/2515245919847202>
- Görtzlehner, R., Borst, C., Ellerbroek, J., Westin, C., van Paassen, M. M., & Mulder, M. (2014). Effects of transparency on the acceptance of automated resolution advisories. 2014 *IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 2965–2970). Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/SMC.2014.6974381>
- Guznov, S., Lyons, J., Pfahler, M., Heironimus, A., Woolley, M., Friedman, J., & Neimeier, A. (2020). Robot transparency and team orientation effects on human-robot teaming. *International Journal of Human-Computer Interaction*, 36(7), 650–660. <https://doi.org/10.1080/10447318.2019.1676519>
- Hancock, G. M., Longo, L., Young, M. S., & Hancock, P. A. (2021). Mental workload. *Handbook of human factors and ergonomics*



- (5th ed., pp. 203–226). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781119636113.ch7>
- Hancock, P. A., Jagacinski, R. J., Parasuraman, R., Wickens, C. D., Wilson, G. F., & Kaber, D. B. (2013). Human-automation interaction research: Past, present, and future. *Ergonomics in Design*, 21(2), 9–14. <https://doi.org/10.1177/1064804613477099>
- Heldlin, T., Ohlander, U., Falkman, G., & Riveiro, M. (2014). Transparency of Automated Combat Classification. In D. Harris (Ed.) *Engineering psychology and cognitive ergonomics*, (pp. 22–33). Springer International Publishing.
- Hepworth, A. J., Baxter, D. P., Hussein, A., Yaxley, K. J., Debie, E., & Abbass, H. A. (2020). Human-swarm-teaming transparency and trust architecture. *IEEE/CAA Journal of Automatica Sinica*, 8(7), 1–15. <https://doi.org/10.1109/JAS.2020.1003545>.
- Hergeth, S., Lorenz, L., & Kreams, J. F. (2017). Prior familiarization with takeover requests affects drivers' takeover performance and automation trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 59(3), 457–470. <https://doi.org/10.1177/0018720816678714>
- Hilburn, B., Westin, C., & Borst, C. (2014). Will controllers accept a machine that thinks like they think? The role of strategic conformance in decision aiding automation. *Air Traffic Control Quarterly*, 22(2), 115–136. <https://doi.org/10.2514/atcq.22.2.115>
- Hocraffer, A., & Nam, C. S. (2017). A meta-analysis of human-system interfaces in unmanned aerial vehicle (UAV) swarm management. *Applied Ergonomics*, 58, 66–80. <https://doi.org/10.1016/j.apergo.2016.05.011>.
- IMO (2018). *Maritime Safety Committee (MSC), 100th session, 3-7 December 2018*. International Maritime Organisation. <https://www.imo.org/en/MediaCentre/MeetingSummaries/Pages/MSC-100th-session.aspx>
- Jamieson, G. A., & Skraaning, G. (2020). The absence of degree of automation trade-offs in complex work settings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 62(4), 516–529. <https://doi.org/10.1177/0018720819842709>
- Johnson, M., Bradshaw, J. M., Feltoch, P. J., Jonker, C. M., van Riemsdijk, M. B., & Sierhuis, M. (2014). Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, 3(1), 43–69. <https://doi.org/10.5898/JHRI.3.1.Johnson>
- Kim, S.-H., Prinzel, L. J., Kaber, D. B., Alexander, A. L., Stelzer, E. M., Kaufmann, K., & Veil, T. (2011). Multidimensional measure of display clutter and pilot performance for advanced head-up display. *Aviation, Space, and Environmental Medicine*, 82(11), 1013–1022. <https://doi.org/10.3357/ASEM.3017.2011>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1), 50–80. [https://doi.org/10.1518/hfes.46.1.50\\_30392](https://doi.org/10.1518/hfes.46.1.50_30392)
- Li, W. C., Zakarija, M., Yu, C. S., & McCarty, P. (2020). Interface design on cabin pressurization system affecting pilot's situation awareness: The comparison between digital displays and pointed displays. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 30(2), 103–113. <https://doi.org/10.1002/hfm.20826>
- Lipton, Z. C. (2017). *The mythos of model interpretability*. ArXiv: 1606.03490 [Cs, Stat] <http://arxiv.org/abs/1606.03490>
- Loftus, T. J., Filiberto, A. C., Balch, J., Ayzengart, A. L., Tighe, P. J., Rashidi, P., Bihorac, A., & Upchurch, G. R. (2020). Intelligent, autonomous machines in surgery. *Journal of Surgical Research*, 253, 92–99. <https://doi.org/10.1016/j.jss.2020.03.046>
- Lyons, J. B. (2013). Being transparent about transparency. *Proceedings of the AAAI Spring Symposium*, 48–53. <https://www.aaai.org/ocs/index.php/SSS/SSS13/paper/download/5712/6000>
- Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, 6(1), 57–87. <https://doi.org/10.1177/1555343411433844>.
- Meister, D. (1999). *The history of human factors and ergonomics*. Lawrence Erlbaum Associates.
- Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human-agent teaming for multi-UxV management. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 58(3), 401–415. <https://doi.org/10.1177/0018720815621206>
- Metzger, U., & Parasuraman, R. (2005). Automation in future air traffic management: Effects of decision aid reliability on controller performance and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 47(1), 35–49. <https://doi.org/10.1518/0018720053653802>
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, T. P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine*, 6(7), 6. <https://doi.org/10.1371/journal.pmed.1000097>
- Moher, D., Shamseer, L., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. PRISMA-P Group (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, 4(1), 1. <https://doi.org/10.1186/2046-4053-4-1>
- National Academies of Sciences, Engineering and Medicine (2021). *Human-AI teaming: State of the art and research needs*. The National Academies Press. <https://doi.org/10.17226/26355>
- Norman, D. A. (1990). The “problem” with automation: Inappropriate feedback and interaction, not “over-automation”. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 327(1241), 585–593. <https://doi.org/10.1098/rstb.1990.0101>
- Oliveira, L., Burns, C., Luton, J., Iyer, S., & Birrell, S. (2020). The influence of system transparency on trust: Evaluating interfaces in a highly automated vehicle. *Transportation Research Part F: Traffic Psychology and Behaviour*, 72, 280–296. <https://doi.org/10.1016/j.trf.2020.06.001>
- O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2020). Human-autonomy teaming: A review and analysis of the empirical literature. *Human Factors*. Advance online publication <https://doi.org/10.1177/0018720820960865>.
- Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 56(3), 476–488. <https://doi.org/10.1177/0018720813501549>
- Osofsky, S., Sanders, T., Jentsch, F., Hancock, P., & Chen, J. Y. C. (2014). Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In R. E. Karlson, D. W. Gage, C. M. Shoemaker & G. R. Gerhart (Eds.), *Proceedings volume 9084: Unmanned systems technology XVI*. SPIE. <https://doi.org/10.1117/12.2050622>
- Panganiban, A. R., Matthews, G., & Long, M. D. (2020). Transparency in autonomous teammates: Intention to support as teaming information. *Journal of Cognitive Engineering and Decision Making*, 14(2), 174–190. <https://doi.org/10.1177/1555343419881563>

- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39(2), 230–253. <https://doi.org/10.1518/001872097778543886>
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3), 286–297. <https://doi.org/10.1109/3468.844354>
- Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Wiley-Blackwell.
- Pokam, R., Debernard, S., Chauvin, C., & Langlois, S. (2019). Principles of transparency for autonomous vehicles: First results of an experiment with an augmented reality human-machine interface. *Cognition, Technology & Work*, 21(4), 643–656. <https://doi.org/10.1007/s10111-019-00552-9>
- Rajabiyazdi, F., & Jamieson, G. A. (2020). A Review of Transparency (seeing-into) Models. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 302–308). IEEE. <https://doi.org/10.1109/SMC42975.2020.9282970>
- Rao, A. S., & Georgeff, M. P. (1995). BDI agents: From theory to practice. *Proceedings of the First International Conference on Multiagent Systems* (pp. 312–319). <https://www.aai.org/Papers/ICMAS/1995/ICMAS95-042.pdf>
- Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-13*(3), 257–266. <https://doi.org/10.1109/TSMC.1983.6313160>
- Rieger, T., & Manzey, D. (2020). Human performance consequences of automated decision aids: The impact of time pressure. *Human Factors*. Advance online publication. <https://doi.org/10.1177/0018720820965019>
- Roth, G., Schulte, A., Schmitt, F., & Brand, Y. (2020). Transparency for a workload-adaptive cognitive agent in a manned-unmanned teaming application. *IEEE Transactions on Human-Machine Systems*, 50(3), 225–233. <https://doi.org/10.1109/THMS.2019.2914667>
- Russell, S. J., & Norvig, P. (2022). *Artificial intelligence: A modern approach* (4th ed, global ed.). Pearson.
- Sadler, G., Battiste, H., Ho, N., Hoffmann, L., Johnson, W., Shively, R., Lyons, J., & Smith, D. (2016). Effects of transparency on pilot trust and agreement in the autonomous constrained flight planner. *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)* (pp. 1–9). IEEE. <https://doi.org/10.1109/DASC.2016.7777998>
- Sanders, T. L., Wixon, T., Schafer, K. E., Chen, J. Y. C., & Hancock, P. A. (2014). The influence of modality and transparency on trust in human-robot interaction. In 2014 IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), San Antonio, TX, USA, 3–6 March, 2014, pp. 156–159. <https://doi.org/10.1109/CogSIMA.2014.6816556>
- Sanneman, L., & Shah, J. A. (2020). A situation awareness-based framework for design and evaluation of explainable AI. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems* (pp. 94–110). Springer. [https://doi.org/10.1007/978-3-030-51924-7\\_6](https://doi.org/10.1007/978-3-030-51924-7_6)
- Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, 29(4), 260–278. <https://doi.org/10.1080/12460125.2020.1819094>
- Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, 47(5), 609–612. <https://doi.org/10.1016/j.jrp.2013.05.009>
- Selkowitz, A., Lakhmani, S., Chen, J. Y. C., & Boyce, M. (2015). The effects of agent transparency on human interaction with an autonomous robotic agent. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 59(1), 806–810. <https://doi.org/10.1177/1541931215591246>
- Selkowitz, A. R., Lakhmani, S. G., & Chen, J. Y. C. (2017). Using agent transparency to support situation awareness of the Autonomous Squad Member. *Cognitive Systems Research*, 46(194), 13–25. <https://doi.org/10.1016/j.cogsys.2017.02.003>
- Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators*. Defense Technical Information Center. <https://doi.org/10.21236/ADA057655>
- Skraaning, G., Jamieson, G., & Jeffrey, J. (2020). *Towards a deeper understanding of automation transparency in the operation of nuclear plants (INL/EXT-20-59469)*. U.S. Department of Energy. <https://doi.org/10.2172/1668828>
- Skraaning, G., & Jamieson, G. A. (2021). Human performance benefits of the automation transparency design principle: validation and variation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 63(3), 379–401. <https://doi.org/10.1177/0018720819887252>
- Society of Automotive Engineers (2021). *Taxonomy and definitions for terms Related to driving automation systems for on-road motor vehicles (J3016\_202104* (pp. 1–41). Society of Automotive Engineers. [https://doi.org/10.4271/J3016\\_202104](https://doi.org/10.4271/J3016_202104)
- Statheros, T., Howells, G., & Maier, K. M. (2008). Autonomous ship collision avoidance navigation concepts, technologies and techniques. *The Journal of Navigation*, 61(1), 129–142. <https://doi.org/10.1017/S037346330700447X>
- Stowers, K., Kasdaglis, N., Rupp, M. A., Newton, O. B., Chen, J. Y. C., & Barnes, M. J. (2020). The IMPACT of agent transparency on human performance. *IEEE Transactions on Human-Machine Systems*, 50(3), 245–253. <https://doi.org/10.1109/THMS.2020.2978041>
- Strauch, B. (2018). Ironies of automation: Still unresolved after all these years. *IEEE Transactions on Human-Machine Systems*, 48(5), 419–433. <https://doi.org/10.1109/THMS.2017.2732506>
- van de Merwe, K., Oprins, E., Eriksson, F., & van der Plaats, A. (2012). The influence of automation support on performance, workload, and situation awareness of air traffic controllers. *The International Journal of Aviation Psychology*, 22(2), 120–143. <https://doi.org/10.1080/10508414.2012.663241>
- van Doorn, E., Horváth, I., & Rusák, Z. (2021). Effects of coherent, integrated, and context-dependent adaptable user interfaces on operators' situation awareness, performance, and workload. *Cognition, Technology & Work*, 23(3), 403–418. <https://doi.org/10.1007/s10111-020-00642-z>
- van Doorn, E., Rusák, Z., & Horváth, I. (2017). A situation awareness analysis scheme to identify deficiencies of complex man-machine interactions. *International Journal of Information Technology and Management*, 16(1), 53–72. <https://doi.org/10.1504/IJITM.2017.080958>
- Vicente, K. J. (2002). Ecological interface design: Progress and challenges. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 44(1), 62–78. <https://doi.org/10.1518/0018720024494829>
- Warden, T., Carayon, P., Roth, E. M., Chen, J., Clancey, W. J., Hoffman, R., & Steinberg, M. L. (2019). The national academies board on human system integration (BOHSI) panel: Explainable AI, system transparency, and human machine teaming. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63(1), 631–635. <https://doi.org/10.1177/1071181319631100>

- Weaver, B. W., & DeLucia, P. R. (2020). A systematic review and meta-analysis of takeover performance during conditionally automated driving. *Human Factors*. Advance online publication. <https://doi.org/10.1177/0018720820976476>
- Westin, C., Borst, C., & Hilburn, B. (2015). Strategic conformance: Overcoming acceptance issues of decision aiding automation? *IEEE Transactions on Human-Machine Systems*, 46(1), 41–52. <https://doi.org/10.1109/THMS.2015.2482480>
- Wickens, C. (2018). Automation stages & levels, 20 years after. *Journal of Cognitive Engineering and Decision Making*, 12(1), 35–41. <https://doi.org/10.1177/1555343417727438>
- Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2013). *Engineering psychology and human performance* (4th ed.). Pearson.
- Wright, J. L., Chen, J. Y. C., Barnes, M. J., & Hancock, P. A. (2017). *Agent reasoning transparency: The influence of information level on automation induced complacency (ARL-TR-8044* (p. 214). US Army Research Laboratory. Human Research and Engineering Directorate. <https://apps.dtic.mil/dtic/tr/fulltext/u2/1035306.pdf>
- Wright, J. L., Chen, J. Y. C., & Lakhmani, S. G. (2020). Agent transparency and reliability in human–robot interaction: The influence on user confidence and perceived reliability. *IEEE Transactions on Human-Machine Systems*, 50(3), 254–263. <https://doi.org/10.1109/THMS.2019.2925717>

Koen van de Merwe is a principal researcher at DNV Group R&D at Høvik, Norway. He received his MSc in Cognitive Psychology in 2004 and an MSc in Industrial Ecology in 2006 from Leiden University, The Netherlands, and he is currently pursuing his PhD in Nautical Operations at the University of South-Eastern Norway.

Steven Mallam is an Associate Professor of Maritime Human Factors at the Faculty of Technology Natural Sciences and Maritime Sciences at The University of South-Eastern Norway. He received his PhD in Human Factors in 2016 from Chalmers University of Technology, Sweden.

Salman Nazir is a Professor in Training and Assessment at Department of Maritime Operations at the University of South-Eastern Norway. He received his PhD in Industrial Chemistry and Chemical Engineering from Politecnico di Milano, Italy, in 2014.

*Date received: August 23, 2021*

*Date accepted: January 17, 2022*



**Article 2**

van de Merwe, K., Mallam, S., Nazir, S., & Engelhardtsen, Ø. (2024). Supporting human supervision in autonomous collision avoidance through agent transparency. *Safety Science*, 169, 13.

<https://doi.org/10.1016/j.ssci.2023.106329>





# Supporting human supervision in autonomous collision avoidance through agent transparency

Koen van de Merwe<sup>a,b,\*</sup>, Steven Mallam<sup>b,c</sup>, Salman Nazir<sup>b</sup>, Øystein Engelhardtson<sup>a</sup>

<sup>a</sup> Group Research and Development, DNV, Veritasveien 1, 1363 Høvik, Norway

<sup>b</sup> Department for Maritime Operations, University of South-Eastern Norway, Raveien 215, 3184 Borre, Norway

<sup>c</sup> Fisheries & Marine Institute, Memorial University of Newfoundland, St. John's, NL A1B1T5, Canada

## ARTICLE INFO

### Keywords:

Transparency  
Autonomous shipping  
Collision avoidance  
Goal-directed task analysis  
Cognitive task analysis  
Human-Machine Interaction  
Safety  
Situation Awareness  
Autonomous agents

## ABSTRACT

Ongoing trends in society point towards the adoption of intelligent agents across safety critical industries. In the maritime domain, artificially intelligent agents may soon be capable of autonomously performing collision and grounding avoidance (CAGA); a task traditionally performed by humans. Consequently, the role of humans is anticipated to change from those performing collision avoidance to those supervising an agent performing collision avoidance. One of the key concerns with regards to human factors is avoiding the out-of-the-loop performance problem where humans lose situation awareness (SA) and become susceptible to misinterpreting the agent's decisions and planned actions. Despite previous research addressing human factors in autonomous shipping and remote control, few studies have focused on how to support the humans' mental processes in this new role. Therefore, this study performed a goal-directed task analysis addressing goals, decisions, and SA requirements for human-supervised collision avoidance. Data was obtained from in situ observations and interviews with nine navigators onboard passenger ferries, an appraisal of the collision regulations, and of relevant company documentation. The task analysis identified specific SA requirements to make agents, capable of collision and grounding avoidance, transparent to their users. The results further indicate a change towards increased cognitive activities required to verify agent performance. Therefore, providing insight into the agents' internal reasoning and actions becomes a key consideration in supporting future supervisors. Given the trends towards the application of artificially intelligent agents capable of autonomous behaviour, this study anticipates that transparency becomes an essential prerequisite for safe and effective human-autonomy system oversight.

## 1. Introduction

### 1.1. Towards autonomous shipping

In recent years, the maritime industry has shown increased interest in developing autonomous solutions with the aim to achieve more efficient, punctual, and safer operations (Kretschmann et al., 2017; Wróbel et al., 2017). To illustrate, the MUNIN research project (Maritime Unmanned Navigation through Intelligence in Networks) explored safety and autonomy in a dry bulk carrier for deep-sea shipping (Burmeister et al., 2014) and DNV demonstrated its ReVolt concept to explore crewless short-sea shipping (DNV, 2018). Furthermore, Rolls Royce proposed an autonomous ferry in Finland showing its capabilities for fusing sensor information, detecting obstacles, avoiding conflicts and berthing automatically (Rolls Royce, 2018). The AUTOSHIP research

project aimed to build, test, and operate two autonomous vessels with capabilities for short sea shipping and inland waterway scenarios (AUTOSHIP, 2019). In Japan the commercial ship Suzaku conducted a 790-kilometre trial using a container ship, testing its autonomous navigation capabilities (NYK, 2022). Finally, in Norway, the Yara Birkeland container ship and ASKO barges have commenced service with the aim is to sail without crew onboard, with remote supervision, in the near future (ASKO, 2022; Yara International, 2022). Table 1

Although the reasons for pursuing autonomous operations are diverse, the prospect of reduced manning has sparked the interest of the industry. Autonomous and unmanned ships may allow for new and more efficient ship designs enabling lighter structures, reduced voyage costs, and/or increased payload capacity (Kretschmann et al., 2017; Kurt and Aymelek, 2022). In addition, the prospects of reduced crew (Kooij and Hekkenberg, 2020), and reduced number of fatalities by removing

\* Corresponding author.

E-mail address: [koen.van.de.merwe@dnv.com](mailto:koen.van.de.merwe@dnv.com) (K. van de Merwe).

<https://doi.org/10.1016/j.ssci.2023.106329>

Received 27 June 2023; Received in revised form 1 September 2023; Accepted 24 September 2023

Available online 30 September 2023

0925-7535/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

**Table 1**  
Interviewee demographics and experience with selected technologies.

	Min	Max	Mean	Median	Std. dev.	Yes	No
Navigational licence (D2/D1)						9	0
Navigational experience (yrs.)	7	48	30,6	35,0	13,9		
Experience at sea (yrs.)	15	53	36,3	41,0	14,2		
Experience with:							
Track control autopilot						9	0
Auto-docking						6	3
Auto-crossing						9	0
Auto-departure						5	4

personnel from the sharp end of the operations are also motivating factors (Wróbel et al., 2017). One key challenge to be resolved in moving towards autonomous and potentially unmanned shipping is collision and grounding avoidance. The ability to safely sail from port-to-port, resolve potential collisions with other ships, whilst avoiding grounding is an essential part of ship navigation. As such this piece of the autonomy puzzle has received much attention in ship autonomy research (Chaal et al., 2023; Li et al., 2023; Ramos et al., 2019; Statheros et al., 2008; Wróbel et al., 2017).

### 1.2. The role of humans in collision avoidance

The current legal framework concerning the obligations of the ship's master (Vojković and Milenković, 2019) and the International Maritime Organisation's Resolution A.1047 – Principles of minimum safe manning (IMO, 2011) have requirements for the presence of humans onboard ships. The objective of this regulation is “to ensure that a ship is sufficiently, effectively and efficiently manned to provide safety and security of the ship, safe navigation and operations at sea [...]” (2011, p. 3). However, the regulation also stipulates that the ship's manning level should be based on an evaluation of factors including the ship's level of automation and the degree of shore-side support. Notwithstanding other regulations, such as International Convention for the Safety of Life at Sea (SOLAS; IMO, 1974) and the International Convention on Standards of Training, Certification and Watchkeeping for Seafarers (STCW; IMO, 1978), this means that to allow for unmanned and autonomous ships, the level of ship automation and shore-side support needs to be sufficient. However, as there are challenges related to developing highly reliable systems capable of performing collision avoidance in all relevant situations, most of the maritime autonomy concepts currently under development employ a human supervisor that monitors the ship's operation and can intervene when the system's performance is insufficient (Rødseth et al., 2021). Although the idea of having a human as a backup to compensate for system limitations is attractive, the introduction of humans in a supervisory position to monitor advanced autonomous systems can introduce new and unknown risks (Ramos et al., 2019, 2018; Veitch and Alsos, 2022).

At present, navigators determine the presence of collision risk and perform relevant avoidance manoeuvres supported by a range of instrumentation and control systems (Boissier, 2018; Cockcroft and Lameijer, 2011). In the future, autonomous ships are envisioned to deploy artificially intelligent agents capable of sensing its environment and executing goal-directed behaviour using actuators (Russell and Norvig, 2022). Such agents are anticipated to be able to perform the navigational tasks, independent of human input, by having an ability to detect targets, determine collision risks, decide how to avoid collisions, and execute avoidance manoeuvres. To understand how navigational risk changes when transitioning from human- to autonomous collision avoidance, Fan et al. (2020) explored a range of factors influencing the navigational risk of autonomous ships. They found that humans supervising autonomous systems may, among others, be prone to information overload, reduced vigilance, automation bias, data misinterpretation, inappropriate SA, lack of knowledge or skills, and lack of mechanisms to intervene. As such, among the human-, ship-, environmental-, and

technological factors considered, human-related factors were most frequent among these four. These findings corroborate with the state-of-the-art regarding human factors related challenges associated with monitoring complex systems, such as complacency (Parasuraman and Manzey, 2010), automation bias (Wickens et al., 2015), reduced vigilance (Wohleber et al., 2019), increased workload during manual take-overs (Endsley, 2017), and SA related issues (Endsley and Kiris, 1995). Given these findings in the context of autonomous shipping, it becomes clear that further understanding is needed regarding the role of the human as a supervisor in autonomous collision avoidance.

### 1.3. Human-supervised collision avoidance

Collision avoidance is internationally regulated through the collision avoidance rules and were developed to provide the “rules of the road” for maritime traffic. The collision regulations (“COLREGs”) came into force in 1977 and, although the rules have been amended several times since, they do not consider autonomous ships (IMO, 1977). However, ongoing work at the IMO is considering how to amend the rules to accommodate for the presence of autonomous ships in national and international waters (IMO, 2022, 2018). Nevertheless, for the foreseeable future, it is assumed that the COLREGs will apply for all ships; autonomous and conventional alike (Ringbom, 2019; Zhou et al., 2020).

An important challenge to consider in accommodating the COLREGs for autonomous ships is how to address the qualitative and interpretative nature of the rules (Porathe, 2019a). For example, Rule 8 “Action to avoid collision” states: “any action to avoid collision shall be taken in accordance with the Rules of this Part and shall, if the circumstances of the case admit, be positive, made in ample time and with due regard to the observance of good seamanship.” This rule is a good example of the use of terminology that may make sense to a human, but that an autonomous system may find difficult to conform with. That is, what does “ample time” mean in this context? And how does one quantify “good seamanship” (Porathe, 2019b; Zhou et al., 2020)? Furthermore, Rule 2 “Responsibility” states: “In construing and complying with these Rules due regard shall be had to all dangers of navigation and collision and to any special circumstances, including the limitations of the vessels involved, which may make a departure from these Rules necessary to avoid immediate danger.” Interestingly, this rule takes interpretability one step further by stating that every ship has the responsibility to avoid collisions even if this implies breaking the rules. It may be a challenge to envision an autonomous CAGA system designed to adhere to the rules whilst at the same time designed to break them (Miyoshi et al., 2022).

One could consider setting boundaries for the autonomous system by defining when collision avoidance responsibility is performed by the system and when it is performed by the navigator. This would allow the system to operate autonomously within clearly defined limits and delegate responsibility when it cannot perform its function to a sufficient degree. For example, some authors have argued for an “operational envelope” concept where the ship's design, intended operations and environment are defined to scope the intended operations of the system (Rødseth et al., 2021). Within these limits, all foreseen tasks and operations reside, including tasks by the system and the human supervisor. When an autonomous CAGA system is unable to resolve a collision situation, it can prompt the supervisor by providing a take-over request allowing the supervisor to assume the collision avoidance task or otherwise take control of the ship. Alternatively, a supervisor may be unsatisfied with the performance of the system and initiate an intervention at the supervisor's own discretion. Regardless of the type of intervention, the introduction of a CAGA system changes the role of navigators from the ones performing collision avoidance to ones supervising an agent performing collision avoidance.

To better understand this change, some studies have focused on how to design remote control centres to support humans in performing remote supervision of autonomous vessels (Man et al., 2018, 2016; Porathe, 2021, 2014). In addition, some have built and evaluated



control centre research facilities to study human factors in potential remote supervision scenarios (Alsos et al., 2022; Hoem et al., 2022). Others have focused on developing risk models and methods to accommodate human reliability in maritime autonomous collision avoidance with remote supervision (Ramos et al., 2020, 2019; Thieme and Utne, 2017; Ventikos et al., 2020). However, few studies have focused on the human’s cognitive changes related to this role change. In addition, few studies have focused on how to support the cognitive processes to aid human supervisory performance in an autonomy context. To this end, this paper investigates how goals, decision making, and information needs change and which requirements need to be set on the information provided by the autonomous system to support effective human oversight.

1.4. This study

This study aims to explore human-supervised autonomous collision and grounding avoidance by systematically investigating two case studies: a case where collision and grounding avoidance is performed by a navigator, and a case where this task is performed by a human-supervised system. In particular, this study aims to systematically explore how human supervisors will be able to understand the system’s collision and grounding avoidance decisions and actions through mapping goals, decision, and information needs. By focusing on the cognitive aspects rather than the technology, the analysis is not constrained by how autonomous collision avoidance technology is implemented.

For example, according to the degrees of autonomy as defined by IMO (2018), CAGA systems may be deployed on a manned bridge where only the collision avoidance and grounding task is allocated to the CAGA

system (degree 1). The system may be deployed on a remotely controlled manned ship where local seafarers can be called to the bridge and assess the CAGA system’s plans (degree 2). The system may be deployed on an unmanned ship with remote supervision by operators from a remote-control centre (degree 3). Finally, the system may be deployed on an unmanned and autonomous ship capable of independent operations (degree 4). Although the latter case is the most futuristic and visionary, for the foreseeable future it is unlikely that such vessels will be devoid of human interaction and are thus likely to have some form of supervision.

Given the multitude of potential control configurations, function allocation possibilities, and degrees of autonomy, this study has chosen to focus on the common denominator, i.e., CAGA systems and their interaction with human users. That is, the study concentrates on identifying the information required to make CAGA systems understandable and predictable to the human user irrespective of how autonomy is solved. This allows the analysis to focus on what humans ideally would like to know to make decisions and achieve their goals within their operational context (Endsley et al., 2003). As such, this study aims to establish an information basis for developing human supervised CAGA systems in the ship autonomy context.

2. Method

2.1. Establishing a framework

A Goal-Directed Task Analysis (GDTA) was performed to determine and describe the cognitive processes involved in collision avoidance (Endsley et al., 2003). Specifically, the goals, decisions, and SA requirements relevant for avoiding collisions were captured providing an

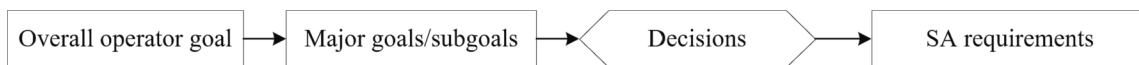


Fig. 1. The Goal-Directed Task Analysis approach (Endsley et al., 2003).

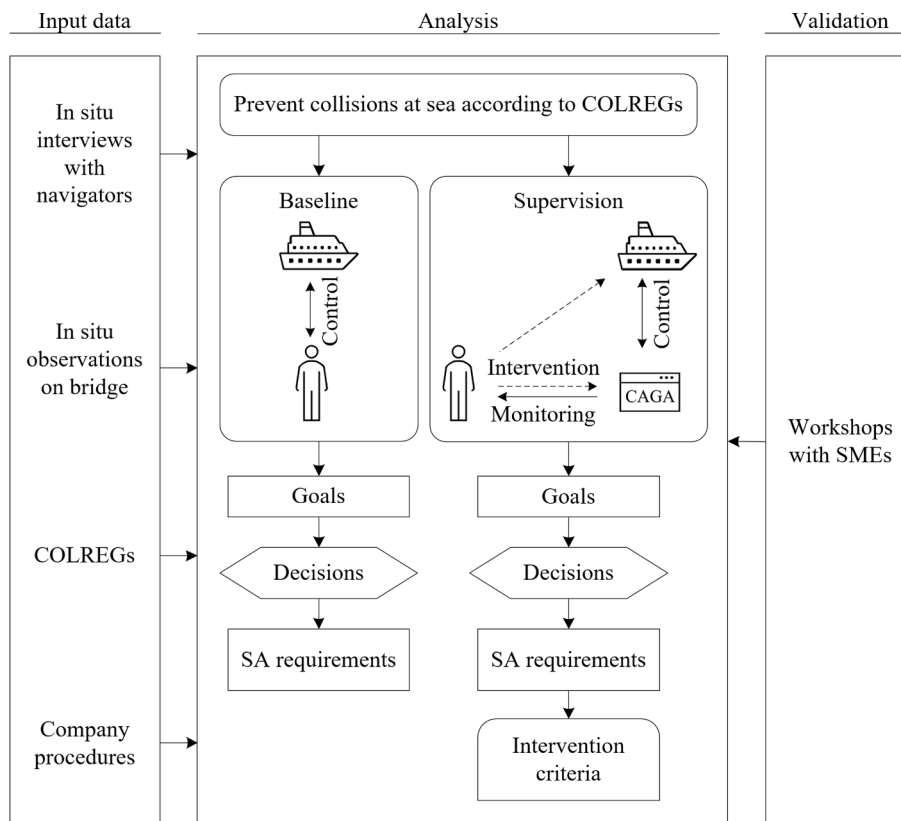


Fig. 2. The analysis framework employed in this study.

understanding of how decisions are made when handling vessel collision situations (see Fig. 1). Task analysis is typically performed prior to system design to understand what humans are required to do, in terms of tasks and/or cognitive processes, to achieve an overall goal. With this method, a detailed picture of human involvement in a system is created, and requirements are established to ensure the goals can be achieved (Kirwan and Ainsworth, 1992).

As depicted in Fig. 2, the analysis encompassed two case studies: a baseline case, where a navigator was in direct control of the vessel, and a case where a system performs collision avoidance autonomously under human supervision, including intervention. The baseline case was used to analyse how collision avoidance is performed in the present, and to form the information basis for analysing this task in future supervised autonomous collision avoidance. Based on four data sources, the analysis was performed for each case, and the results were validated with independent SMEs.

## 2.2. Input data

Four information sources were used as input data: in situ interviews with navigators, in situ observations on ship bridges, the COLREGS, and collision avoidance procedures of a ferry operator.

### 2.2.1. In situ interviews with navigators

Nine navigators, all of whom had an active navigational license at the time of data collection (see Table 2) were interviewed using a semi-structured format on the bridge of ro-ro ferries whilst on active duty. The interviews were conducted over a period of two months and were performed on multiple ships operated by the same company, sailing the same route. The first part of the interview focused on conventional collision avoidance and aimed to capture goals, decisions, and the information elements navigators typically use in collision avoidance (i.e., the baseline case). The questions for this part of the interview focused on how navigators establish an awareness of the surrounding traffic, determine collision risk, and decide on the actions needed to avoid collisions. In addition, the interviews captured which information and equipment navigators typically need to perform these tasks. The second part of the interviews focused on the navigators' potential interaction with an autonomous CAGA system (i.e., the supervision case). For this part, a series of questions based on a modified MITRE Human-Machine Teaming Systems Engineering Guide (MITRE, 2018) was used to identify SA requirements when humans team with advanced automation to perform a task.

Participation in the interviews was based on informed written consent, voluntary, and was approved by the Norwegian Centre for Research Data's ethics committee (reference nr. 579620). As the interviews were performed whilst the navigators were on active duty, questions were only asked when these did not interfere with their work, and at their discretion. As such, the total interview time was between four and six hours per interview.

### 2.2.2. In situ observations

In situ observations were performed of how potential collision situations were handled by navigators on duty. The observations were

**Table 2**  
Workshop participant demographics and experience with selected technologies.

	Min	Max	Mean	Std. dev.	Yes	No
Navigational license (D2/D1)					2	0
Navigational experience (yrs.)	5	18	11,5	9,2		
Experience at sea (yrs.)	5	10	7,5	3,5		
Experience with:						
Track control autopilot					2	0
Auto-docking					0	2
Auto-crossing					1	1
Auto-departure					0	2

performed whilst interviewing the navigators and were conveniently used as examples and objects of enquiry during the interviews. Therefore, potential collision situations that arose during the visits were observed, noted, and discussed in detail.

### 2.2.3. COLREGs

An appraisal of the COLREGs was performed to identify goals, decisions, tasks, and information needs provided in the rule descriptions. For this study, the analysis focused on the general conduct of ships in any cases of visibility. The COLREGs describe, to a degree, the tasks to be performed in ship-to-ship encounters. As such, the information already embedded in the rule descriptions was used to understand how navigators establish an awareness of the traffic, estimate safe speed, determine collision risk, and decide on which actions are needed to avoid collisions. Specifically, the following rules described in COLREGs Part B – “Steering and Sailing Rules, Section I - Conduct of vessels in any condition of visibility” were within the scope of the analysis (IMO, 1977):

- Rule 5 – Look-out
- Rule 6 – Safe speed
- Rule 7 – Risk of collision
- Rule 8 – Action to avoid collision

This subset of the COLREGs is concerned with the process of establishing awareness of the environment, how to choose a safe sailing speed given the circumstances, how to assess if a risk of collision is present, and how to decide which actions to take to avoid collisions. This subset of the rules lays the foundation of the collision avoidance process, applicable in all circumstances, traffic situations, and visibilities. However, these rules do not describe how to address specific collision situation types, e.g., overtaking, head-on, and crossing situations, as these are described in rules 14, 15, and 16 respectively. Therefore, this analysis does not address specific traffic encounters, but focuses on the generic processes of building SA through look-out, establishing a risk picture, and performing collision avoidance manoeuvring through avoidance actions and safe speed. Therefore, as this study is interested in the goals, decision-making, and SA requirements for collision avoidance, the analysis was limited to this subset of the COLREGs only. Also, this study focused on ship-to-ship situations where both parties are in motion and collision was therefore out of scope of this analysis.

### 2.2.4. Company procedures

Operational procedures from a ferry operator were reviewed to supplement and support the COLREGs appraisal and identify any concrete operationalisations relevant for collision avoidance. Specific focus was on written documentation about the ferry operator's interpretation of the rules and the roles, responsibilities, and actions relevant in avoiding collisions.

## 2.3. Analysis

Two cases were analysed: a baseline case, and a supervision case. The baseline case concentrated on collision avoidance performed by a navigator on the bridge and assumed a modern sea-going ship sailing in open waters. The collision avoidance task was assumed to be exclusively performed by the navigator, supported by modern navigational equipment, e.g., a radar with automatic plotting aids (ARPA), automatic identification system (AIS), and electronic chart display and information system (ECDIS). The analysis focused on the navigator's cognitive activities associated with building and maintaining SA of the ship in its operational environment (Rule 5), determining safe sailing speed (Rule 6), determining collision risk (Rule 7), and planning and executing avoidance manoeuvres (Rule 8) according to COLREGs (see Fig. 3).

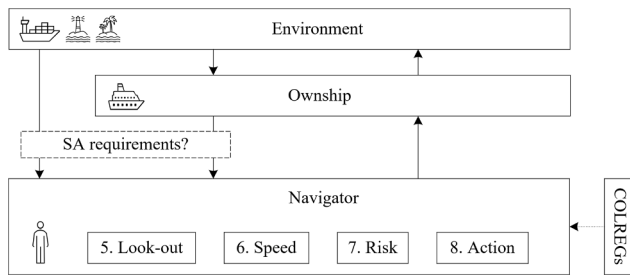


Fig. 3. The baseline case, in which the navigator performs the look-out task, determines safe sailing speed, identifies potential collision risks, and performs collision avoidance manoeuvring.

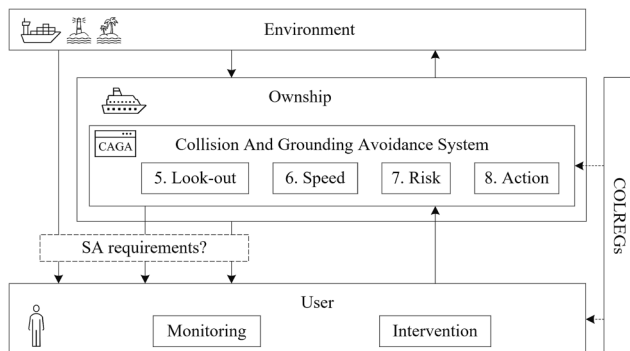


Fig. 4. The supervision case, in which the user is tasked with supervising the CAGA system and perform intervention when needed.

The supervision case concentrated on the human supervision of a CAGA system to which the collision avoidance function was allocated. This system was assumed to be able to sense and keep track of its environment through a suite of sensors, estimate a safe sailing speed, determine collision risk, and calculate, plan, and execute avoidance manoeuvres, in accordance with COLREGs. In this case, the navigator was assumed to retain the final authority of the collision avoidance manoeuvring of the ship and therefore had the role of the user of the CAGA system tasked with its supervision. The location of the user was not defined as this case aimed to answer the question as to what information the system should disclose about its itself to provide understandability and predictability to its user. This case is therefore implementation agnostic, and the analysis is therefore limited to identifying the SA requirements for CAGA such that monitoring, and intervention can be performed by a human user (see Fig. 4).

#### 2.4. Validation

Data gathering and analysis was performed by the first author. To validate the data, workshops were held with two independent navigators (see Table 2 for demographics and experience). In these workshops, the analysed data was reviewed and discussed, and changes, clarifications, and revisions were made where needed. The use of independent navigators provided validation of the data set and additional confidence in its contents.

### 3. Results

The results for the analyses are described for each of the COLREGs rules 5 to 8 respectively.

#### 3.1. Rule 5: Look-out

According to COLREGS rule 5 - Lookout, “every vessel shall at all times maintain a proper look-out by sight and hearing as well as by all available means appropriate in the prevailing circumstances and conditions so as to make a full appraisal of the situation and of the risk of collision” (IMO, 1977). Establishing and maintaining a lookout is a continuous task that lies at the foundation of collision risk avoidance. For the baseline case, as depicted in Fig. 5, key goals for the navigator to perform lookout include determining the presence of vessels and other navigational constraints to perform an appraisal of the situation and determine collision risk. To achieve these goals, many information sources are available to the navigator, including information obtained from direct sensory perception (i.e., through sight and hearing), and information obtained through instruments (e.g., radar, AIS). Initial collision risk estimation is achieved by observing the relative motion of targets, e.g., by taking a visual bearing, if possible, and through available radar and AIS functionality on the bridge, e.g., (Time to) Closest Point of Approach (TCPA/CPA) estimates, Bow Crossing Range/Time (BCR/BCT) estimates, the use of true- and relative vectors, target vessel course changes, messages from Vessel Traffic Services (VTS) and other vessels.

In the supervision case, the look-out function is now performed by the CAGA system, and the aim of the user of CAGA is to verify that the system has performed an adequate appraisal of the situation and determined collision risk. In order for the user to obtain insight into the system’s perception of its environment, the CAGA system should, at minimum, show which elements in its environment it has identified (e.g., vessels, objects, terrain, and other navigational constraints) and which of these pose a collision risk (see Fig. 5). By providing insight into what the system perceives and how it appraises its environment, the user should have an adequate information basis to understand the system’s interpretation of its surroundings. An intervention by the user could be driven by a mismatch between the information the system depicts and the user’s perception of the environment. As such, Fig. 5 also depicts the need for the information from the baseline case to be available to the user. For example, an incompatibility between the targets, objects, or terrain identified by the CAGA system and reality may trigger scrutiny by the user and initiate a possible intervention. This may especially be relevant when a missed target has consequences for the collision risk. Therefore, for a user to be able to assess the veracity of the information provided access should be provided to independent information sources. In a degree 1 level of autonomy, this may be provided by systems already available on the ship’s bridge, including the outside view. However, in a remote supervision case (degree of autonomy 2, 3, and 4), this information should be provided through sources independent of the CAGA system.

#### 3.2. Rule 6: Safe speed

Rule 6 – Safe speed is described as “every vessel shall at all times proceed at a safe speed so that she can take proper and effective action to avoid collision and be stopped within a distance appropriate to the prevailing circumstances and conditions” (IMO, 1977). In determining safe speed, the navigator shall consider several influencing factors that shall be taken into account such as the state of visibility, traffic density, manoeuvrability, background lights, meteorological conditions, and draught in relation to water depth. In addition, limitations of equipment, i.e., radar, shall be considered. The resulting safe speed is therefore not a value that can be determined upfront (notwithstanding speed restrictions imposed by local authorities) but is derived from a continuous cognitive process performed by the navigator in which these variables are mentally weighted resulting at a speed that is deemed safe (see Fig. 6).

In the supervision case, the primary goal of the user is to verify if the CAGA system provides adequate speed input to the control system of the vessel such that the vessel’s speed is adapted to the circumstances. This

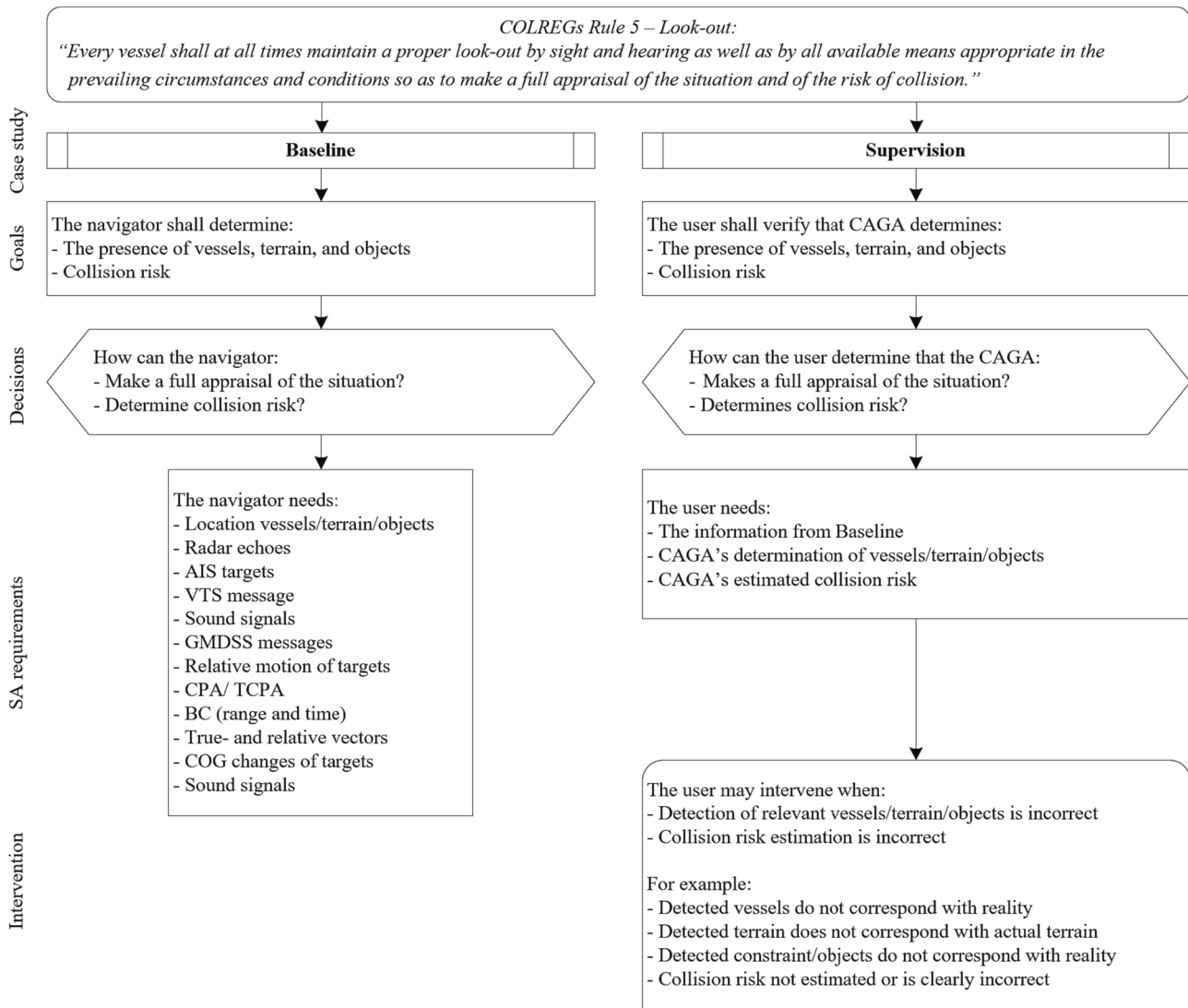


Fig. 5. GDTA for COLREGs Rule 5 – Look-out. Key: COG: Course Over Ground.

means that the user should be able to evaluate if the speed the CAGA system has chosen is within an acceptable operational range. Any deviations from this range (e.g., speed too high or too low) may trigger intervention, or at a minimum, a search for better understanding the reason for the speed setting. For example, as depicted in Fig. 6, rule 6a describes the range of parameters to be considered in the estimation of safe speed. An error in CAGA’s interpretation of one of these parameters may result in incorrect speed estimations. Also, rule 6b states that limitations to radar equipment should be taken into consideration when determining safe speed. For the CAGA system, this means uncertainties in its sensor systems (e.g., cameras) shall be considered when estimating speed. In cases of uncertainties in the sensor information, or errors in the CAGA system’s interpretation of sensor uncertainty, user intervention can be considered.

### 3.3. Rule 7: Risk of collision

Rule 7 – Risk of collision is described as “a. every vessel shall use all available means appropriate to the prevailing circumstances and conditions to determine if risk of collision exists. If there is any doubt such risk shall be deemed to exist” (IMO, 1977). In addition, the subsections to rule 7 describe that “b. [...] proper use shall be made of radar equipment [...]”, “c.

[...] assumptions shall not be made on the basis of scanty information [...]”, and d. that compass bearing estimations shall be used whilst being aware of its limitations at close range, when target vessel is large, or when target vessel uses a tow. After the initial risk estimation as described in rule 5, rule 7 describes the risk assessment process in more detail. Here, the goal for the navigator is ultimately to determine the collision risk through a process of obtaining sufficient and reliable information and applying risk estimation techniques. An important element here is how uncertainty is handled. As the navigator estimates risk through primary senses (sight and hearing) and with the help of decision aids (e.g., radar and AIS), uncertainty in the information available to the navigator can hamper accurate risk estimation. Reliable and multiple independent information sources should minimize the need for assuming collision risk, but if these are not available, a risk of collision shall be deemed to exist (see Fig. 7).

In the supervision case, the primary goal of the user is to verify that the CAGA system has correctly determined a risk of collision. This task is at the centre of the CAGA system as correct risk estimation is a prerequisite for correct risk avoidance. Essentially, this means the user should be able to verify that the system has determined collision risk given the detected targets, their estimated future tracks, and CPA estimates. A mismatch between the system’s estimation of collision risk and

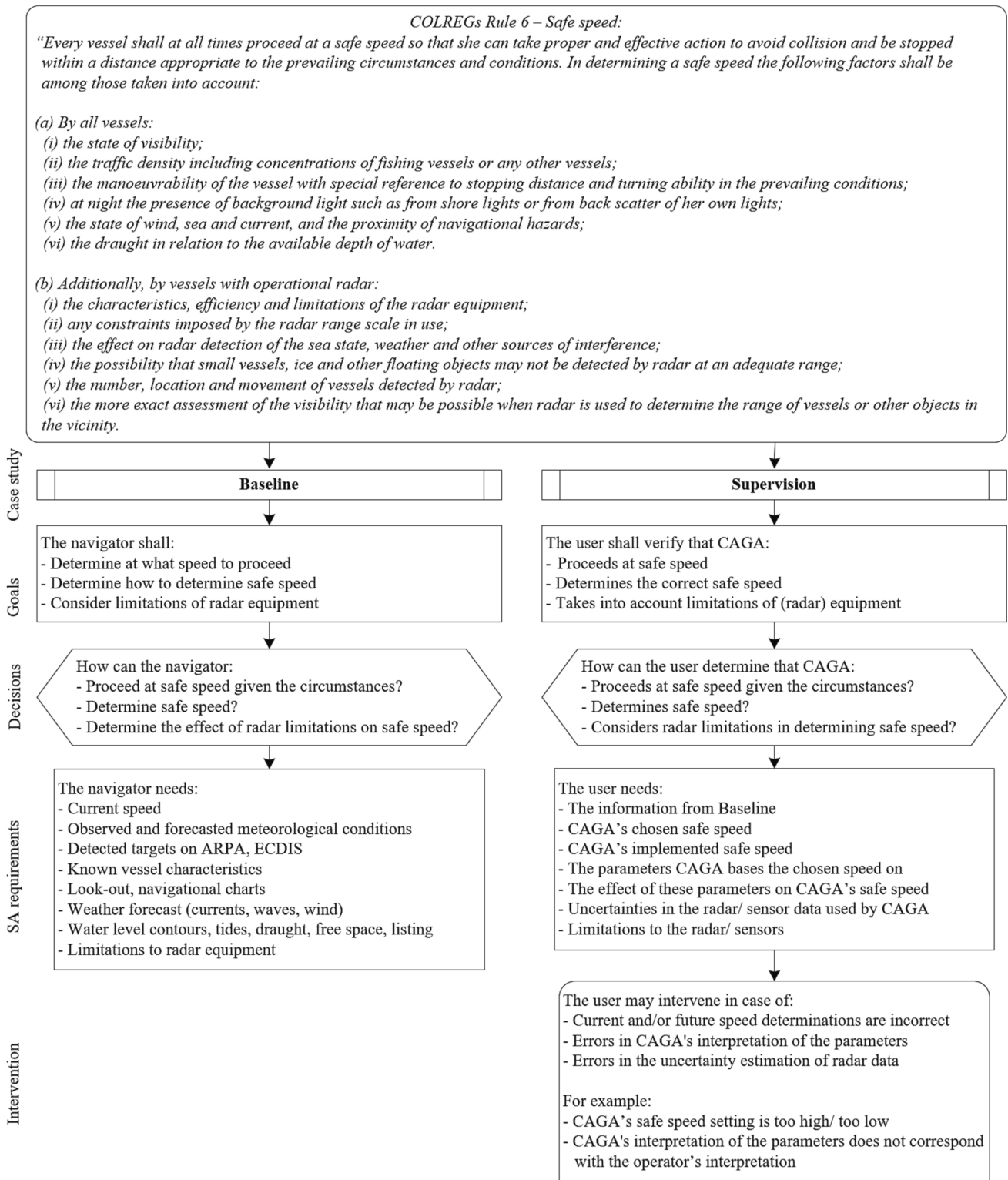


Fig. 6. GDTA for COLREGs Rule 6 – Safe speed.

the user’s estimation may be a reason for the user to intervene or at minimum to understand the background for the system’s collision risk estimate. For example, rule 7b, c, and d state the need to evaluate risk of collision as early as possible, not to make assumptions based on scanty information, use compass bearings, and that risk estimations should consider vessel size, tow, or proximity to the target vessel. This means that a late detection, a detection based on uncertain sensor information, or one that fails to consider vessel type, size, and distance may lead to

insufficient collision estimation and may require intervention (see Fig. 7).

### 3.4. Rule 8: Action to avoid collision

Rule 8 – Action to avoid collision is described as “a. any action to avoid collision shall be taken in accordance with the Rules of this Part and shall, if the circumstances of the case admit, be positive, made in ample time



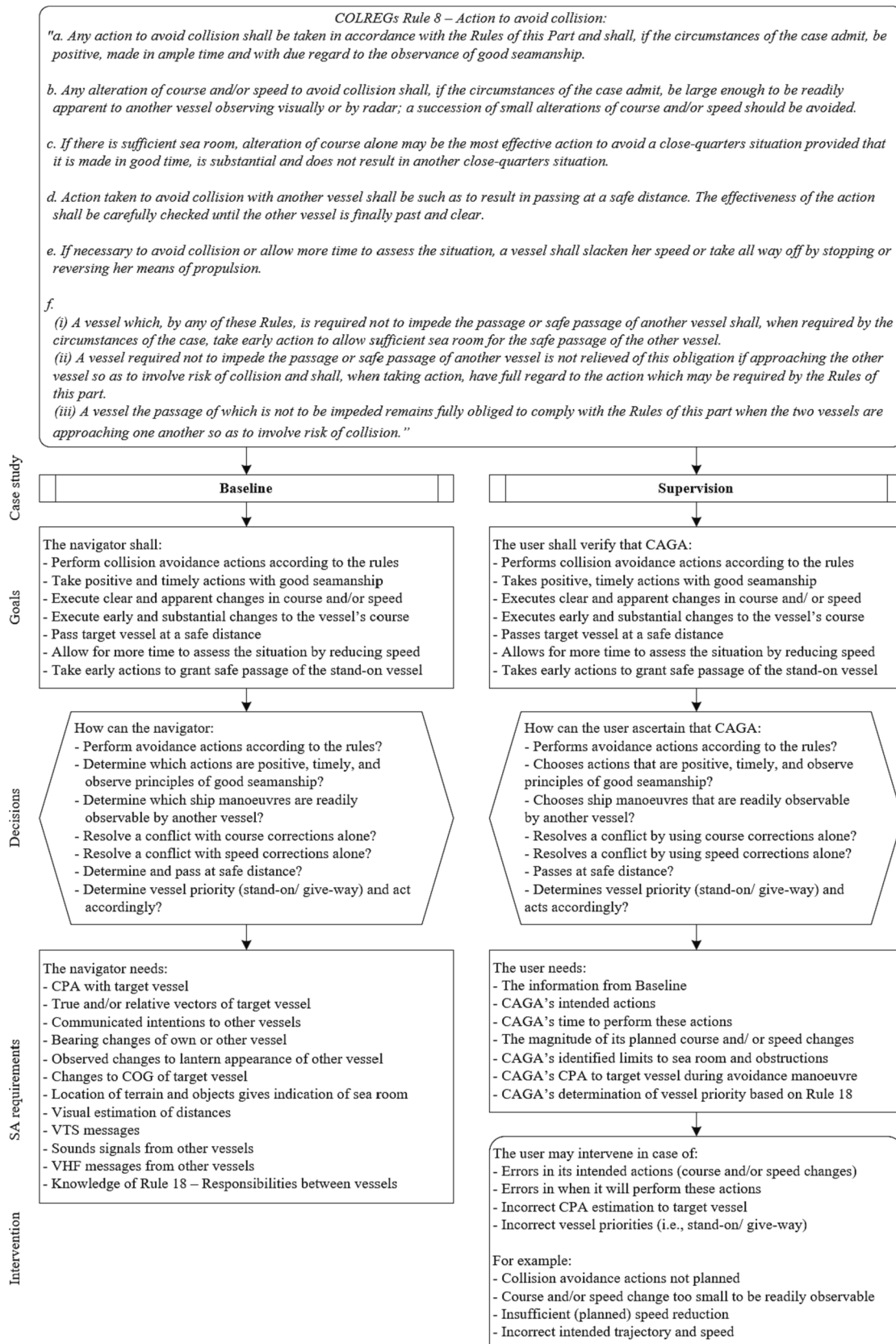


Fig. 8. GDTA for COLREGs Rule 8 – Actions to avoid collision.

e.g., a minimum 10-degree course change, and/or a 5–10 knots speed change was suggested.

For the supervision case, the primary goal of the supervisor is to verify that the CAGA system executes collision avoidance manoeuvring according to the COLREGs. That is, according to rule 8a, collision avoidance manoeuvres should be executed using positive, timely, and actions with good seamanship. As mentioned in the baseline case, there are no explicit acceptance criteria for what constitutes a manoeuvre with good seamanship, but it was clear that movements that clearly demonstrate the vessel's intention are central to this notion. This means that a CAGA system built to manoeuvre according to COLREGs should be able to make its avoidance manoeuvres apparent to other vessels (rule 8b), using substantial route and/or speed changes (rule 8c and 8e), ensuring sufficient distance to the target vessel (rule 8d), such that its intentions are clearly understood. This also applies to cases where the CAGA system is the stand-on vessel and required to continue its course (rule 8f), and where it is the other vessel's obligation to take evasive action. Therefore, when collision avoidance is required, the task of the user is to verify if the CAGA system's current and intended avoidance manoeuvres satisfy the notion of good seamanship. Incorrect, insufficient, or untimely (intended) actions may trigger a decision to intervene (see Fig. 8).

## 4. Discussion

### 4.1. Information support for decision-making in a dynamic context

Transitioning from conventional collision avoidance to human-supervised autonomous collision avoidance has consequences for the locus of the decisions and actions associated with this task. In conventional collision avoidance, decisions, and avoidance actions are performed by the navigator. In the supervision case, these have now been outsourced to an artificially intelligent agent capable of collision and grounding avoidance. In turn, the supervisor is left with the task of overseeing the agent's decision-making and actions. The CAGA system continuously detects and analyses its environment in search for traffic conflicts, and subsequently makes decisions and executes manoeuvres to avoid collisions whilst informing the human supervisor about its actions. As the system is not dependent on input from the supervisor to perform its actions unless deemed necessary, this type of automation implementation requires a high degree of sophistication. The role of the human has therefore changed from an active one to supervisory one that manages a system by means of exception (Cummings et al., 2007; Sheridan and Verplank, 1978).

In this type of supervisory control, the system informs the supervisor of a collision avoidance solution and allows for a restricted time to veto before the solution is executed. In our supervision case, the system may detect a ship on collision course, determines that own-ship is the give-way vessel, calculate evasion alternatives and select the optimal option based on a cost function. This alternative is subsequently presented to the supervisor to evaluate, and the supervisor is given a limited amount of time to veto the solution before it is executed. However, as navigating is an activity performed in a continuously evolving dynamic context, vessel-to-vessel encounters may occur at any given moment or change over time. As such, a supervisor should not depend on, and wait for, avoidance solutions from the system to present themselves. For example, as identified for rule 8a in Fig. 8, a supervisor may intervene in the system in cases of errors in the CAGA system's current actions, planned actions, or timeliness of its actions. Furthermore, rule 7a states that a supervisor may intervene if there is an error in the risk estimation capabilities of the system. For such cases, it would not be prudent for a supervisor to wait for the system to come up with a (potentially unreliable) solution, but to act proactively and intervene at own initiative.

Self-paced transitions have demonstrated potential benefits in terms of control performance, compared to system-initiated transitions in automated driving tasks (Eriksson and Stanton, 2017; Kircher et al., 2014). Self-paced take-overs, however, imply that the supervisor can

assess system performance, i.e., able to assess the system's and/or ship's behaviour close to the edge of the operational envelope but which has not (yet) led to a take-over request. For example, upon detecting a collision situation with another vessel, the supervisor may notice the absence of an avoidance manoeuvre by their own ship, despite expecting there to be one (Fig. 8, rule 8a). Alternatively, the system may perform an avoidance manoeuvre that, according to the supervisor, is not deemed sufficiently safe in terms of distance (Fig. 8, rule 8d), leading the supervisor to decide to perform an avoidance manoeuvre manually. Such deviations between system behaviour and the supervisor's expectations of the system's behaviour may prompt a decision to intervene at own discretion. To make such decisions in an ever-changing context, and to intervene at any moment and at one's own initiative, the supervisor needs to have appropriate SA based on access to continuous, sufficient, and relevant information (Endsley et al., 2003; Sheridan, 2002). This study has made explicit which information elements would be needed to support supervisors performing this task.

### 4.2. Disclosing the agent's reasoning

Evaluating agent behaviour in a dynamic high-risk context puts responsibility on the human to adequately perform the supervisory task. Essentially, in the supervisory case the role of the human is transformed from one actively performing collision avoidance, to a supervisor vetoing the CAGA system's solutions (Veitch et al., 2022). Earlier and ongoing research have explored the complexities of humans interacting with highly sophisticated automation in terms of human performance (Bainbridge, 1983; Endsley, 2017; Strauch, 2018). An intricate challenge exists in cases where humans interact with systems to which more automation is added. The addition of automation adds to the overall reliability and robustness of the system, whilst the human's ability to take over manually decreases due to reduced SA. This "automation conundrum" and its effects need to be well understood when developing human-supervised CAGA systems in which humans take a management-by-exception role (Endsley, 2017). That is, despite some evidence for improved overall operator performance, management-by-exception may, amongst others, come at the cost of increased levels of automation bias (Cummings and Mitchell, 2007).

Automation bias occurs when operators do not search for information disconfirming the proposed solution by the system (Parasuraman and Manzey, 2010). The consequences of automation bias in supervisory control of autonomous CAGA systems are a concern because of the risks associated with erroneously executed avoidance manoeuvres. Automation bias may manifest itself when humans interact with agents designed to aid decision making in complex environments. The tendency to rely on the agent's decisions makes humans less critical to scrutinise the background information for the proposed solution. Therefore, placing too high trust in the agent's proposals becomes a problem when the agent is not fully reliable (Bowden et al., 2021; Hutchinson et al., 2022; Lee and See, 2004). However, some have suggested that insight into the agent's reasoning, allowing a supervisor to understand its internal activities, may alleviate some of the effects of automation bias (Gegoff et al., 2023; Wright et al., 2016). Also, in a recent study with navigators using a tool to perform collision avoidance manoeuvring, agent transparency was suggested as a means to increase trust and reliance in the technology (Aylward et al., 2022).

Making an agent's internal reasoning available to its user, i.e., making it transparent, should provide the ability for a user to understand what the agent is doing, why it is doing it, and what it will do next (Endsley, 2017). Several recent reviews have investigated the relation between automation transparency and typical human factors variables, demonstrating a promising effect in terms of improved operator performance variables (e.g., decision making) and SA without the added cost of mental workload (Bhaskara et al., 2020; van de Merwe et al., 2022). Despite variations in theoretical models underpinning the transparency concept (Rajabiyazdi and Jamieson, 2020), most



definitions agree that transparency constitutes an agent property (e.g., Chen et al., 2014; Christoffersen and Woods, 2002; Endsley et al., 2003; Norman, 1990; Skraaning and Jamieson, 2021). In other words, understandability and predictability of the agent's actions is something that can, theoretically, be designed in its human machine interface.

For supervising autonomous CAGA systems, the results of this study depict the SA requirements to allow the user to understand and predict the systems' actions. For example, to understand that the CAGA system is performing the lookout function (rule 5), it should show which vessels, terrain, and other objects and constraints it has detected. In determining which speed is safe given the circumstances (rule 6), the system should provide the chosen safe speed, the parameters this is based on, and any uncertainties in the sensor data affecting the chosen safe speed. In terms of estimating collision risk (rule 7), the system should show which vessels, in the short to long range, form a collision risk, including its understanding of vessel type, size, distance to target and the presence of a tow. Finally, in determining the collision avoidance manoeuvre (rule 8), the system should show, amongst others, which actions it intends to perform, when it intends to perform these actions, including its understanding of vessel priorities, i.e., which ship will stand-on, and which ship will give-way. As such, by using a cognitive task analysis approach, this study derived at SA requirements for making autonomous CAGA systems transparent to its user for this type of supervisory control.

#### 4.3. Practical implications and future work

This study assumed a change in function allocation and task distribution between the baseline- and the supervision case. That is, in the baseline case, the collision avoidance function was allocated to the human navigator, and in the supervision case, this function was performed to the CAGA system. Furthermore, the supervision case was analysed independent of how autonomy was solved. That is, as this study aimed to specifically derive at requirements to CAGA systems, we intended to avoid defining and analysing information sources that were specific to the context in which the system was deployed. In other words, whether CAGA systems are used as decision support systems for navigators on a manned bridge or deployed on autonomous ships remotely supervised by operators, should not influence the information depicting the system's *internal reasoning* regarding a collision situation. Still, when practically implementing autonomous solutions in their operational context, additional requirements to how the system can be independently verified need to be defined. In addition, future studies should address means for how supervisors can effectively intervene in the system. For example, on a bridge, a navigator may look out the window to cross-check the information from the collision avoidance system and may intervene using the ship's existing control options. In a remote-control centre, a supervisor may need access to information other than that coming from the collision avoidance system itself. Also, options for intervention depend on the control philosophy of the control facility and the autonomous ship defining capabilities for manual control. Therefore, to complement this work, future studies should extend this analysis to derive at context-specific requirements relevant for independent verification of CAGA system information and its intervention.

This study has not explicitly addressed how to implement and operationalise the SA requirements in a practical human machine interface. Designing a human machine interface for supervising a CAGA system requires a structured, human centred, and iterative process (e.g., ISO, 2019), and was outside the scope of this study. However, based on our results, acknowledging the study's limit to rules 5 to 8 only, it is clear that care should be taken not to overload the user with information. Implementing transparency is "as much an art as it is a science", and the potential for visual clutter and distraction should be considered as this may potentially offset transparency's benefits (Wickens, 2018, p. 39). Presenting all information identified in this study, as depicted in Figs. 5, 6, 7, 8, continuously is likely not prudent, and further work

should focus on teasing out which information has precedent over other information and in which circumstances (see van de Merwe et al., 2023 for preliminary results).

## 5. Conclusions

This study focused on deriving SA requirements for supervising autonomous CAGA systems based on a task analysis approach. The analysis showed the changes in goals, decisions, cognitive tasks, and SA requirements when transitioning towards supervisory control. The study also depicted changes to the navigators' role from those performing collision avoidance to those verifying a system performing collision avoidance, including deciding on whether to intervene. Decisions of this kind require the supervisor to perceive the current and anticipated actions of the CAGA system, create a mental model of its behaviour, and evaluate its adequacy in its context. Given the foreseen supervisory role of the human in autonomous shipping, these types of decisions are therefore likely to become an increasingly important part of the human's task repertoire (Banks et al., 2014). Therefore, it is essential that future supervisors of autonomous systems are supported in this role by ensuring agent reasoning is disclosed such that they remain on the information loop.

Current societal trends point towards the application of intelligent agents across safety critical industries, e.g., automobile (Society of Automotive Engineers, 2021), healthcare (Coronato et al., 2020; Loftus et al., 2020), and manufacturing (Elghoneimy and Gruver, 2012). Also in the maritime industry, ships with advanced autonomous capabilities are impending (IMO, 2021). Despite ongoing progress and advances, humans are anticipated to play a central role as supervisors tasked with the assurance of safe system performance. Given the well-known human performance challenges with supervisory control of highly automated systems, finding ways to assist supervisors in performing this task therefore becomes a key focus area going forward. This study has elucidated the relevance of affording human supervisors with insight into an autonomous system's reasoning to support human-autonomy system oversight and discussed transparency as an important prerequisite on the path towards safe and effective human-supervisory control.

### CRedit authorship contribution statement

**Koen van de Merwe:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Visualization, Validation, Writing – original draft, Writing – review & editing. **Steven Mallam:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Salman Nazir:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Oystein Engelhardtson:** Conceptualization, Methodology, Supervision, Writing – review & editing.

### Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Data availability

Data will be made available on request.

### Acknowledgements

The authors would like to thank the ferry operators and their navigators for the opportunity to perform interviews and observations. In addition, we would like to thank the independent navigators for their efforts in quality assuring the data. Finally, we would like to thank the independent reviewers for their valuable input to this manuscript.

This paper has been thoroughly expanded from an earlier version presented at the MTEC-ICMASS 2022 conference in Singapore (6th & 7th of April 2022) and published within its conference proceedings: see van de Merwe et al. (2022).

### Funding

This work was supported by the Research Council of Norway under grant number: 311365.

### References

- Alsos, O., Veitch, E., Pantelatos, L., Vasstein, K., Eide, E., Petermann, F.-M., Breivik, M., 2022. NTNU Shore Control Lab: Designing shore control centres in the age of autonomous ships. *J. Phys. Conf. Ser.* 2311, 012030 <https://doi.org/10.1088/1742-6596/2311/1/012030>.
- ASKO, 2022. Verdens første batterielektriske autonome sjodroner har ankommet Norge! URL <https://asko.no/nyhetsarkiv/verdens-forste-autonome-sjodroner-h-ar-ankommet-norge/> (accessed 5.10.22).
- AUTOSHIP, 2019. AUTOSHIP - Autonomous shipping initiative for European waters. URL <https://www.autoship-project.eu> (accessed 8.29.23).
- Aylward, K., Weber, R., Lundh, M., MacKinnon, S.N., Dahlman, J., 2022. Navigators' views of a collision avoidance decision support system for maritime navigation. *J. Navig.* 1–14 <https://doi.org/10.1017/S0373463322000510>.
- Bainbridge, L., 1983. Ironies of automation. *Automatica* 19, 775–779. [https://doi.org/10.1016/0005-1098\(83\)90046-8](https://doi.org/10.1016/0005-1098(83)90046-8).
- Banks, V.A., Stanton, N.A., Harvey, C., 2014. Sub-systems on the road to vehicle automation: Hands and feet free but not 'mind' free driving. *Saf. Sci.* 62, 505–514. <https://doi.org/10.1016/j.ssci.2013.10.014>.
- Bhaskara, A., Skinner, M., Loft, S., 2020. Agent Transparency: A Review of Current Theory and Evidence. *IEEE Trans. Hum.-Mach. Syst.* 50, 215–224. <https://doi.org/10.1109/THMS.2020.2965529>.
- Boissier, P., 2018. Learn the nautical rules of the road: an expert guide to the COLREGS, Second, edition. ed. Fernhurst Books Limited, La Vergne.
- Bowden, V.K., Griffiths, N., Strickland, L., Loft, S., 2021. Detecting a Single Automation Failure: The Impact of Expected (But Not Experienced) Automation Reliability. *Hum. Factors*. <https://doi.org/10.1177/00187208211037188>.
- Burmeister, H.-C., Bruhn, W., Rødseth, Ø.J., Porathe, T., 2014. Autonomous Unmanned Merchant Vessel and its Contribution towards the e-Navigation Implementation: The MUNIN Perspective. *Int. J. E-Navig. Marit. Econ.* 1, 1–13. <https://doi.org/10.1016/j.enavi.2014.12.002>.
- Chaal, M., Ren, X., BahooToroody, A., Basnet, S., Bolbot, V., Banda, O.A.V., Gelder, P.V., 2023. Research on risk, safety, and reliability of autonomous ships: A bibliometric review. *Saf. Sci.* 167, 106256 <https://doi.org/10.1016/j.ssci.2023.106256>.
- Chen, J.Y.C., Procci, K., Boyce, M., Wright, J., Garcia, A., Barnes, M.J., 2014. Situation Awareness-Based Agent Transparency (No. ARL-TR-6905). U.S. Army Research Laboratory, Aberdeen Proving Ground. Doi: 10.21236/ADA600351.
- Christoffersen, K., Woods, D.D., 2002. 1. How to make automated systems team players. In: *Advances in Human Performance and Cognitive Engineering Research*. Emerald Group Publishing Limited, pp. 1–12.
- Cockcroft, A.N., Lameijer, J.N.F., 2011. A guide to the collision avoidance rules: international regulations for preventing collisions at sea, Seventh edition. ed. Elsevier, Butterworth-Heinemann, Amsterdam, [Netherlands].
- Coronato, A., Naem, M., De Pietro, G., Paragliola, G., 2020. Reinforcement learning for intelligent healthcare applications: A survey. *Artif. Intell. Med.* 109, 101964 <https://doi.org/10.1016/j.artmed.2020.101964>.
- Cummings, M.L., Mitchell, P.J., 2007. Operator scheduling strategies in supervisory control of multiple UAVs. *Aerosp. Sci. Technol.* 11, 339–348. <https://doi.org/10.1016/j.ast.2006.10.007>.
- Cummings, M.L., Bruni, S., Mercier, S., Mitchell, P.J., 2007. Automation architecture for single operator, multiple UAV command and control. *Int. C2 J.* 1, 1–24.
- DNV, 2018. The ReVolt - A new inspirational ship concept. URL <https://www.dnv.com/technology-innovation/revolt/> (accessed 5.16.22).
- Elghoneimy, E., Gruver, W.A., 2012. Agent-Based Decision Support and Simulation for Wood Products Manufacturing. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* 42, 1656–1668. <https://doi.org/10.1109/TSMCC.2012.2213809>.
- Endsley, M.R., Kiris, E.O., 1995. The out-of-the-loop performance problem and level of control in automation. *Hum. Factors* 37, 381–394. <https://doi.org/10.1518/001872095779064555>.
- Endsley, M.R., Bolté, B., Jones, D.G., 2003. Designing for situation awareness: an approach to user-centered design. Taylor & Francis, London; New York.
- Endsley, M.R., 2017. From Here to Autonomy: Lessons Learned from Human-Automation Research. *Hum. Factors* 59, 5–27. <https://doi.org/10.1177/0018720816681350>.
- Eriksson, A., Stanton, N.A., 2017. Driving Performance After Self-Regulated Control Transitions in Highly Automated Vehicles. *Hum. Factors* 59, 1233–1248. <https://doi.org/10.1177/0018720817728774>.
- Fan, C., Wróbel, K., Montewka, J., Gil, M., Wan, C., Zhang, D., 2020. A framework to identify factors influencing navigational risk for Maritime Autonomous Surface Ships. *Ocean Eng.* 202 <https://doi.org/10.1016/j.oceaneng.2020.107188>.
- Gegoff, I., Tatasciore, M., Bowden, V., McCarley, J., Loft, S., 2023. Transparent Automated Advice to Mitigate the Impact of Variation in Automation Reliability. *Hum. Factors*. <https://doi.org/10.1177/00187208231196738>.
- Hoem, Å.S., Veitch, E., Vasstein, K., 2022. Human-centred risk assessment for a land-based control interface for an autonomous vessel. *WMU J. Marit. Aff.* 21, 179–211. <https://doi.org/10.1007/s13437-022-00278-y>.
- Hutchinson, J., Strickland, L., Farrell, S., Loft, S., 2022. Human behavioral response to fluctuating automation reliability. *Appl. Ergon.* 105, 103835 <https://doi.org/10.1016/j.apergo.2022.103835>.
- IMO, 1974. International Convention for the Safety of Life at Sea. URL <https://www.imo.org/en/KnowledgeCentre/ConferencesMeetings/Pages/SOLAS.aspx> (accessed 5.10.22).
- IMO, 1977. Convention on the International Regulations for Preventing Collisions at Sea, 1972 (COLREGS).
- IMO, 1978. International Convention on Standards of Training, Certification and Watchkeeping for Seafarers. URL <https://www.imo.org/en/OurWork/HumanElement/Pages/STCW-Conv-LINK.aspx> (accessed 5.10.22).
- IMO, 2011. Resolution A.1047(27) Principles of safe manning.
- IMO, 2018. Maritime Safety Committee (MSC), 100th session, 3-7 December 2018. URL <https://www.imo.org/en/MediaCentre/MeetingSummaries/Pages/MSC-100th-session.aspx> (accessed 5.31.21).
- IMO, 2021. Autonomous ships: regulatory scoping exercise completed. URL <https://www.imo.org/en/MediaCentre/PressBriefings/pages/MASSRSE2021.aspx> (accessed 8.30.23).
- IMO, 2022. Maritime Safety Committee (MSC 105), 20-29 April 2022. URL <https://www.imo.org/en/MediaCentre/MeetingSummaries/Pages/MSC-105th-session.aspx> (accessed 11.5.22).
- ISO, 2019. ISO 9241-210:2019 Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems. ISO.
- Kircher, K., Larsson, A., Hultgren, J.A., 2014. Tactical Driving Behavior With Different Levels of Automation. *IEEE Trans. Intell. Transp. Syst.* 15, 158–167. <https://doi.org/10.1109/TITS.2013.2277725>.
- Kirwan, B., Ainsworth, L.K. (Eds.), 1992. *A Guide to Task Analysis*. Taylor & Francis, London; Washington, DC.
- Kooij, C., Hekkenberg, R., 2020. The effect of autonomous systems on the crew size of ships – a case study. *Marit. Policy Manag.* 1–17 <https://doi.org/10.1080/03088839.2020.1805645>.
- Kretschmann, L., Burmeister, H.C., Jahn, C., 2017. Analyzing the economic benefit of unmanned autonomous ships: An exploratory cost-comparison between an autonomous and a conventional bulk carrier. *Res. Transp. Bus. Manag.* 25, 76–86. <https://doi.org/10.1016/j.rtbm.2017.06.002>.
- Kurt, I., Aymelek, M., 2022. Operational and economic advantages of autonomous ships and their perceived impacts on port operations. *Marit. Econ. Logist.* 24, 302–326. <https://doi.org/10.1057/s41278-022-00213-1>.
- Lee, J.D., See, K.A., 2004. Trust in automation: Designing for appropriate reliance. *Hum. Factors* 46, 50–80. <https://doi.org/10.1518/hfes.46.1.50.30392>.
- Li, Z., Zhang, D., Han, B., Wan, C., 2023. Risk and reliability analysis for maritime autonomous surface ship: A bibliometric review of literature from 2015 to 2022. *Accid. Anal. Prev.* 187, 107090 <https://doi.org/10.1016/j.aap.2023.107090>.
- Loftus, T.J., Filiberto, A.C., Balch, J., Ayzengart, A.L., Tighe, P.J., Rashidi, P., Bihorac, A., Upchurch, G.R., 2020. Intelligent, Autonomous Machines in Surgery. *J. Surg. Res.* 253, 92–99. <https://doi.org/10.1016/j.jss.2020.03.046>.
- Man, Y., Lundh, M., Porathe, T., 2016. Seeking harmony in shore-based unmanned ship handling: From the perspective of human factors, what is the difference we need to focus on from being onboard to onshore? *Hum. Fact. Transp.: Soc. Technol. Evol. Across Maritime, Road, Rail, Aviation Domains.* 61–70. <https://doi.org/10.1201/9781315370460>.
- Man, Y., Weber, R., Cimbritz, J., Lundh, M., MacKinnon, S.N., 2018. Human factor issues during remote ship monitoring tasks: An ecological lesson for system design in a distributed context. *Int. J. Ind. Ergon.* 68, 231–244. <https://doi.org/10.1016/j.ergon.2018.08.005>.
- MITRE, 2018. Human-Machine Teaming Systems Engineering Guide (No. MP180941). MITRE Corporation.
- Miyoshi, T., Fujimoto, S., Rooks, M., Konishi, T., Suzuki, R., 2022. Rules required for operating maritime autonomous surface ships from the viewpoint of seafarers. *J. Navig.* 75, 384–399. <https://doi.org/10.1017/S0373463321000928>.
- Norman, D.A., 1990. The "problem" with automation: inappropriate feedback and interaction, not "over-automation". *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 327, 585–593. <https://doi.org/10.1098/rstb.1990.0101>.
- NYK, 2022. NYK Group Companies Participate in Trial to Simulate the Actual Operation of Fully Autonomous Ship. URL [https://www.nyk.com/english/news/2022/20220303\\_02.html](https://www.nyk.com/english/news/2022/20220303_02.html) (accessed 8.25.22).
- Parasuraman, R., Manzey, D.H., 2010. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Hum. Factors J. Hum. Factors Ergon. Soc.* 52, 381–410. <https://doi.org/10.1177/0018720810376055>.
- Porathe, T., 2014. Remote Monitoring and Control of Unmanned Vessels –The MUNIN Shore Control Centre. *Proc. 13th Int. Conf. Comput. Appl. Inf. Technol. Marit. Ind. COMPIT 14*, 460–467.
- Porathe, T., 2019b. Maritime Autonomous Surface Ships (MASS) and the COLREGS: Do We Need Quantified Rules Or Is "the Ordinary Practice of Seamen" Specific Enough? *TransNav Int. J. Mar. Navig. Saf. Sea Transp.* 13.
- Porathe, T., 2019a. Safety of Autonomous Shipping: COLREGS and Interaction between Manned and Unmanned Ships, in: *Proceedings of the 29th European Safety and Reliability Conference (ESREL)*. Presented at the Proceedings of the 29th European Safety and Reliability Conference (ESREL), pp. 4146–4153. Doi: 10.3850/978-981-11-2724-3\_0655-cd.
- Porathe, T., 2021. No-one in Control: Unmanned Control Rooms for Unmanned Ships?. In: *In: 20th International Conference on Computer and IT Applications in the*

- Maritime Industries. Presented at the COMPIT'21. Hamburg University of Technology, Mülheim, Germany, p. 7.
- Rajabiyazdi, F., Jamieson, G.A., 2020. A Review of Transparency (seeing-into) Models, in: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Presented at the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 302–308. Doi: 10.1109/SMC42975.2020.9282970.
- Ramos, M.A., Utne, I.B., Vinnem, J.E., Mosleh, A., 2018. Accounting for human failure in autonomous ship operations. *Saf. Reliab. - Safe Soc. Chang. World - Proc. 28th Int. Eur. Saf. Reliab. Conf. ESREL 2018 2016*, 355–364. Doi: 10.1201/9781351174664-45.
- Ramos, M.A., Utne, I.B., Mosleh, A., 2019. Collision avoidance on maritime autonomous surface ships: Operators' tasks and human failure events. *Saf. Sci.* 116, 33–44. <https://doi.org/10.1016/j.ssci.2019.02.038>.
- Ramos, M.A., Thieme, C.A., Utne, I.B., Mosleh, A., 2020. A generic approach to analysing failures in human – System interaction in autonomy. *Saf. Sci.* 129, 104808 <https://doi.org/10.1016/j.ssci.2020.104808>.
- Ringbom, H., 2019. Regulating Autonomous Ships—Concepts, Challenges and Precedents. *Ocean Dev. Int. Law* 50, 141–169. <https://doi.org/10.1080/00908320.2019.1582593>.
- Rolls Royce, 2018. Rolls-Royce and Finferries demonstrate world's first Fully Autonomous Ferry. URL <https://www.rolls-royce.com/media/press-releases/2018/03-12-2018-rr-and-finferries-demonstrate-worlds-first-fully-autonomous-ferry.aspx> (accessed 5.16.22)..
- Russell, S.J., Norvig, P., 2022. *Artificial intelligence: a modern approach, Fourth edition, global, edition*. ed. Pearson series in artificial intelligence, Pearson, Harlow.
- Rødseth, Ø.J., Lien Wenersberg, L.A., Nordahl, H., 2021. Towards approval of autonomous ship systems by their operational envelope. *J. Mar. Sci. Technol.* <https://doi.org/10.1007/s00773-021-00815-z>.
- Sheridan, T.B., Verplank, W.L., 1978. Human and Computer Control of Undersea Teleoperators: Defense Technical Information Center. Fort Belvoir, VA. <https://doi.org/10.21236/ADA057655>.
- Sheridan, T.B., 2002. *Humans and Automation: System Design and Research Issues*, 1st ed. Wiley-Interscience.
- Skraaning, G., Jamieson, G.A., 2021. Human Performance Benefits of The Automation Transparency Design Principle: Validation and Variation. *Hum. Factors* 63, 379–401. <https://doi.org/10.1177/0018720819887252>.
- Society of Automotive Engineers, 2021. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles (No. J3016\_202104).
- Statheros, T., Howells, G., Maier, K.M., 2008. Autonomous Ship Collision Avoidance Navigation Concepts, Technologies, and Techniques. *J. Navig.* 61, 129–142. <https://doi.org/10.1017/S037346330700447X>.
- Strauch, B., 2018. Ironies of Automation: Still Unresolved After All These Years. *IEEE Trans. Hum.-Mach. Syst.* 48, 419–433. <https://doi.org/10.1109/THMS.2017.2732506>.
- Thieme, C.A., Utne, I.B., 2017. A risk model for autonomous marine systems and operation focusing on human–autonomy collaboration. *Proc. Inst. Mech. Eng. Part O J. Risk Reliab.* 231, 446–464. <https://doi.org/10.1177/1748006X17709377>.
- van de Merwe, K., Mallam, S., Nazir, S., 2022. Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review. *Hum. Factors.* <https://doi.org/10.1177/001872082211077804>.
- Van de Merwe, K., Mallam, S.C., Engelhardt, Ø., Nazir, S., 2022. Exploring navigator roles and tasks in transitioning towards supervisory control of autonomous collision avoidance systems. *J. Phys.: Conf. Ser.* 2311, 012017 <https://doi.org/10.1088/1742-6596/2311/1/012017>.
- van de Merwe, K., Mallam, S., Engelhardt, Ø., Nazir, S., 2023. Operationalising Automation Transparency for Maritime Collision Avoidance. *TransNav Int. J. Mar. Navig. Saf. Sea Transp.* 17 <https://doi.org/10.12716/1001.17.02.09>.
- Veitch, E., Alsos, O.A., 2022. A systematic review of human-AI interaction in autonomous ship systems. *Saf. Sci.* 152, 105778 <https://doi.org/10.1016/j.ssci.2022.105778>.
- Veitch, E., Christensen, K., Log, M., Valestrand, E., Hilmo Lundheim, S., Nesse, M., Alsos, O., Steinert, M., 2022. From captain to button-presser: operators' perspectives on navigating highly automated ferries. *J. Phys. Conf. Ser.* 2311, 012028 <https://doi.org/10.1088/1742-6596/2311/1/012028>.
- Ventikos, N.P., Chmurski, A., Louzis, K., 2020. A systems-based application for autonomous vessels safety: Hazard identification as a function of increasing autonomy levels. *Saf. Sci.* 131, 104919 <https://doi.org/10.1016/j.ssci.2020.104919>.
- Vojković, G., Milenković, M., 2019. Autonomous ships and legal authorities of the ship master. *Case Stud. Transp. Policy.* Doi: 10.1016/j.cstp.2019.12.001.
- Wickens, C.D., Clegg, B.A., Vieane, A.Z., Sebok, A.L., 2015. Complacency and Automation Bias in the Use of Imperfect Automation. *Hum. Factors J. Hum. Factors Ergon. Soc.* 57, 728–739. <https://doi.org/10.1177/0018720815581940>.
- Wickens, C.D., 2018. Automation Stages & Levels, 20 Years After. *J. Cogn. Eng. Decis. Mak.* 12, 35–41. <https://doi.org/10.1177/1555343417727438>.
- Wohleber, R.W., Matthews, G., Lin, J., Szalma, J.L., Calhoun, G.L., Funke, G.J., Chiu, C.-Y.-P., Ruff, H.A., 2019. Vigilance and Automation Dependence in Operation of Multiple Unmanned Aerial Systems (UAS): A Simulation Study. *Hum. Factors J. Hum. Factors Ergon. Soc.* 61, 488–505. <https://doi.org/10.1177/0018720818799468>.
- Wright, J.L., Chen, J.Y.C., Barnes, M.J., Hancock, P.A., 2016. The Effect of Agent Reasoning Transparency on Automation Bias: An Analysis of Response Performance. In: Lackey, S., Shumaker, R. (Eds.), *Virtual, Augmented and Mixed Reality*. Springer International Publishing, Cham, pp. 465–477.
- Wróbel, K., Montewka, J., Kujala, P., 2017. Towards the assessment of potential impact of unmanned vessels on maritime transportation safety. *Reliab. Eng. Syst. Saf.* 165, 155–169. <https://doi.org/10.1016/j.res.2017.03.029>.
- Yara International, 2022. Crown Prince and youths christen world's first emission-free container ship. URL <https://www.yara.com/corporate-releases/crown-prince-and-youths-christen-worlds-first-emission-free-container-ship/> (accessed 5.16.22).
- Zhou, X.-Y., Huang, J.-J., Wang, F.-W., Wu, Z.-L., Liu, Z.-J., 2020. A Study of the Application Barriers to the Use of Autonomous Ships Posed by the Good Seamanship Requirement of COLREGS. *J. Navig.* 73, 710–725. <https://doi.org/10.1017/S0373463319000924>.



**Article 3**

van de Merwe, K., Mallam, S., Engelhardtsen, Ø., & Nazir, S. (2023). Towards an approach to define transparency requirements for maritime collision avoidance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 67(1), 483–488.  
<https://doi.org/10.1177/21695067231192862>



# Towards an approach to define transparency requirements for maritime collision avoidance

Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2023, Vol. 67(1) 483–488  
Copyright © 2023 Human Factors and Ergonomics Society  
DOI: 10.1177/21695067231192862  
journals.sagepub.com/home/pro



Koen van de Merwe<sup>1,2</sup>, Steven Mallam<sup>2,3</sup>, Øystein Engelhardtson<sup>1</sup>, and Salman Nazir<sup>2</sup>

## Abstract

This study discusses an approach to support human supervision of autonomous maritime collision avoidance systems by disclosing the system's perceived information, internal reasoning, decisions, and planned actions as layers of transparency. Information requirements, identified through a cognitive task analysis, were structured using the information processing model by Parasuraman, Sheridan, and Wickens (2000). This model was contextualized to the maritime collision avoidance setting such that the information from the analysis could be structured into unique and distinct layers. A set of minimum information requirements was identified depicting the system's decisions and planned action, supported by additional layers to reveal its internal reasoning. This approach aims at supporting humans in effectively supervising autonomous collision avoidance systems in their operational context by providing understandability and predictability about what the system is doing, why it is doing it, and what it will do next, i.e., transparency.

## Keywords

Automation Transparency, Collision avoidance, Human Machine Interaction, Agent transparency, Cognitive task analysis, Supervisory control

## Introduction

Continuous technological development is pushing the boundaries of automation capabilities in the maritime industry. The deployment of advanced technologies is envisioned to be able to allow unmanned ships to navigate from port to port without human intervention, avoiding collisions along the way. As collision avoidance is a complex and multi-faceted aspect of navigation, solving this problem is critical to realizing autonomous shipping. Consequently, autonomous collision avoidance has received much focus in recent years (Aylward et al., 2022; Miyoshi et al., 2022; Ramos et al., 2019).

At present, collision avoidance manoeuvring is a task that relies primarily on human performance. Navigators, located on the ship's bridge, determine the presence of collision risk during the ship's journey and perform avoidance manoeuvres when needed. On most modern ships, navigators are supported by a range of instrumentation and control systems to ensure traffic is detected early and effective avoidance manoeuvres are executed in accordance with the "rules of the road" for maritime traffic, i.e., the collision regulations (IMO, 1977).

Autonomous collision avoidance systems are envisioned to perform this task by having the capability to perceive their

surroundings, estimate collision risks, decide on how to resolve a collision situation, and execute timely avoidance manoeuvres. However, as there are challenges related to developing reliable automated collision avoidance systems that can resolve conflict situations under all circumstances, most autonomy concepts use some kind of human oversight to monitor the system's performance (Mackinnon et al., 2015; Pietrzykowski et al., 2017; Wróbel et al., 2022). However, there are well-known issues related to the role of human supervisors of automated systems that puts the feasibility of this vision into question (e.g., Endsley, 2017; Onnasch et al., 2014; Strauch, 2018).

Some of the main challenges with supervising autonomous systems is the impact on operator's situation awareness (SA) and the ability to successfully intervene when

<sup>1</sup>Group Research and Development, DNV, Høvik, Norway

<sup>2</sup>Department of Maritime Operations, University of South-Eastern Norway, Borre, Norway

<sup>3</sup>Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, Canada

## Corresponding Author:

Koen van de Merwe, Group Research and Development, DNV, Høvik, Akershus, 1322, Norway.

Email: Koen.van.de.Merwe@dnv.com

needed (Endsley & Kiris, 1995). A key factor contributing to facilitating SA is the feedback the system gives to the human user (Norman, 1990). When a system does not provide sufficient and useful information to the supervisor, the human is left with an incomplete understanding with regards to the system under control (Endsley, 2017). Therefore, disclosing information with regards to the system's actions and behaviours becomes an important factor in building and maintaining situation awareness where humans take a supervisory role of autonomous systems (Endsley, 2023). In other words, by making the reasoning of a system available to its user, the system should allow for understandability of its decisions, predictability of its actions, and consequently support human supervision of such systems (Chen et al., 2014, 2018; Christoffersen & Woods, 2002; Endsley, 2017; Lyons, 2013).

Automation transparency is described as the ability of a system to convey what it is doing, why it is doing it, and what it will do next (Endsley, 2017). Recent reviews have explored the relationship between automation transparency and human performance variables based on the findings from several empirical studies from various safety critical domains (Bhaskara et al., 2020; Rajabiyazdi & Jamieson, 2020; van de Merwe, Mallam, & Nazir, 2022). A promising effect of transparency was identified in terms of increased operator situation awareness, and task performance, without adding additional mental workload. Applying these findings to the autonomous shipping domain implies that providing human supervisors of autonomous collision avoidance systems with insight into the reasoning process behind the system's decisions and actions should enhance system understanding without mentally burdening the supervisor. However, given the lack of maritime-specific empirical evidence in the abovementioned studies, it is unclear what the specific implications are for transparency of autonomous collision avoidance systems for this domain.

To consider this gap, a recent study led by the first author focused on establishing information requirements for supervising a hypothetical autonomous collision avoidance systems using a cognitive task analysis (van de Merwe et al., under review; van de Merwe, Mallam, Engelhardt, et al., 2022). The analysis made explicit which information elements a supervisor would need to effectively perform the supervisory task. However, given the dynamic nature of collision avoidance, the information needed to supervise the system may vary given the circumstances.

To address this, the current study examines how the use of a layered approach, based on an information processing model, can be used as a means for the supervisor to scrutinize the system's internal reasoning. The method discloses the system's information acquisition, -analysis, decision making, and action implementation in a stepwise manner, such that understandability and predictability are supported in a dynamic setting.

## Method

To determine understandability and predictability in a collision avoidance context, a method was established that consisted of three steps. First, the data from the cognitive task analysis was used as input to this study. Second, the information processing model was adapted to the collision avoidance context. Third, the information to be disclosed to the human supervisor was generated (see Figure 1).

First, the information requirements described in Van de Merwe et al. (under review; 2022) provided the information elements a supervisor requires to understand the system's reasoning in the collision avoidance context. In contrast with conventional collision avoidance, the primary task of the supervisor is to verify the collision avoidance system's functioning through an assessment of the information provided by it. As such, the cognitive task analysis provided the information requirements for determining whether the system was performing its task in accordance with the collision regulations.

Second, the information processing model by Parasuraman, Sheridan, and Wickens ("PSW model") was used as a framework to represent the inner reasoning of the system (Parasuraman et al., 2000). The PSW model was originally developed to represent human information processing steps and to assist in function allocation between humans and systems. For this study, it was assumed that all functions are performed by a hypothetical collision avoidance system, and that the human takes the role of a supervisor. As such, the model was adapted to this context allowing the categories of information belonging to each stage of the system's information processing to be defined and consequently, which type of information the system to be provided to the supervisor.

Third, the information requirements identified in the task analysis were structured and organised according to the categories of the PSW model. As a result, the model was used to make the individual information processing stages of the collision avoidance system visible, and thereby support understandability and predictability of its actions.

## Results

### *Establishing information needs for human supervision*

The cognitive task analysis compared two cases: a conventional collision avoidance, and a supervision case where a human took the role of a supervisor of an autonomous collision avoidance system (see Figure 2). Based on a series of in situ interviews, in situ observations, a review of ferry operators' procedures, and a detailed and systematic analysis of the collision regulations, a set of goals, decisions, tasks, and information needs were established for conventional collision avoidance maneuvering. These results were subsequently



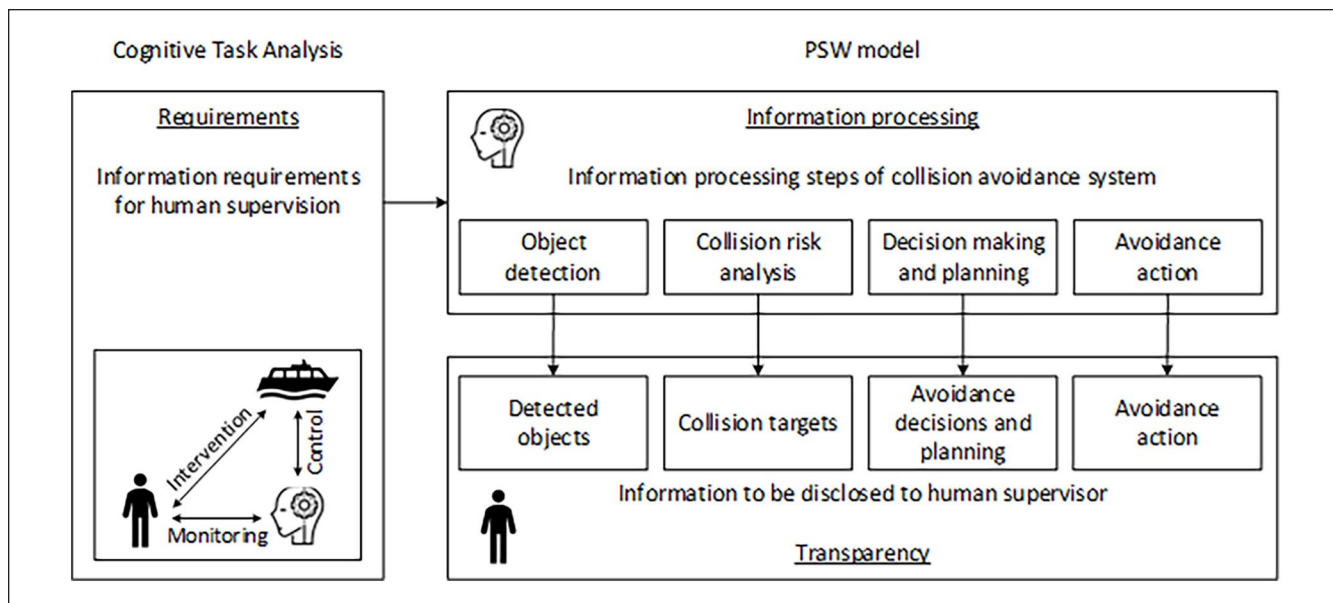


Figure 1. The framework for the analysis.

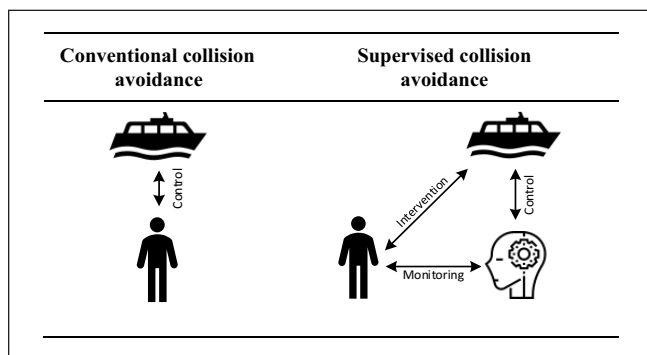


Figure 2. Conventional and supervised collision avoidance.

extrapolated to the supervised collision avoidance case. The detailed results are discussed in Van de Merwe et al. (2022) and Van de Merwe et al. (under review).

In the supervised collision avoidance case, the supervisor’s primary role is to verify that the system’s actions are in accordance with the collision regulations, i.e., that objects are accurately detected early, an adequate risk assessment is performed, an avoidance maneuver is planned, and safely executed. Since the supervisor is only monitoring the collision avoidance system’s actions, the supervisor is dependent on the information coming from the system to perform this verification task.

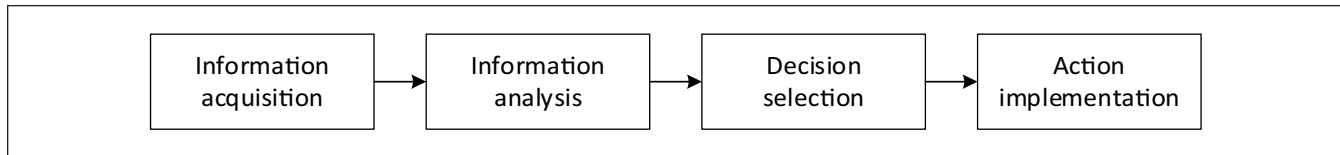
The analysis identified the need for the system to convey its perception of the situation, its identified collision risks herein, and its plans to resolve these. This includes, amongst others, depicting the location of detected terrain and objects in the short- and long-term, terrain, estimated collision risks and type (i.e., crossing, head-on, overtaking/overtaken), information about target vessel type and size, (time to

closest point of approach for risk targets, what actions it intends to perform against these targets, when it will perform these actions, its own priority against risk targets (give-way or stand-on), the target ship’s predicted track, and any limitations to sea-room it has identified that will affect its maneuvering capabilities. Given the dynamic- and potentially complex nature of maritime traffic situations, conveying this amount of information may prove to be challenging to effectively depict on a Human Machine Interface (HMI). Also, regardless of the graphical presentation, making sense of a continuously changing stream of information, which the supervisor is only monitoring, is a challenging task in itself. As such, to determine how the system can provide understandability and observability of its actions whilst limiting the information conveyed to the supervisor, the PSW model was used to provide structure to this approach.

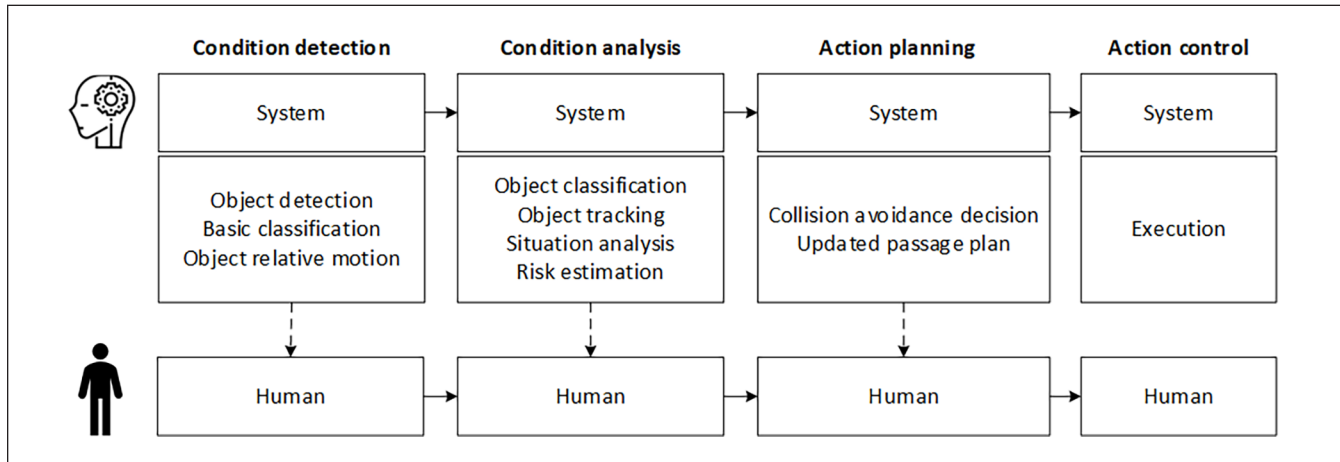
### Information processing in autonomous collision avoidance

In the PSW model, information processing is assumed to take place in four stages: information acquisition, information analysis, decision selection, and action implementation (see Figure 3, Parasuraman et al., 2000).

The initial information processing stage refers to the step where information is acquired, registered, and initial data processing is performed. Adapting this model to the collision avoidance context, it can be inferred that in this stage the system determines the presence of objects, provides basic information regarding these objects, such as whether the object is moving or stationary, and provides own ship relative motion with regards to these objects (e.g., whether these objects are crossing, head-on, or overtaking).



**Figure 3.** A model for human information processing (adopted from Parasuraman et al., 2000).



**Figure 4.** Adapting the PSW model to a collision avoidance context in which a human supervises a collision avoidance system. The dashed arrows indicate information flow from the system to the supervisor.

In the information analysis stage, the analytical and inferential processes are performed, and algorithms are applied to allow for extrapolation of the data over time, such that predictions can be made about future states. In our case, this means that target objects are tracked (i.e., followed over time), classified (e.g., in terms of ship types and their manoeuvrability), and future states are predicted (e.g., based on target course and speed). Based on this analysis, the collision avoidance system can determine the risk of collision with own ship.

In the decision selection stage solution alternatives are determined and decisions are made with regards to the optimal solution among these. For the collision avoidance context, this means that actions are planned based on the outcomes of the risk analysis with which the risk of collision can best be avoided whilst adhering to the collision regulations. This includes determining own ship- and target ship priority (e.g., give-way or stand-on) and the changes to course and speed needed (if any) to clearly indicate own ship's intention to avoid collision.

Finally, the action implementation stage refers to the execution of a response or action that is consistent with the decision choice, i.e., the actual execution of the action. For our context, this means that the collision avoidance system sends the parameters of the updated passage plan to the control- and machinery system for effectuation. Limited information processing is performed in this stage as it is primarily concerned with the execution of decisions made earlier.

Adapting the PSW model to the collision avoidance context allowed the model to be used as a method to categorize the information from the cognitive task analysis. This way, a contextualised adaptation of the theoretical model was used to assist in creating structure in the information requirements for supervising the collision avoidance system (see Figure 4).

### *A layered approach to information disclosure*

Table 1 describes excerpts of the information requirements that were obtained in the task analysis and that were structured using the contextualised PSW model. As a result, a set of information elements was derived, per stage of information processing, that depicts the information to be provided to the supervisor of a collision avoidance system.

The results show that each of the information processing stages of the system can be depicted using a unique set of information parameters such that each stage is clearly distinguishable. Also, each set of information elements are additive in terms of that they convey information that other elements do not. For example, in the condition detection stage, the purpose of the information is to provide insight into what the system has detected in its surroundings and how this is interpreted, including some rudimentary information processing. Therefore, the information provided here aims to convey how the system perceives the world.

The aim of the information provided in the condition analysis stage is to convey how the system interprets this

**Table 1.** Applying the PSW model to information requirements from the task analysis. Key: OT= overtaking/overtaken, HO=head-on, CR=crossing, GW=give-way, SO=stand-on.

Information processing stage	Information requirements for supervision (excerpts)
<b>1. Condition detection:</b>	
- Object detection	- Detected objects short & long range
- Basic classification	- Identified target ship
- Object relative motion	- Target object type and size
	- Identified target object as OT/HO/CR
	- Uncertainties in the radar/ sensor data
	- Status of sensors
<b>2. Condition analysis:</b>	
- Object classification	- Objects that pose risk
- Object tracking	- Plotted objects
- Situation analysis	- Risk object type and size
- Risk estimation	- Risk object priority
	- Risk object course and speed
	- Risk object intended trajectory
	- Risk object conflict type
	- Safe speed parameters
<b>3. Action planning:</b>	
- Collision avoidance decision	- Own ship priority (GW/SO)
- Updated passage plan	- Target vessel priority (GW/SO)
	- Own ship intended track and speed
<b>4. Action control:</b>	
- Execution of plan	N/A: only action implementation

information. Here, the primary focus lies on conveying own ship collision risk with other vessels by estimating future trajectories of target vessels. Hence, the information depicted here aims to convey how the system understands the world around it in terms of collision risk.

The aim of the action planning stage is to convey how the system sees the solution to the collision risk situation. Here, priorities are determined guiding the decision to give-way or stand-on. Finally, and most importantly, here it is conveyed what actions the system intends to perform, including if this means keeping current course and speed. Hence, in this phase, the system conveys how it plans to safely manoeuvre through its surroundings given its interpretation of it.

The final stage, action control, aims to convey that the system is executing the plan. Since there is limited information processed in this stage, i.e., the execution of the plan is monitored only and any deviations from the original plan, detected, analysed, decided upon and planned in the preceding stages, there is limited information depicted to the supervisor in this stage.

## Discussion

This paper described the rationale and approach to derive information requirements to disclose the inner reasoning of autonomous collision avoidance systems to human supervisors. The PSW model was used to provide a framework to drive the categorisation of the information requirements from the task analysis into unique and distinguishable categories.

Given the dynamic nature of the collision avoidance task, a flexible information solution is needed that supports

supervisors in attaining and maintaining SA of the autonomous system and the environment it is operating in. Passive information presentation depicting the reasoning of the collision avoidance system may not adequately represent the dynamic nature of the collision avoidance task and may therefore not fully support the supervisor herein. As such, the structuring of system reasoning into layers of transparency allows the information provision to be adjusted to its context. That is, depending on the needs of the task, the layers allow for adjusting the degree of transparency depending on the level of understandability and predictability required for effective supervision.

An approach to determining which degree of transparency supports which degree of understandability and predictability can be derived from the cognitive task analysis. That is, although the model describes the different layers of information processing pertaining to the collision avoidance system, a supervisor may choose to prioritise one aspect of the system's processing over another depending on the task and its context. For example, depending on the complexity of the traffic situation, a supervisor may only be interested in the system's decisions and actions without requiring understanding of the underlying reasoning. Conversely, a supervisor may want to understand why the system has chosen a particular solution and as a result, requires a full overview of the system's sensor input for this. Consequently, the supervision task executed in the dynamic context of collision avoidance, determines which layer of transparency take precedence over another. Still, an attempt is made here to argue for a minimum level of transparency to support the supervision task.

A potential starting point is to determine which information a supervisor would like to know as a minimum to perform the supervision task. Given the collision avoidance context, a minimum degree of information a supervisor requires is to determine whether the system is capable of resolving a collision situation at all. Understanding the intentions of the collision avoidance system requires information regarding the system's decisions and planned actions. In other words, by depicting the system's decisions (i.e., give-way or stand-on) and its actions (i.e., intended changes to its course and speed) a supervisor should be able to verify the system's intentions regarding the collision situation. Consequently, the action planning stage should be able to provide a level of information sufficient to comprehend the primary intentions of the system, and therefore could serve as the minimum degree of transparency for this type of system.

Additional information from the information analysis, and information acquisition stages may be made available to the supervisor should the task, and its context, require this. As such, a cumulative approach, starting at the action planning stage, may be used to "dig deeper" into the reasoning process behind the system's decisions and planned actions. This way, each level allows for increased understandability of the system's decision making and action implementation

by making the system's information inputs and internal reasoning incrementally observable.

## Conclusion

By organising the information from the cognitive task analysis using the context specific PSW model, a set of information layers were defined for supervision of the autonomous collision avoidance system that is adaptable to its dynamic context. Using this approach, HMIs can be developed that support supervisors in understanding the system by providing information depending on user needs in the given situation. Since each of the information processing steps reflects the system's internal reasoning, the layered approach driven by the PSW model enables a structured means to create transparent systems.

## Acknowledgments

This research is sponsored by the Research Council of Norway, project nr. 311365.

## References

- Aylward, K., Weber, R., Lundh, M., MacKinnon, S. N., & Dahlman, J. (2022). Navigators' views of a collision avoidance decision support system for maritime navigation. *Journal of Navigation*, 1–14. <https://doi.org/10.1017/S0373463322000510>
- Bhaskara, A., Skinner, M., & Loft, S. (2020). Agent Transparency: A Review of Current Theory and Evidence. *IEEE Transactions on Human-Machine Systems*, 50(3), 215–224. <https://doi.org/10.1109/THMS.2020.2965529>
- Chen, J. Y. C., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. J. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, 19(3), 259–282. <https://doi.org/10.1080/1463922X.2017.1315750>
- Chen, J. Y. C., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. J. (2014). *Situation Awareness-Based Agent Transparency* (ARL-TR-6905). U.S. Army Research Laboratory. <https://doi.org/10.21236/ADA600351>
- Christoffersen, K., & Woods, D. D. (2002). 1. How to make automated systems team players. In *Advances in Human Performance and Cognitive Engineering Research (Vol. 2, pp. 1–12)*. Emerald Group Publishing Limited. [https://doi.org/10.1016/S1479-3601\(02\)02003-9](https://doi.org/10.1016/S1479-3601(02)02003-9)
- Endsley, M. R. (2017). From Here to Autonomy: Lessons Learned from Human-Automation Research. *Human Factors*, 59(1), 5–27. <https://doi.org/10.1177/0018720816681350>
- Endsley, M. R. (2023). Supporting Human-AI Teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior*, 140, 107574. <https://doi.org/10.1016/j.chb.2022.107574>
- Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors*, 37(2), 381–394. <https://doi.org/10.1518/001872095779064555>
- IMO. (1977). *Convention of the international regulations for preventing collisions at sea (COLREGS)*. International Maritime Organisation.
- Lyons, J. B. (2013). Being transparent about transparency: A model for human-robot interaction. *2013 AAAI Spring Symposium Series*.
- Mackinnon, S. N., Man, Y., Lundh, M., & Porathe, T. (2015). Command and control of unmanned vessels: Keeping shore based operators in-the-loop. *18th International Conference on Ships and Shipping Research, NAV 2015*, 612–619.
- Miyoshi, T., Fujimoto, S., Rooks, M., Konishi, T., & Suzuki, R. (2022). Rules required for operating maritime autonomous surface ships from the viewpoint of seafarers. *The Journal of Navigation*, 75(2), 384–399. <https://doi.org/10.1017/S0373463321000928>
- Norman, D. A. (1990). The “problem” with automation: Inappropriate feedback and interaction, not “over-automation.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 327(1241), 585–593. <https://doi.org/10.1098/rstb.1990.0101>
- Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 56(3), 476–488. <https://doi.org/10.1177/0018720813501549>
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, 30(3), 286–297. <https://doi.org/10.1109/3468.844354>
- Pietrzykowski, Z., Wolejsza, P., & Borkowski, P. (2017). Decision Support in Collision Situations at Sea. *The Journal of Navigation*, 70(3), 447–464. <https://doi.org/10.1017/S0373463316000746>
- Rajabiyazdi, F., & Jamieson, G. A. (2020). A Review of Transparency (seeing-into) Models. *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 302–308. <https://doi.org/10.1109/SMC42975.2020.9282970>
- Ramos, M. A., Utne, I. B., & Mosleh, A. (2019). Collision avoidance on maritime autonomous surface ships: Operators' tasks and human failure events. *Safety Science*, 116, 33–44. <https://doi.org/10.1016/j.ssci.2019.02.038>
- Strauch, B. (2018). Ironies of Automation: Still Unresolved After All These Years. *IEEE Transactions on Human-Machine Systems*, 48(5), 419–433. <https://doi.org/10.1109/THMS.2017.2732506>
- van de Merwe, K., Mallam, S. C., Engelhardt, Ø., & Nazir, S. (2022). Exploring navigator roles and tasks in transitioning towards supervisory control of autonomous collision avoidance systems. *Journal of Physics: Conference Series*, 2311(1), 012017. <https://doi.org/10.1088/1742-6596/2311/1/012017>
- van de Merwe, K., Mallam, S., Engelhardt, Ø., & Nazir, S. (under review). *Supporting human supervision in autonomous collision avoidance through system transparency: A structured and systematic approach*.
- van de Merwe, K., Mallam, S., & Nazir, S. (2022). Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review. *Human Factors*, 00187208221077804. <https://doi.org/10.1177/00187208221077804>
- Wróbel, K., Gil, M., Huang, Y., & Wawruch, R. (2022). The Vagueness of COLREG versus Collision Avoidance Techniques—A Discussion on the Current State and Future Challenges Concerning the Operation of Autonomous Ships. *Sustainability*, 14(24), Article 24. <https://doi.org/10.3390/su142416516>

**Article 4**

van de Merwe, K., Mallam, S., Engelhardtsen, Ø., & Nazir, S. (2023). Operationalising Automation Transparency for Maritime Collision Avoidance. *TransNav, International Journal on Marine Navigation and Safety of Sea Transportation*, 17(2). <https://doi.org/10.12716/1001.17.02.09>



# Operationalising Automation Transparency for Maritime Collision Avoidance

K. van de Merwe<sup>1,2</sup>, S. Mallam<sup>2,3</sup>, Ø. Engelhardtson<sup>1</sup> & S. Nazir<sup>2</sup>

<sup>1</sup> DNV, Høvik, Norway

<sup>2</sup> University of South-Eastern Norway, Borre, Norway

<sup>3</sup> Memorial University of Newfoundland, St. John's, Canada

**ABSTRACT:** Automation transparency is a means to provide understandability and predictability of autonomous systems by disclosing what the system is currently doing, why it is doing it, and what it will do next. To support human supervision of autonomous collision avoidance systems, insight into the system's internal reasoning is an important prerequisite. However, there is limited knowledge regarding transparency in this domain and its relationship to human supervisory performance. Therefore, this paper aims to investigate how an information processing model and a cognitive task analysis could be used to drive the development of transparency concepts. Also, realistic traffic situations, reflecting the variation in collision type and context that can occur in real-life, were developed to empirically evaluate these concepts. Together, these activities provide the groundwork for exploring the relation between transparency and human performance variables in the autonomous maritime context.

## 1 INTRODUCTION

### 1.1 *Human supervision in autonomous collision avoidance*

The last decade has shown an increasing interest in research and development efforts towards use of autonomy in the maritime industry. The purpose of increased automation is diverse, but improvements in cost, efficiency and safety for sharp-end personnel are major drivers [1]–[3]. Yara Birkeland, and the ASKO barges are examples of the ambition of the industry when it comes to the application of highly automated functions to support and/or substitute onboard personnel [4], [5]. In this development, remote-control centres are foreseen to play a role from where operators can perform oversight of autonomous ships and can make critical decisions with regards to the operations of the ship [6].

The purpose of remote-control centres is to provide shore-side support for autonomous ships, to be compliant with current regulations on minimum safe manning, and to provide an equivalent level of safety (or better) compared to conventional ship operations [7], [8]. The idea is that from a remote-control position operators can supervise the ship's operations and monitor, assist, and take over from the autonomous systems when the circumstances require this. In this case, it is assumed that humans can perceive and understand the information concerning the ship under supervision such that adequate situation awareness can be attained and maintained.

A key challenge to be resolved in moving towards autonomous, and potentially unmanned, shipping is how unforeseen circumstances, such as collision and grounding situations, are handled without the presence of navigators onboard the ship [9]–[12]. At present, navigators determine collision risk and

perform relevant avoidance manoeuvres supported by a range of systems, e.g., radar, AIS, and ECDIS. Also, collision and grounding avoidance requires knowledge, skills, and experience to be performed in accordance with the collision regulations. When this task is performed by an autonomous Artificial Intelligence-powered collision avoidance system, adequate and sufficient contextual information is essential to support human oversight (see Figure 1) [13].

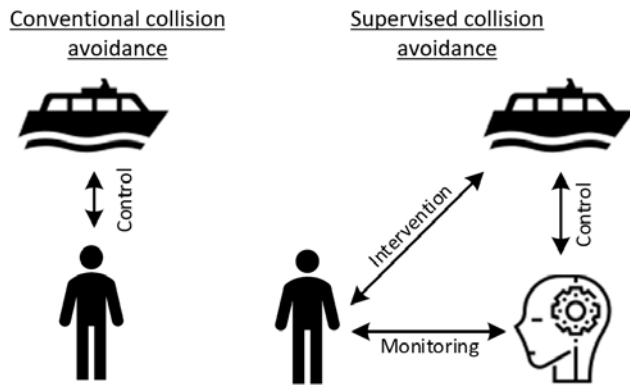


Figure 1. Conceptualization of control in conventional- and supervised collision avoidance.

An earlier study led by the first author identified the information required to supervise the performance of an autonomous collision avoidance system through a mapping and assessment of relevant cognitive tasks [12], [14]. This study concluded that adequately supervising an autonomous collision and grounding avoidance system requires insight into the system's information processing to understand its decisions and actions. Based on the knowledge that human supervision of automated functions has challenges in terms of human performance, keeping humans in the loop, or rather "on the loop", becomes an essential design requirement [15], [16]. Thus, providing sufficient information about the automated system's reasoning process has been proposed as one of the elements that could support humans in such a role. In other words, by disclosing the system's internal decision-making process to its supervisor, the system is made transparent with regards to its intent, performance, future plans, and reasoning process [17].

Automation transparency is concerned with making the inner reasoning of systems observable, such that its actions are understandable and predictable [15], [18], [19]. Therefore, transparency should make it clear to human supervisors what the system is currently doing, why it is doing it, and what it will do next [15]. Earlier reviews have indicated that transparency has a promising effect on human performance and situation awareness [20]–[22]. However, there is limited knowledge regarding transparency in the maritime domain, especially in relation to autonomous collision and grounding avoidance. To this end, further work is needed to investigate the role of transparency in supervised autonomous shipping and to explore its relationship with human performance in this context.

This paper discusses ongoing work towards performing an empirical evaluation to study differing levels and types of transparency concepts in a realistic

traffic collision avoidance setting. An empirical evaluation is planned in which participants take the role of a supervisor of an autonomous collision avoidance system. An approach is used in which participants are tasked with evaluating traffic situations for their understandability, whilst being measured on human performance variables. The purpose of this evaluation is to better understand which levels and types of transparency information support human supervisors and how this knowledge can be applied to a dynamic collision avoidance context. This paper describes the groundwork for this evaluation by describing the systematic development process behind the traffic situations, as well as the levels and types of transparency concepts developed for this.

## 2 DEVELOPING TRAFFIC SITUATIONS

### 2.1 Defining criteria to ensure variation

To provide participants of the planned empirical evaluation with realistic conflicts, traffic situations were developed that reflected the variation in collision type and context that may occur in real-life. Also, to avoid familiarisation with the traffic situations, and thereby unintentionally influencing the results of the evaluation, multiple variants of traffic situations were developed based on a set of criteria (see Table 1).

Table 1. Criteria for establishing a varied set of traffic situations.

Criterion	Variation
Complexity avoidance manoeuvre own ship	Low - No limitations High - Limitations manoeuvre
Collision type	CR - Crossing HO - Head-on OT - Overtaking/overtaken NC - No collision
Avoidance actions own ship	Give-way Stand-on
Restrictions target	No restrictions Restricted in manoeuvrability
Traffic density	Few other ship and objects Many other ships and objects
Geography	Land Open water

Variability was ensured through differing levels of complexity, collision types, the avoidance actions of own ship, restrictions to target ships, traffic density, and geography. First, in high complex situations, own ship was restricted in its avoidance manoeuvring ability compared to low complex situations. That is, in low complexity situations, own ship was free to manoeuvre in any direction to avoid a collision, whereas in high complexity situations, there were obstacles prohibiting own ship to perform certain manoeuvres. Second, for collision type, traffic situations consisted of crossing-, head-on-, overtaking-/overtaken situations. Also, situations were developed in which no collision was present. Third, for avoidance actions, situations were developed for which own ship was the give-way vessel or the stand-on vessel. Fourth, for some situations, target ships were restricted in their manoeuvrability, e.g., because of ongoing bunkering.



Fifth, situations were developed with low- and high traffic densities. Finally, traffic situations were developed in which contextual factors were varied that were external to the traffic situations (i.e., land formations or open water).

To constrain the amount of variation and retain controllability in the traffic situations, some limitations were set in terms of number of ships posing a collision risk and the number of simultaneous collision situations. That is, own ship could only be in direct conflict with one other ship for one collision type (e.g., not in a crossing and head-on situation simultaneously), own ship could not be in both a give-way and stand-on situation simultaneously, and own ship was never restricted in its manoeuvrability. Also, although it is recognised that grounding avoidance is an essential part of collision avoidance, the traffic situations in this paper were limited to collision situations only. Finally, external factors that could affect the collision situation or own-ship's capabilities, such as weather or technical failures, were not included.

## 2.2 Development process

For each criterion in Table 1, two scenarios were created resulting in a set of 70 situations (see Table 2). The traffic situations were created in a desktop simulator from a popular equipment manufacturer by a navy-certified navigator with five years of navigational experience. Upon creating an initial set of traffic situations, a review was performed with independent navigators.

Table 2. The traffic situations created based on the set of criteria. Key: HO = Head-on, CR = Crossing, OT = Overtaking/overtaken, NC = No collision, L = Low, H = High, T = Total. \*Note: in a head-on situation with one motorised target ship and no other exceptions, own ship cannot be stand-on.

Variant/Complexity	HO		CR		OT		T
	L	H	L	H	L	H	
Type (HO/CR/OT)	5	5	4	4	4	4	26
Type (NC)	2	2	2	2	2	2	12
Own ship stand-on*	0	0	2	2	2	2	8
Restrictions target	2	2	2	2	2	2	12
Geography (land)	2	2	2	2	2	2	12
Total	11	11	12	12	12	12	70

## 2.3 Verification and validation workshop

The final verification and validation were performed with two independent navigators holding active navigational licenses (D1/D2), with an average of 6.5 years of navigational experience (SD=2.1, min=5, max=8). The review was performed in the form of a 1,5-day workshop.

In the workshop traffic situations were shown on a display and participants were asked to state if own ship was in a collision situation, if yes, which type (HO/CR/OT), and the avoidance action required by own ship (give-way/ stand-on). In addition, three questions were asked, using a 7-point Likert scale, probing the situation's realism, complexity, and likelihood of occurrence. With these questions, a comparison between the situation's intended

depiction and the navigator's perception was obtained. Discrepancies were discussed and suggestions for improving the design of the traffic situations were noted. A final set of traffic situations were produced, incorporating the inputs from the workshop (see Figures 2, 3, 4, and 5 for examples).

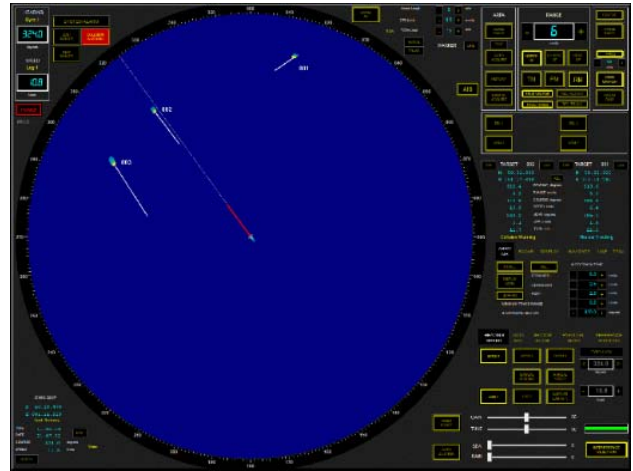


Figure 2. Own ship is in a head-on situation in open water where it is required to give-way. The situation is of low complexity as there are no restrictions to own ship's avoidance manoeuvrability.



Figure 3. Own ship is in an overtaking situation in open waters where it is required to give-way. The situation is of high complexity as there are restrictions to own ship's avoidance manoeuvrability (both port and starboard).



Figure 4. Own ship is in a crossing situation in open waters where it is required to stand-on. The situation is of high complexity as there are restrictions to the target ship's avoidance manoeuvrability (the ship crossing at port side).

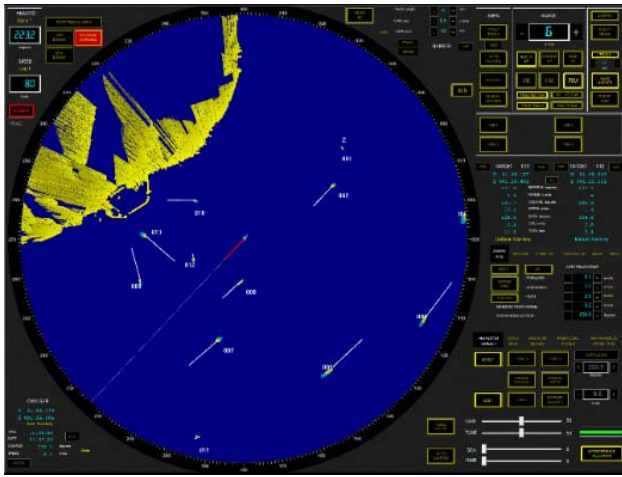


Figure 5. Own ship is in a crossing situation in open waters where it is required to give-way. The situation is of high complexity as there are restrictions to the target ship's avoidance manoeuvrability (a buoy).

### 3 DEVELOPING TRANSPARENCY FOR COLLISION AVOIDANCE

#### 3.1 Defining transparency layers

An earlier study led by the first author performed a cognitive task analysis to identify the information required to perform supervision of a collision avoidance system [12], [14]. The analysis describes the information pertaining to the supervisory task and depicts which information should be disclosed to human supervisors to make the internal reasoning of the collision avoidance system observable. However, the analysis only describes what information should be made available and it does not dictate which type, or how much of the identified information should be disclosed. Simply depicting all information simultaneously will likely put too large a cognitive burden on the supervisor's information processing capabilities, resulting in high mental workload. At the same time, only limiting the information from the system to single information elements may not provide the full picture about the system's internal reasoning either. In addition, considering the dynamic nature of the collision avoidance task, the information needed to effectively supervise the system may vary given the circumstances and the task analysis does not define which information should be disclosed when. As such, providing transparency to supervisors means making choices as to which information is made available to allow supervisors to understand the system's behaviour.

The rationale for specifying what constitutes transparency information in a collision avoidance context, together with how this information can be categorised into distinct information types is discussed in a separate study [23]. In brief, a simple information processing model was used (see Figure 6), consisting of information acquisition, information analysis, decision selection, and action implementation stages, to identify and categorise the information into discrete steps [24]. As such, a layered approach to transparency was used allowing supervisors to observe the different facets of the

system's input parameters, reasoning, decisions, and actions pertaining to the collision situation.

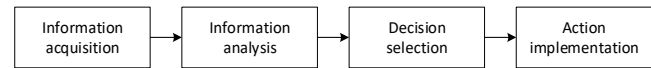


Figure 6. A simple model of human information processing adopted from [24].

This model provides, at minimum, a means to organize the information describing the system's information processing into several distinct parts. However, the model does not provide guidance as to which information takes priority over the other. A potential starting point is to try answer the question of what information supervisors would like to know at a minimum, before adding layers of transparency to allow for increased understandability. A plausible means for human supervisors to obtain an understanding of the collision avoidance system's performance is to be informed whether the system can avoid a potential collision at all. In other words, supervisors likely need to be informed about the system's decisions and actions first, before needing to "dig deeper" into the system's underlying reasoning. This indicates that the starting point for providing transparency to supervisors is thus the "decision selection" step of the information processing model depicted in Figure 6 and not the "information acquisition" step. (Note that in the "action implementation" step there is no information processing, only execution.) Further understanding of how and why the system has derived at its decision and planned actions can subsequently be obtained by "going backwards" through the model. That is, the "information analysis" stage of the model provides the relevant information pertaining to the analysis that underlie the system's decisions and actions. Finally, when the full picture is required for understanding the system's decisions and actions, the "information acquisition" stage of the model provides all the input data the system uses in its information processing.

#### 3.2 Development process

A concept illustration is provided of a radar screen depicting a traffic situation in which own ship, in the centre of the radar screen, is involved in a head-on situation (see Figure 7). Own ship depicts its intended avoidance manoeuvre by drawing its planned track for the next three manoeuvring steps (each step corresponds to one vector length and equals six minutes). It also states "GW" indicating it intends to give-way. Additional information about current and next actions, including speed, are depicted on the left side of the figure. With this information, minimum transparency is provided to allow supervisors to understand that the system is about to initiate a 12-degree starboard turn and that it intends to give-way. The information provided in Figure 7 was proposed as the minimum information needed to obtain an understanding of the own ship's decisions and actions.



Figure 7. Traffic situation with transparency information overlaid (decision selection).

Figure 8 depicts that own ship considers two targets as especially relevant in this traffic situation. The target ship in red is depicted as the highest risk as this ship is the one considered to be on collision course with own ship (minimum predicted CPA exceeded). The target in orange is also highlighted as own ship has considered this target to be of importance during the avoidance manoeuvre. Further information regarding the targets that own ship considers is provided through the indicators next to the targets depicting the conflict situation (e.g., HO for head-on, and MV for motor vessel). In addition, further information regarding the system's reasoning is provided through a manoeuvrability indicator around own ship indicating where it can manoeuvre within one vector length. Finally, tables to the left of the radar screen depict additional target information and the variables own ship has considered in determining safe speed.

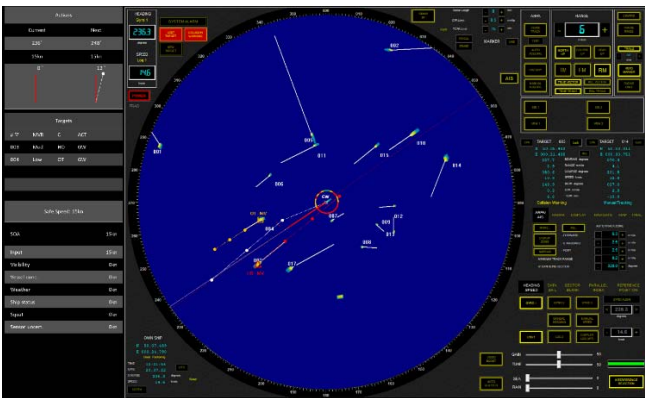


Figure 8. Traffic situation with transparency information overlaid (decision selection + information analysis).

Figure 9 provides a depiction of what a transparent collision avoidance system could look like when all transparency information described in the task analysis is provided. Here, all targets have received identifiers (green circles), and initial classifications (ship types and relevant conflict type indicators). In addition, information regarding the status of the system's sensors are provided in the tables to the left of the radar screen.



Figure 9. Traffic situation with transparency information overlaid (decision selection + information analysis + information acquisition).

### 3.3 Verification and validation workshop

The transparency concepts were developed through a series of iterations based on the information from the task analysis and the information processing model. Final verification and validation of the interfaces was performed in a second workshop with two independent navigators holding active navigational licenses with an average of 12 years of navigational experience (SD=9.9, min=5, max=19).

The purpose of this second workshop was to evaluate a selected set of traffic situations that included the transparency layers as described above. A representative subset of five traffic situations were included for review in this workshop, including head-on, crossing with own ship as stand-on, overtaken by a ship restricted in its manoeuvrability, crossing with speed-only as the avoidance manoeuvre and overtaking a slower ship when approaching a harbour. A talk aloud protocol was used where participants were asked to describe their interpretation of the traffic situation with primary focus on the information the system provided through the Human Machine Interface (HMI). In other words, the focus in the workshop was on how they perceived the collision avoidance system would solve the conflict situation, and not how they would solve it. The independent navigator's interpretations were noted, including all comments related to recommendations for improvement, corrections, and additions which were included in the final transparency iteration.

## 4 SUMMARY AND FURTHER WORK

When a collision situation occurs that requires human intervention, the collision avoidance system needs to facilitate human supervisors in gaining SA such that successful decisions can be made. This paper described the systematic development of a realistic and validated foundation for evaluating the relationship between automation transparency and human supervisory performance in an autonomous collision avoidance context. First, a set of traffic situations were developed based on navigational experience aimed at capturing the variability encountered in real-life situations. Second, a set of

transparency concepts were developed based on a cognitive task analysis and a model for human information processing. Together, these preparations provide the groundwork for the planned empirical work to explore this relationship.

As the maritime industry moves towards increased use of automation, including deploying systems that can perform (part of) the collision and grounding avoidance functions, there is an urgent need to understand how humans will interact with these systems. Automation transparency has been proposed as a critical element that can support human supervisors in obtaining situation awareness of the system's behaviours and actions [16]. Conversely, without transparency, i.e., systems that have low degrees of observability and predictability, humans will be highly challenged in understanding what the system is doing, why it is doing it, and what it will do next. As such, given the critical nature of the supervisory task for autonomous maritime collision and grounding avoidance systems, it is pertinent that further understanding is needed with regards to the application of the transparency in this domain.

This paper aimed to address this need by investigating how an information processing model could be used to drive the development of transparency layers. Given the dynamic nature of collision and grounding avoidance the amount and type of information needed to understand the system may depend on the type of situation, the degree of human oversight, the complexity of the situation, or the time available to intervene. The transparency concepts discussed in this paper have attempted to address this. In addition, an empirical evaluation is underway in which the relationship between automation transparency and human performance variables are evaluated in a collision avoidance context. This way, the relation between transparency and human performance variables can be explored, and its practical benefits can be assessed.

## ACKNOWLEDGEMENTS

The authors would like to express their gratitude to the navigators for their participation in the workshops. Also, we would like to express our sincere gratitude to Koen Houweling for his contribution in developing the traffic situations and the transparency illustrations.

## REFERENCES

[1] L. Kretschmann, H. C. Burmeister, and C. Jahn, "Analyzing the economic benefit of unmanned autonomous ships: An exploratory cost-comparison between an autonomous and a conventional bulk carrier," *Research in Transportation Business and Management*, vol. 25, pp. 76–86, 2017, doi: 10.1016/j.rtbm.2017.06.002.

[2] I. Kurt and M. Aymelek, "Operational and economic advantages of autonomous ships and their perceived impacts on port operations," *Marit Econ Logist*, vol. 24, no. 2, pp. 302–326, Jun. 2022, doi: 10.1057/s41278-022-00213-1.

[3] K. Wróbel, J. Montewka, and P. Kujala, "Towards the assessment of potential impact of unmanned vessels on

maritime transportation safety," *Reliab Eng Syst Saf*, vol. 165, pp. 155–169, 2017, doi: 10.1016/j.res.2017.03.029.

[4] Kongsberg, "Kongsberg maritime and Massterly to equip and operate two zero-emission autonomous vessels for ASKO," Sep. 01, 2020. <https://www.kongsberg.com/maritime/about-us/news-and-media/news-archive/2020/zero-emission-autonomous-vessels/> (accessed Nov. 18, 2020).

[5] Yara International, "Yara Birkeland," 2021. <https://www.yara.com/news-and-media/media-library/press-kits/yara-birkeland-press-kit/> (accessed Jan. 02, 2023).

[6] DNV, "DNVGL-CG-0264: Autonomous and remotely operated ships." 2018. [Online]. Available: <http://rules.dnvgl.com/docs/pdf/dnvgl/cg/2018-09/dnvgl-cg-0264.pdf>

[7] IMO, Resolution A.1047(27) Principles of safe manning. 2011.

[8] IMO, "Guidelines for the approval of alternatives and equivalents as provided for in various IMO instruments," MSC.1/Circ.1455, Jun. 2013.

[9] K. Aylward, R. Weber, M. Lundh, S. N. MacKinnon, and J. Dahlman, "Navigators' views of a collision avoidance decision support system for maritime navigation," *J. Navigation*, pp. 1–14, Sep. 2022, doi: 10.1017/S0373463322000510.

[10] E. Hannaford, P. Maes, and E. Van Hassel, "Autonomous ships and the collision avoidance regulations: a licensed deck officer survey," *WMU J Marit Affairs*, vol. 21, no. 2, pp. 233–266, Jun. 2022, doi: 10.1007/s13437-022-00269-z.

[11] M. A. Ramos, I. B. Utne, and A. Mosleh, "Collision avoidance on maritime autonomous surface ships: Operators' tasks and human failure events," *Saf Sci*, vol. 116, pp. 33–44, 2019, doi: 10.1016/j.ssci.2019.02.038.

[12] K. van de Merwe, S. C. Mallam, Ø. Engelhardtson, and S. Nazir, "Exploring navigator roles and tasks in transitioning towards supervisory control of autonomous collision avoidance systems," *J. Phys.: Conf. Ser.*, vol. 2311, no. 1, p. 012017, Jul. 2022, doi: <https://doi.org/10.1088/1742-6596/2311/1/012017>.

[13] S. N. Mackinnon, Y. Man, M. Lundh, and T. Porathe, "Command and control of unmanned vessels: Keeping shore based operators in-the-loop," 18th International Conference on Ships and Shipping Research, NAV 2015, pp. 612–619, 2015.

[14] K. van de Merwe, S. Mallam, Ø. Engelhardtson, and S. Nazir, "Supporting human supervision in autonomous collision avoidance through system transparency: a structured and systematic approach," under review.

[15] M. R. Endsley, "From Here to Autonomy: Lessons Learned from Human-Automation Research," *Hum Factors*, vol. 59, no. 1, pp. 5–27, 2017, doi: 10.1177/0018720816681350.

[16] M. R. Endsley, "Supporting Human-AI Teams: Transparency, explainability, and situation awareness," *Computers in Human Behavior*, vol. 140, p. 107574, Mar. 2023, doi: 10.1016/j.chb.2022.107574.

[17] J. Y. C. Chen, K. Procci, M. Boyce, J. Wright, A. Garcia, and M. J. Barnes, "Situation Awareness-Based Agent Transparency," U.S. Army Research Laboratory, Aberdeen Proving Ground, ARL-TR-6905, Apr. 2014. doi: 10.21236/ADA600351.

[18] J. Y. C. Chen, S. G. Lakhmani, K. Stowers, A. R. Selkowitz, J. L. Wright, and M. J. Barnes, "Situation awareness-based agent transparency and human-autonomy teaming effectiveness," *Theor Issues Ergon Sci*, vol. 19, no. 3, pp. 259–282, May 2018, doi: 10.1080/1463922X.2017.1315750.

[19] M. R. Endsley, B. Bolté, and D. G. Jones, *Designing for situation awareness: an approach to user-centered design*. London ; New York: Taylor & Francis, 2003.

[20] A. Bhaskara, M. Skinner, and S. Loft, "Agent Transparency: A Review of Current Theory and

- Evidence," *IEEE Trans Hum Mach Syst*, vol. 50, no. 3, pp. 215–224, Jun. 2020, doi: 10.1109/THMS.2020.2965529.
- [21] F. Rajabiyazdi and G. A. Jamieson, "A Review of Transparency (seeing-into) Models," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct. 2020, pp. 302–308. doi: 10.1109/SMC42975.2020.9282970.
- [22] K. van de Merwe, S. Mallam, and S. Nazir, "Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review," *Hum Factors*, p. 00187208221077804, Mar. 2022, doi: 10.1177/00187208221077804.
- [23] K. van de Merwe, S. Mallam, Ø. Engelhardtson, and S. Nazir, "Supporting human supervisory performance through information disclosure: establishing transparency requirements for maritime collision avoidance," in *Proceedings of the Human Factors Society Annual Meeting, Orlando, FL*, submitted.
- [24] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Trans Syst Man Cybern*, vol. 30, no. 3, pp. 286–297, May 2000, doi: 10.1109/3468.844354.



**Article 5**

van de Merwe, K., Mallam, S., Nazir, S., & Engelhardtson, Ø. (2024). The Influence of Agent Transparency and Complexity on Situation Awareness, Mental Workload, and Task Performance. *Journal of Cognitive Engineering and Decision Making*, 18(2), 156–184.  
<https://doi.org/10.1177/15553434241240553>





# The Influence of Agent Transparency and Complexity on Situation Awareness, Mental Workload, and Task Performance

Journal of Cognitive Engineering  
and Decision Making  
2024, Vol. 18(2) 156–184  
© 2024, Human Factors  
and Ergonomics Society



Article reuse guidelines:  
[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)  
DOI: 10.1177/15553434241240553  
[journals.sagepub.com/home/edm](https://journals.sagepub.com/home/edm)



Koen van de Merwe<sup>1,2</sup> , Steven Mallam<sup>2,3</sup>, Salman Nazir<sup>2</sup>, and Øystein Engelhardtson<sup>1</sup>

## Abstract

Transparency is a design principle intended to make the inner workings of autonomous agents visible to end-users such that humans can evaluate the reasoning behind its decisions and actions. To test the effect of agent transparency on situation awareness, mental workload, and task performance, an experiment was performed where 34 nautical navigators were tasked with interpreting the information provided by an autonomous collision and grounding avoidance system. Sixteen traffic situations were created with two levels of complexity. Four levels of transparency varied the amount and type of information in terms of the system's decisions, planned actions, reasoning, and input parameters. The results show that increased transparency improves SA without increasing mental workload. However, the time to comprehend the system's decisions and planned actions increased when its reasoning was depicted. Traffic complexity impaired SA, mental workload, and time-to-comprehension regardless of transparency level. However, for level 2 SA, transparency was found to negate the influence of complexity, resulting in improved comprehension of the agent's reasoning despite high traffic complexity. These outcomes demonstrate the merits of agent transparency as a design principle in supporting human supervision of autonomous agents. However, developers should take care when extending these principles to time-critical applications.

## Keywords

human-automation interaction, autonomous agents, supervisory control, human-machine interface, experimental design

## Introduction

### *Autonomous Shipping and Human Supervisory Control*

Systems with autonomous capabilities, typically based on Artificial Intelligence (AI) and Machine Learning algorithms, are proliferating across society and industries. In the maritime domain, ships are envisioned to deploy advanced automation, or 'agents', capable of sensing their environment and executing goal-directed behaviour using actuators,

allowing for advanced functions to be performed with increasing levels of autonomy (IMO, 2018; Russell & Norvig, 2022). For example, in Japan, a

---

<sup>1</sup>DNV, Norway

<sup>2</sup>University of South-Eastern Norway, Norway,

<sup>3</sup>Memorial University of Newfoundland, NL, Canada

#### Corresponding Author:

Koen van de Merwe, DNV, Veritasveien 1, Høvik 1363, Norway.

Email: [koen.van.de.merwe@dnv.com](mailto:koen.van.de.merwe@dnv.com)

commercial container ship conducted a 790-km trial to test its autonomous navigation capabilities without human intervention (Nippon Yusen Kaisha; NYK, 2022). In Norway, the Yara Birkeland container ship and the ASKO barges have commenced operation with the aim to navigate autonomously within a few years (AS Kolonialgrossistene; ASKO, 2022; Yara International, 2022). Here, operators are envisioned to work in positions from which single or multiple autonomous ships can be continuously monitored and supervised (e.g. see Massterly, 2023). In this context, supervisory performance is dependent on the operator's ability to '[perceive] elements in the environment within a volume of time and space, [comprehend] their meaning, and [project] their status in the near future', that is, to obtain and maintain situation awareness (SA; Endsley, 1995, p. 36). In other words, this means that operators should be able to perceive critical parameters made available through the control and safety systems, analyse the ship's current and planned behaviour, and evaluate the plan's adequacy considering its context (van de Merwe, et al, 2024a). To support operators in achieving and maintaining SA of an autonomous ship's performance, it is critical to understand how effective human supervisory performance can be achieved whilst avoiding potential human performance pitfalls.

Challenges related to the human supervision of highly automated systems are well documented in the scientific literature (Endsley, 2017). For example, the out-of-the-loop (OOTL) performance problem is attributed to a loss of skills and SA, and occurs when operators are no longer an active part of a system's information loop (Endsley & Kiris, 1995; Metzger & Parasuraman, 1999). In addition, transitioning back into the information loop often results in high workload because of the need to build SA and regain manual control (Endsley, 2017; Onnasch, Wickens, Li, & Manzey, 2014; Weaver & DeLucia, 2020). Taken together, these challenges are described as the 'automation conundrum' which states that 'the more automation is added to a system, and the more reliable and robust that automation is, the less likely the human operators overseeing the automation will be aware of critical information and able to take over manual control when needed' (Endsley, 2017, p. 8). This means the safe implementation of systems with autonomous capabilities thus depends on the

degree to which humans can oversee the agent's decisions and actions, and the agent's ability to afford humans insight into its reasoning processes (J. Y. C. Chen, Procci, et al., 2014a).

### *Human Performance and Agent Transparency*

'Agent transparency' (J. Y. C. Chen, Procci, et al., 2014a), 'system transparency' (Ososky, et al, 2014), 'display transparency' (National Academies of Sciences, Engineering and Medicine, 2022), 'automation transparency' (Skraaning & Jamieson, 2021), or simply 'transparency' are terms used to describe the 'understandability and predictability of [a] system' (Endsley, 2023; Endsley, Bolté, & Jones, 2003, p. 146). Endsley (2017) defined transparency as a means to enhance the understandability and predictability of systems by making observable what it is doing, why it is doing it, and what it will do next. J.Y.C. Chen et al. described agent transparency as 'the descriptive quality of an interface pertaining to its abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process' (2014b, p. 2). Finally, Lyons depicted transparency as the ability of an operator to perceive an agent's abilities, intents, and situational constraints (2013). The aim of agent transparency is to provide 'a real-time understanding of the actions of the AI system' (National Academies of Sciences, Engineering and Medicine, 2022, p. 31) and enable 'the operator to maintain proper SA of the system in its tasking environment without becoming overloaded' (Mercado et al., 2016, p. 402). In addition, transparency intends to facilitate human-agent collaboration when humans are tasked with supervising automated systems. That is, when an agent communicates what it does, why it does it, and what it will do next, human supervision should be supported (Endsley, 2023). Conversely, opaque agents can be challenging to supervise as they may be difficult to interpret because of a lack of information provision (Doshi-Velez & Kim, 2017; Lipton, 2017). In other words, when the agent's inner workings are made apparent to the user, the user's comprehension of the agent may be enhanced (Ososky et al., 2014).

In recent years, there has been an increasing interest in understanding the effect of transparency on selected human performance variables including, SA (Selkowitz, Lakhmani, & Chen, 2017; Skraaning & Jamieson, 2021; Wright, Chen, & Lakhmani, 2020), decision making (Bhaskara et al., 2021; Loft et al., 2023), mental workload (Mercado et al., 2016; Stowers et al., 2020), and automation trust (J. Y. C. Chen et al., 2018; Ezenyilimba et al., 2023; Schmidt, Biessmann, & Teubner, 2020). Furthermore, a recent review of the transparency literature studied the relation between agent transparency and human performance variables finding positive effects on SA and task performance, without negatively affecting mental workload, for increasing levels of transparency (van de Merwe et al., 2024b). These findings indicate the potential benefit transparency can have in cases where operators need to understand the behaviour of a system and perform manual intervention when required. Thus, transparency can be especially relevant in safety critical domains where understandability and predictability are essential for safe and effective control of processes (Endsley, 2023; Jamieson, Skraaning, & Joe, 2022).

### *Agent Transparency and Autonomous Shipping*

Several recent studies have addressed agent transparency within the autonomous shipping domain. For example, Ramos et al. (2019) performed a task analysis to derive potential human failures when monitoring autonomous ships. Here, the study identified the importance of the supervisors' ability to collect and evaluate information from the autonomous ship through 'an adequate HMI' (human-machine interface), such that a strategy for intervention could be determined should the automation fail (Ramos et al., 2019, p. 43). Van de Merwe et al. (2024a) identified specific information requirements for supervising autonomous collision and grounding avoidance (CAGA) systems based on a Goal-Directed Task Analysis (GDTA; Endsley et al., 2003). The study highlighted the need for continuous, sufficient, and adequate information about the CAGA system's decisions, planned actions, and underlying

information processing, that is, transparency information, to alleviate some of the human performance issues in supervision and support safe and effective oversight of CAGA systems. Furthermore, Porathe (2021) discussed the use of 'expert systems' to aid operators in supervising one or more autonomous ships. Here, HMI concepts were proposed aiding operators to obtain at-a-glance understanding of how the system perceives and understands the nearby traffic and its intentions for solving collision situations. This includes showing how the CAGA system plans to solve a situation by graphically displaying the various options it has considered, and which solution it intends to execute. Also, Van de Merwe et al. (2023a) operationalised transparency for autonomous ships by developing concepts for how an autonomous CAGA system may display its perception and analysis of its environment, determination of collision risk, and plans to resolve the situation. Moreover, Alsos et al. (2022) examined how the transparency concept could be operationalised for autonomous ships. Here, the aim was to assess how autonomous ships can share intent information to external stakeholders, such as passengers, traffic services, and other nearby ships. Finally, operationalising this idea, Simic and Alsos (2023) developed a concept for autonomous urban ferries in which the ship's perceptions, current state, and future intentions are communicated to external stakeholders through light strips and displays mounted on the outside of the ferry.

Although these studies address the potential benefits of agent transparency in relation to human supervisory control in an autonomous shipping context, they fall short on measuring its purported effects. That is, to the best of our knowledge, no studies have been performed that have empirically tested the effect of transparency on human supervisory performance in an autonomous shipping context and have considered the complexities that can arise in realistic traffic-dense environments. As such, given the concrete developments towards autonomy in the maritime domain, there is a need for knowledge with regards to the application of transparency within this context and study its effect on human performance variables. Therefore, this study aims to extend the literature by empirically evaluating the application of transparency in a maritime autonomous shipping context.

Specifically, this study asks what the effects of agent transparency and traffic complexity are on the supervisor's (1) SA, (2) mental workload, and (3) task performance.

### *Situation Awareness, Mental Workload, and Task Performance*

In complex and dynamic environments, such as shipping, action execution is highly dependent on the human's ability to make accurate and timely decisions in a constantly changing state of the environment. When the collision avoidance task is performed by an autonomous agent, the supervisor's mental model is of particular relevance in understanding if its behaviour is according to expectations or whether intervention is needed (Endsley, 2017). In addition, to effectively assess the real-time and future performance of an autonomous collision and grounding avoidance system, supervisors need to have SA of the system in its tasking environment (Endsley, 2023). To support this, CAGA systems should provide detailed and relevant information regarding its internal processing, for example, which elements in its environment it has perceived (ships, objects, and shallow waters), how these affect the ship's collision and grounding risk (collision, no collision), and how it plans to resolve the situation (give-way and stand-on). This way, SA knowledge, that is, the system's perceptual, comprehended, and projected knowledge (van Doorn, Rusák, & Horváth, 2017), is directly provided to the supervisor and understandability and predictability of the system should be improved (Bhaskara et al., 2020; Endsley, 2023; van de Merwe et al., 2024b). In other words, when information is provided in a manner that supports the cognitive processes needed for supervision, for example, by providing information compatible with how humans process information and make decisions (Westin, Borst, & Hilburn, 2015), improved SA should be expected. Specifically, it is expected that level 1 SA is improved when the CAGA system depicts its perception of the environment, level 2 SA is improved when the system depicts its analysis, and level 3 SA is improved when it provides its decision and planned actions (see Table 1). Furthermore, it is hypothesised that

transparency may especially be beneficial in complex traffic situations where making sense of the system's reasoning may be challenged by the high amount of information available for interpretation. Presenting information that supports transparency, provided this is presented in a salient, well-organised, and integrated manner, is expected to support SA in such cases (Endsley, 2023; Endsley et al., 2003; Skraaning & Jamieson, 2021).

As agent transparency is about disclosing system-internal information, the degree of transparency can typically be varied by increasing or decreasing the amount of information it presents about its internal processes, decisions, and planned actions (see Bhaskara et al. (2021) and Pokam et al. (2019) for examples). Although increased levels of agent transparency imply increased insight into the agent's reasoning, full disclosure of the system's internal state may pose challenges in terms of the user's cognitive processing capabilities (Bhaskara et al., 2020; Wickens, 2018). That is, although increased transparency may benefit SA, this may also add an additional cognitive processing burden due to the resources required for selecting and dividing attention and keeping information in working memory (Wickens & Carswell, 2021). This may be exacerbated in situations where the baseline level of information is already high, that is, in complex traffic situations (Moacdieh & Sarter, 2017). Here, increased levels of transparency information may provide an additional information burden and the risk of overloading the operator with information that supports transparency is high, especially when increased information leads to display clutter (Moacdieh & Sarter, 2015a). However, despite risks of increased workload with agent transparency, recent studies have not found a clear relationship between these variables (Ezenyilimba et al., 2023; Loïck, Guérin, Rauffet, Chauvin, & Éric, 2023; Tatasciore, Bowden, & Loft, 2023), possibly because of the use of graphical symbols and integration of transparency information in task displays (Gegoff, Tatasciore, Bowden, McCarley, & Loft, 2023; van de Merwe, et al, 2024a; van Doorn, Horváth, & Rusák, 2021). Building on these findings, this study anticipates that when, first, information requirements are identified based on an iterative human-centred design approach (Endsley et al., 2003; ISO, 2019), second, symbology is developed based



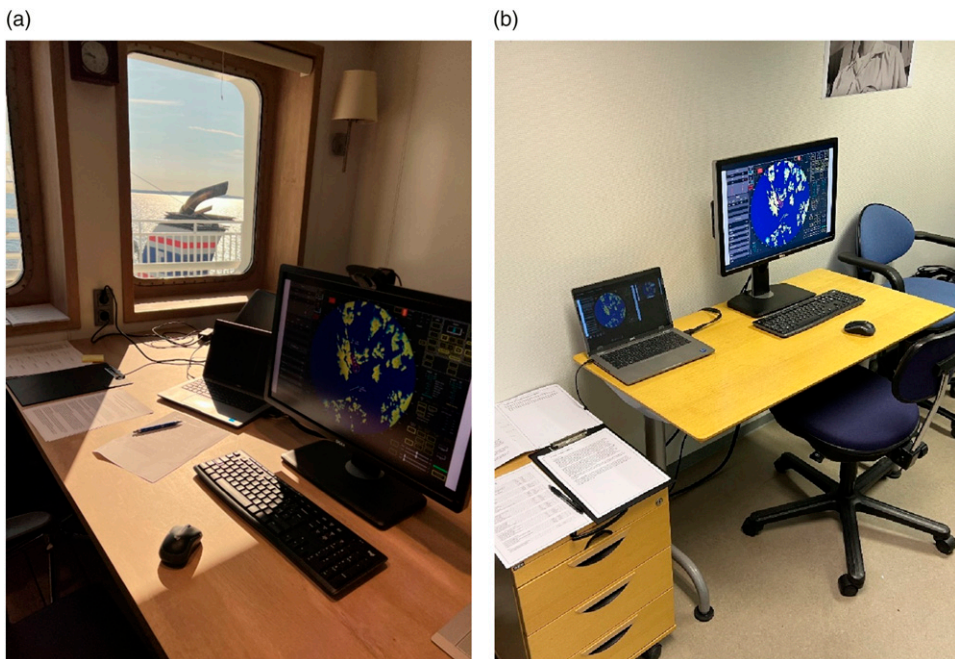
30 participants held an active license whilst 4 participants have navigation experience, but at the time of the study their licenses expired between 1 and 5 years prior.

### Technical Setup

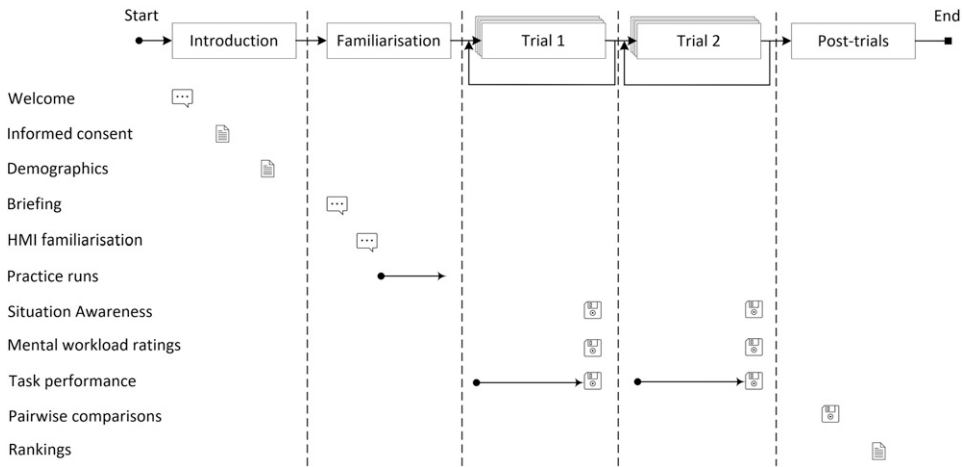
To maximise recruitment, the first author travelled to locations most suitable for the participants to perform the study, including onboard a passenger ferry where participants worked or at various national nautical training institutes. Nevertheless, the technical setup, conditions, and conduct of the experiment was standardised and consistent regardless of the location where the data was gathered (see Figure 1). The experiment was conducted on a standard portable office computer using a 24" screen with 1920x1200 resolution running Windows 10. E-Prime 3.0 served as the experimental platform in which the experimental stimuli were provided and primary data was recorded (Psychology Software Tools, Inc, 2023). Finally, post-experiment interviews were recorded using pen and paper.

### Execution of the Experiment

**Procedure.** Figure 2 depicts the execution of the experiment. After a brief introduction, participants signed an informed consent form stating that participation is voluntary and that they had the liberty to withdraw at any stage during the experiment, without reason or penalty. This research complied with the American Psychological Association Code of Ethics and was approved by the Norwegian Centre for Research Data reference number 986652. Informed consent was obtained from each participant. Participants were briefed on the experimental procedure, what was expected of them, and the HMI used in the experiment. A practice session was performed to familiarise the participants of the execution of the experiment, including stimuli and questionnaires. After this, the experiment commenced, and the experimental trials and measurements were performed. Two trials were performed that were identical in set up, but with new traffic situations to avoid familiarisation. After the trials, the pairwise comparisons, as part of the workload measurements, were performed, and a semi-structured interview was



**Figure 1.** The technical setup used for the experiment: on location onboard of one of the passenger ferries, and at the university's experimental lab.



**Figure 2.** An illustration of the procedure for the experiment.

conducted. Depending on the participant's progress, the entire experiment lasted between one and 2 hours and the experimental trials between 10 and 30 minutes each.

**Experimental Tasks.** Participants took the role of a supervisor of a ship equipped with an autonomous CAGA system. They were tasked with observing and understanding a traffic situation depicting own ship in conflict with a target ship and own ship's proposed solution to resolve it. Once the participant felt they had sufficiently understood the situation, including the system's solution, they were to press a button on the keyboard after which the screen was blanked, and questions were presented concerning SA and mental workload. To provide participants with a sense of urgency, participants were told they had a 90 second time limit to evaluate the traffic situation after which the radar image would disappear automatically. However, in practice, there was no time limit imposed by the researchers to avoid a ceiling effect in the measurements. No time keeping device was available to the participants. Once the questions were answered, the participant pressed a key to continue, and a new traffic situation was shown. This process was repeated until all traffic situations for all experimental conditions were completed.

The traffic situations were developed by a licensed navigator and reviewed by two independent, and licensed navigators (see [Van de Merwe et al. \(2023a\)](#) for further details). The traffic

situations were created on a desktop simulator at a maritime education and training institution. Each traffic situation was configured such that they represented a potential collision situation involving own ship and one other vessel in either a head-on-, crossing-, or overtaking situation. To avoid familiarisation with the traffic situations, multiple variations were developed including conflict situations in coastal- and confined waters, restrictions in target ship's ability to manoeuvre, and own ship as a stand-on vessel ([IMO, 1977](#)). However, to ensure equivalence in difficulty between the situations, they only consisted of one-to-one ship encounters. This meant that, although traffic situations could depict multiple ships, own ship was only in conflict with one other target ship. As such, traffic situations were created with variations in terms of type of conflict situation (head-on, crossing, overtaking/overtaken), who has right of way (own ship is the give-way vessel or the stand-on vessel), type of relevant avoidance actions proposed by the CAGA system (route-and/or speed change), and any restrictions in target ship manoeuvrability (restricted in ability to manoeuvre). In total, 20 unique traffic situations were used for the experiment: four for the familiarisation phase, eight for trial one and eight for trial two, that is, 16 situations for the experimental trials in total. For readers interested in the traffic situations and their configurations, a table is made available as [Supplemental Material](#) on the journal's web site.

**Experimental Design.** This study used a repeated measures approach in which all participants performed all eight experimental conditions: four transparency levels x two complexity levels. Participants were shown one traffic situation for each condition in each trial. Since the experiment comprised of two trials, participants performed 16 experimental runs in total. The data for each experimental condition was averaged between trial one and two. To avoid familiarisation and order effects, the conditions were administered in random order within each trial.

### **Independent Variables**

**Transparency.** For this study, four levels of transparency were defined based on the amount and type of information to disclose to the supervisor. Which information to disclose was identified in an earlier study based on a GDTA of collision avoidance manoeuvring (van de Merwe, et al., 2024b). These information requirements were subsequently structured based on an information processing model (Parasuraman, Sheridan, & Wickens, 2000; van de Merwe, et al, 2023b) (see Figure 3).

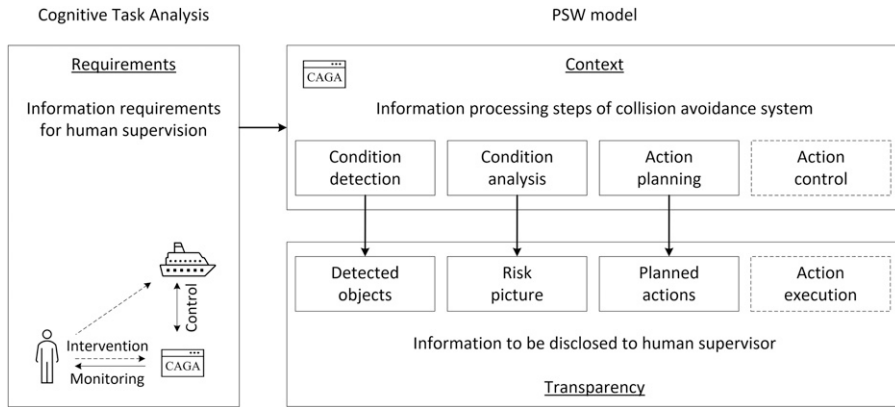
Applying this model to the collision avoidance context allowed the information from the task analysis to be organised into distinct categories with specific information elements belonging to each information processing step (see Table 3). This, in turn, defined which information was depicted (van de Merwe et al., 2023b). In the ‘condition detection’ step, information regarding the system’s input parameters were depicted, including which objects it had detected, a basic classification of these in terms of object type and size, the object’s relative motion to own ship, its sensor status, and any uncertainties in these. In the ‘condition analysis’ step, information regarding the outcome of the system’s analytical process were depicted in terms of objects that posed a collision risk, priorities, intended trajectories and safe speed parameters. In the ‘action planning’ step, the CAGA system depicted its collision avoidance decision and updated passage plan only. Finally, note that the ‘action control’ step was omitted as there was no information processing performed in this stage, only the execution of earlier made decisions and action plans.

Based on this structure, four levels of transparency were defined in which the amount and type of information was varied. These levels were defined based on the minimum requirements for supervising an autonomous CAGA system. A ‘low’ level of transparency was defined as the information an operator needs at minimum for supervision. In this case, the low transparency level depicts the information about the system’s decisions and planned actions (the ‘action planning’ stage in Figure 3). Additional transparency is provided by adding information from previous information processing steps; in other words, by ‘going backwards’ through the model. As such, the ‘medium (A)’ level of transparency is a combination of the ‘action planning’ and ‘condition analysis’ steps and the ‘medium (B)’ layer is a combination of the ‘action planning’ and ‘condition detection’ steps. The latter level was developed to explore the effects of providing participants with information regarding the system’s detection and action planning only, that is, leaving out the analysis part, and thereby not only varying the amount of information but also the type. Finally, the ‘high’ level provides the information from all steps: the ‘action planning’, ‘condition analysis’, and ‘condition detection’ steps (see Table 4).

**Traffic Complexity.** Two levels of complexity were defined for this study: traffic situations with low and with high complexity. Traffic complexity was defined by the degree to which own ship had the space to perform an avoidance manoeuvre. In cases where there was limited manoeuvring space, for example, because of another ship, the vessel was considered ‘boxed in’ and own ship may needed to postpone an avoidance manoeuvre until the obstruction had been passed, change speed, or choose an alternative solution. Given the additional analysis and decision making that was required for such cases, these were considered more complex than those where a single and unobstructed solution could be implemented. As such, complexity was operationalised by adding objects to the traffic situation and ensuring own ship is boxed in.

**Human-Machine Interface.** During the experimental trials, participants were shown traffic situations





**Figure 3.** The framework for establishing transparency requirement for a CAGA system based on a model of human information processing (adapted from Parasuraman et al., 2000; and adopted from van de Merwe et al., 2023b).

**Table 3.** Information Elements Corresponding to Each Information Processing Step (van de Merwe et al., 2023b). Key: OT = overtaking/overtaken, HO = head-On, CR = crossing, GW = give-way, SO = stand-on.

Information Processing Step	Information Elements CAGA Should Depict (Excerpts)
1. Condition detection CAGA performs object detection, basic classification, object tracking, and status	<ul style="list-style-type: none"> <li>- Detected objects short and long range</li> <li>- Identified target ship</li> <li>- Target object type and size</li> <li>- Identified target object as OT/HO/CR</li> <li>- Uncertainties in the radar/sensor data</li> <li>- Status of sensors</li> </ul>
2. Condition analysis CAGA performs object classification, tracking, situation analysis, and risk estimation	<ul style="list-style-type: none"> <li>- Objects that pose a risk</li> <li>- Plotted objects</li> <li>- Risk object type and size</li> <li>- Risk object priority</li> <li>- Risk object course and speed</li> <li>- Risk object intended trajectory</li> <li>- Risk object conflict type</li> <li>- Safe speed parameters</li> </ul>
3. Action planning CAGA decides on collision avoidance manoeuvring and determines an updated passage plan	<ul style="list-style-type: none"> <li>- Own ship priority (GW/SO)</li> <li>- Target vessel priority (GW/SO)</li> <li>- Own ship intended track and speed</li> </ul>

**Table 4.** Levels of Transparency.

Level of Transparency	Information Processing Steps		
	Condition Detection	Condition Analysis	Action Planning
Low			X
Medium (A)		X	X
Medium (B)	X		X
High	X	X	X

in the form of a static radar image depicted on a radar display from a popular maritime equipment manufacturer (see Figure 4 for an example). On this image, vessels, objects and other radar echoes were shown representing a realistic traffic situation. Information such as settings, range, targets, and (time to) closest point of approach limits were also available and could be freely used by the participant to make sense of the traffic situation.

Information about the CAGA system’s information processing was added to the radar display (see Figure 5 for an example) and integrated in the primary task display as much as possible (Endsley, 2023; Endsley et al., 2003). The symbology representing information that supports transparency was developed by a licensed navigator using an iterative development process (ISO, 2019), based on the IEC 62288 standard for maritime navigation and radiocommunication equipment (IEC, 2022), and reviewed by two independent and licensed navigators (see Van de Merwe et al., 2023a for more details on the development process). In this case, central information regarding own

ship actions, risk analysis, and detections, were overlaid onto the primary information source for collision avoidance, that is, the radar display. Depending on the experimental condition, this information varied depending on the relevant level of transparency (see Table 4) and thereby which elements of the system’s information processing were depicted (see Table 3). An example of a traffic situation with four different levels of transparency is made available as Supplemental Material on the journal’s web site.

### Dependent Variables

**Situation Awareness.** SA about the system’s solution and information provision was measured after each experimental run using the Situation Awareness Global Assessment Technique (SAGAT; Endsley & Garland, 2000). The SAGAT is an assessment of a person’s SA that is typically applied in experiments where simulations are used. At random intervals during the simulation, the participant’s screen is blanked and specific queries

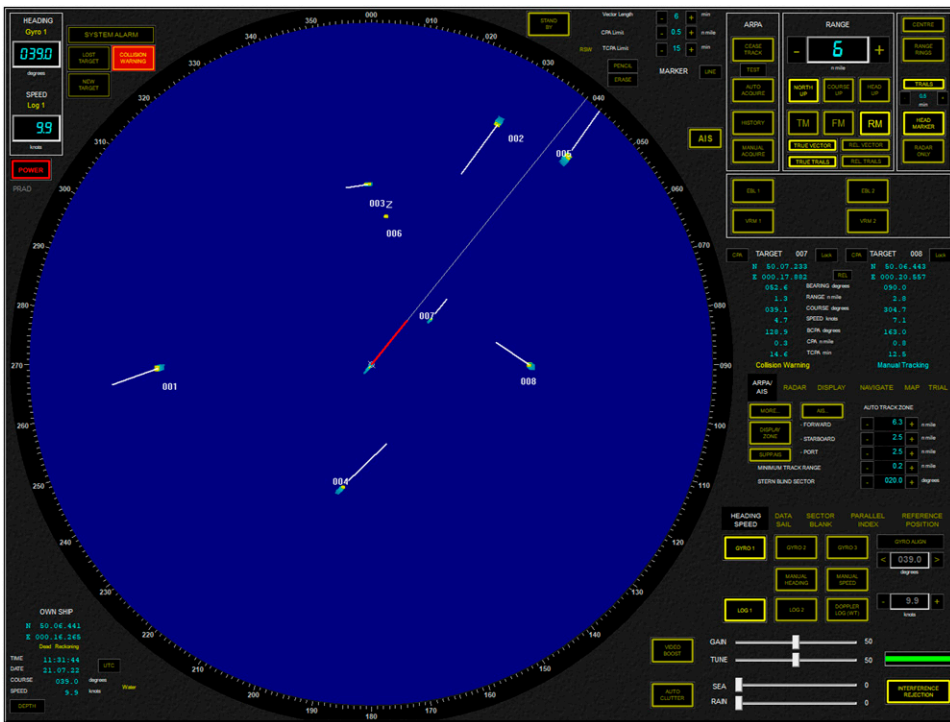
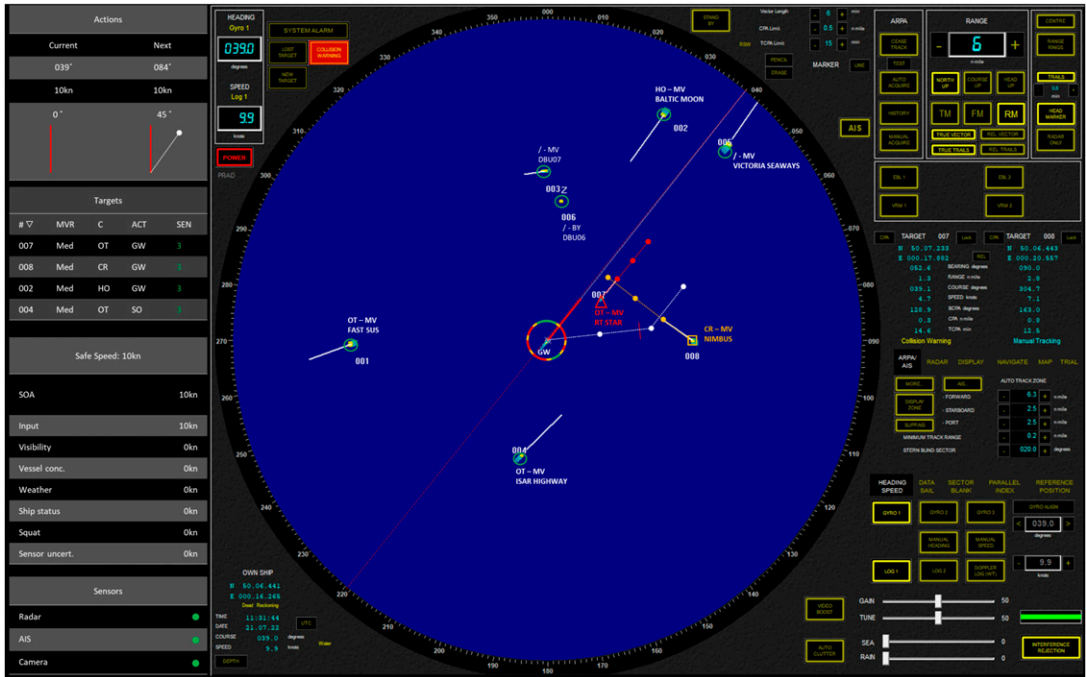


Figure 4. A typical traffic situation representing a collision situation (overtaking) with high complexity (without transparency).



**Figure 5.** A typical traffic situation representing a collision situation (overtaking) with high complexity (with transparency).

about the scenario are asked. Because the participant's answers can be evaluated against the data from the simulator, the SAGAT provides an objective assessment of SA (Endsley & Garland, 2000). For this experiment, a pool of 30 generic SAGAT queries were developed and distributed across the traffic situations. These queries were subsequently tailored to specifically fit the situation (see Table 5 for an example). Three SA queries were administered per traffic situation, one for each level of SA, one at a time, and starting at level 1 SA. Participants answered by selecting one of three multiple-choice alternatives per query of which only one alternative was correct. Scores were recorded per level of SA where '1' was correctly answered and '0' was incorrectly answered.

**Workload.** Workload was measured using the NASA-TLX (Hart & Staveland, 1988). This scale measures self-reported subjective experience of workload across six dimensions (mental demand, physical demand, temporal demand, performance, effort, and frustration level). As part of this scale, participants perform a pairwise comparison to

create weights for the dimensions. The sum of the weighted workload scores for all dimensions defines the total workload score. However, as setting the weights after each run is somewhat time-consuming and as the type of task is constant across the experiment, a version of the NASA-TLX was used where participants only perform pairwise comparisons once, and only after all experimental trials were performed. As such, the weights derived from the pairwise comparison applied to all workload scores for the individual runs.

**Task Performance.** Task performance was defined as the time required for participants to feel they had obtained an understanding of the traffic situation through the information provided by the CAGA system, that is, time-to-comprehension (TTC). Similar to other time-related performance measures, such as eye-tracking, reading speed, search time, and time to task completion, this variable was chosen as an indicator of how quickly humans are able to process information (Gawron, 2019). TTC was self-guided and consisted of the participant deciding that the traffic situation and the visualised

**Table 5.** Examples Situation Awareness Queries for the Traffic Situation Depicted in the Above figures. Correct Answers are in Bold Font.

Level of SA	Query
1	How many targets, within 40 degrees of either side of your bow, are sailing approximately in the opposite direction of you? a) None <b>b) 1</b> c) 3
2	What target ship limits your ability to perform an avoidance manoeuvre? a) 'ISAR HIGHWAY' (target 004) b) 'RT STAR' (target 007) <b>c) 'NIMBUS' (target 008)</b>
3	For 'NIMBUS' (target 008), what is your ship's intention? a) Pass on its starboard side b) Pass on its port side <b>c) Pass on its aft</b>

solution was sufficiently understood. The time measurement started at the moment the traffic situation was displayed and ended upon a key press by the participant after which the screen was blanked. Time was measured in seconds with no time limit imposed. Still, the participants were urged to be as quick and accurate as possible.

**Ranking.** After the experimental trials, one representative high complexity traffic situation from the experiment was shown but with different levels of transparency presented. Participants were asked to rank the four variants for each of the dimensions of transparency: observability and predictability (MITRE, 2018). Definitions for these dimensions were read verbatim to the participants and were available on paper, including an example of its application in the collision avoidance context. A think-aloud protocol was used to record the participant's verbal reasoning of the ranking (Eccles & Aarsal, 2017). The traffic situation with four levels of transparency that was used for the ranking is made available as [Supplemental Material](#) on the journal's web site.

## Results

### Data Analysis and Statistics

In the experiment, two trials were performed (trial 1 and trial 2) that were identical in experimental setup and execution, but for which different traffic

situations were used. The data from these trials were averaged, and screened for missing values, outliers, and tested for normality. Due to technical issues with the experimental setup, recording of TTC was incomplete for the initial set of participants, and led to missing values for six participants. This issue was corrected, and no missing values were reported for the remaining participants. As a result, of the 272 measurements for TTC, 20 measurements (7%) were missing. Finally, there were three participants with outliers for the TTC variable that were removed in the final data analysis. An outlier was defined as a data point lying outside 1.5 times the inter-quartile range of that variable. Thus, the data of 25 participants were used in the analysis of this variable.

The dependent variables were tested for normality using the Shapiro-Wilk test (Shapiro & Wilk, 1965). Significant deviations from normality were found for the SA scores. However, the number of observations per cell for these variables was sufficient and equal for each cell ( $N = 34$ ), such that robustness against normality was attained. As such, it was decided to use the standard Repeated Measures – Analysis of Variance (RM-ANOVA). Main and interaction effects were tested using the F-test, and follow-up pairwise comparisons between the levels of the independent variables were performed using t-tests with Bonferroni corrections.

Table 6 presents the overall descriptive statistics and correlations for the dependent variables. For

calculating the correlations, the overall mean scores per participant were calculated for each dependent variable. This resulted in five new variables representing the mean values for each dependent variable, irrespective of transparency- or complexity level. Subsequently, the Pearson correlation was calculated between these dependent variables. Significant positive correlations were found between TTC and level 1 SA ( $r(23) = .43, p < .05$ ), and TTC and level 3 SA ( $r(23) = .43, p < .05$ ). In other words, increased TTC were positively associated with the ability to perceive elements in the traffic situation, and the ability to project the status of these elements into the future. In addition, a positive correlation between level 2 SA and level 3 SA was found ( $r(32) = .51, p < .05$ ), indicating a positive association between the comprehension of the current traffic situation and its projection into the future. No significant correlations were found between TTC and level 2 SA, TTC and workload, and level 1 SA with level 2 SA, level 3 SA, and mental workload.

Table 7, Table 8, and Table 9 depict the means and standard deviations for the dependent variables as a function of transparency, complexity,

and their interactions. The statistical results for each of these variables, including the figures depicting the interaction between transparency and complexity, are presented in their respective subsections below. For readers interested in the graphs depicting the main effects for transparency and complexity, figures are made available as Supplemental Material on the journal’s web site.

**Situation Awareness**

A main effect for transparency was found for level 1 SA ( $F(3, 31) = 9.37, p < .001, \eta_p^2 = .48$ ). The high transparency level ( $M_{high} = .71$ ) resulted in improved awareness of elements in the environment compared to the low transparency and the medium (B) condition ( $M_{low} = .46, M_{medium (B)} = .57$ ; see Table 7). No differences were found between the medium (A) condition and the other conditions ( $M_{medium (A)} = .60$ ). A main effect for complexity was found where traffic situations with high complexity indicate lower level 1 SA ( $F(1, 33) = 30.35, p < .001, \eta_p^2 = .48; M_{low} = .70, M_{high} = .47$ ; see Table 8). A weak and non-significant interaction was found between the transparency and

**Table 6.** Overall Means, Standard Deviations, and Pearson Correlations Between the Dependent Variables.

	N	Mean	SD	Level 1 SA	Level 2 SA	Level 3 SA	WL	TTC
Level 1 SA	34	.59	.12	--				
Level 2 SA	34	.74	.13	.30	--			
Level 3 SA	34	.76	.14	.33	.51**	--		
WL	34	.59	.21	-.20	-.05	-.18	--	
TTC	25	46.72	13.47	.43*	.38	.43*	.16	--

\* $p < .05$  and \*\* $p < .01$ . Note that TTC is measured in seconds.

**Table 7.** Means and Standard Deviations for the Dependent Variables as a Function of Transparency Level only. Note That TTC is Measured in Seconds.

	Transparency Level							
	Low		Medium (A)		Medium (B)		High	
	M	SD	M	SD	M	SD	M	SD
Level 1 SA	.46	.22	.60	.24	.57	.19	.71	.23
Level 2 SA	.65	.25	.84	.20	.85	.18	.64	.24
Level 3 SA	.73	.19	.69	.23	.77	.20	.85	.21
WL	60.19	20.58	58.37	21.98	60.12	21.67	60.80	24.39
TTC	38.40	14.17	52.62	14.24	44.78	14.75	51.07	14.62

complexity conditions ( $F(3, 31) = 2.89, p = .051, \eta_p^2 = .22$ ; see Table 9). As Figure 6 depicts, there are differences between the level 1 SA scores between the low- and high complexity conditions across the transparency levels, except for medium (A) transparency, albeit this difference is not statistically significant.

A main effect of transparency on level 2 SA was found ( $F(3, 31) = 10.57, p < .001, \eta_p^2 = .51$ ). The SAGAT level 2 scores for medium (A) transparency level ( $M_{medium(A)} = .84$ ) are higher than the low- and high condition ( $M_{low} = .65, M_{high} = .64$ ). Also, the scores for the medium (B) condition are higher than the scores for the low condition and did not differ from the medium (A) condition ( $M_{medium(B)} = .85$ ; see Table 7). Furthermore, a main effect

of complexity on level 2 SA was found ( $F(1, 33) = 24.71, p < .001, \eta_p^2 = .43; M_{low} = .82, M_{high} = .67$ ). This indicates that a lower level 2 SA was achieved in high complexity cases compared to low complexity ones (see Table 8). Finally, a significant interaction between complexity and transparency was found for level 2 SA ( $F(3, 31) = 3.21, p < .037, \eta_p^2 = .24$ ) showing significant differences in level 2 SA scores for medium (A) transparency and complexity (see Table 9 and Figure 7).

A main effect of transparency on level 3 SA was found ( $F(3, 31) = 4.36, p < .011, \eta_p^2 = .30$ ). The scores on SAGAT were highest for the high transparency condition ( $M_{high} = .85$ ) and significantly higher than the low- and medium (A) transparency conditions ( $M_{low} = .73, M_{medium(A)} = .69$ ; see Table 7). No difference between the medium (B) transparency level and the other levels was found ( $M_{medium(B)} = .77$ ). A main effect for complexity was found in which the low complexity level resulted in higher scores on the SAGAT compared to the high complexity level ( $F(1, 33) = 38.60, p < .001, \eta_p^2 = .54; M_{low} = .85, M_{high} = .67$ ; see Table 8). No interaction between complexity and transparency was found for level 3 SA (see Table 9 and Figure 8).

**Table 8.** Means and Standard Deviations for the Dependent Variables as a Function of Complexity only. Note That TTC is Measured in Seconds.

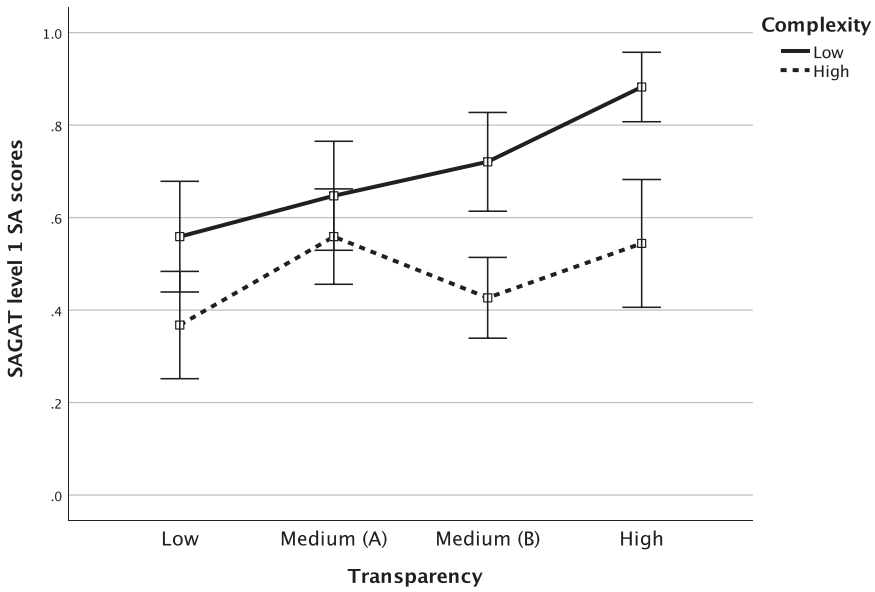
	Complexity			
	Low		High	
	M	SD	M	SD
Level 1 SA	.70	.18	.47	.16
Level 2 SA	.82	.16	.67	.16
Level 3 SA	.85	.15	.67	.18
WL	55.29	21.75	64.45	22.72
TTC	40.30	10.88	53.14	16.99

**Mental Workload**

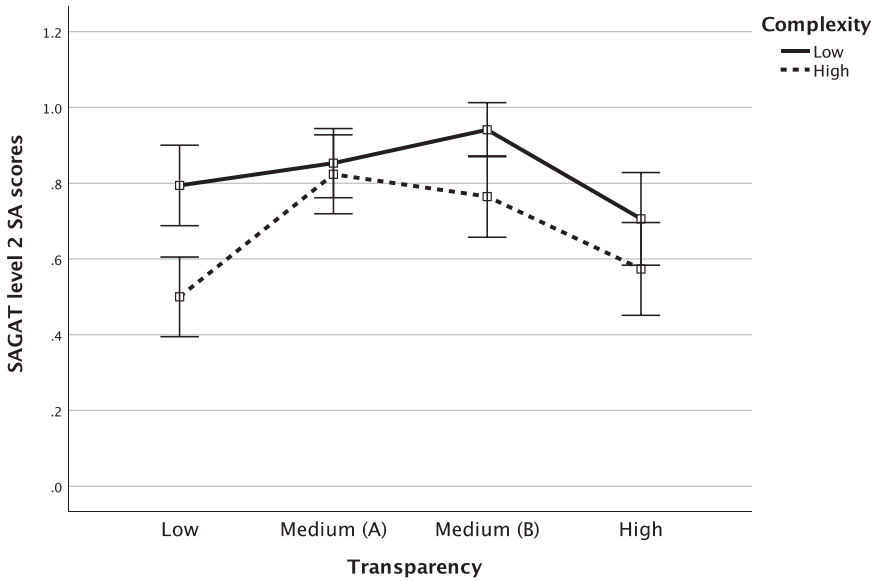
No main effect of transparency on mental workload was found (see Table 6). However,

**Table 9.** Means and Standard Deviations for the Dependent Variables as a Function of Level of Transparency and Complexity. Note That TTC is Measured in Seconds.

Complexity		Transparency Level							
		Low		Medium (A)		Medium (B)		High	
		M	SD	M	SD	M	SD	M	SD
Low	Level 1 SA	.56	.34	.65	.34	.72	.31	.88	.22
	Level 2 SA	.79	.30	.85	.26	.94	.20	.71	.35
	Level 3 SA	.82	.24	.78	.28	.85	.26	.94	.16
	WL	56.07	20.80	55.19	23.81	53.21	22.55	56.69	24.04
	TTC	31.81	13.38	45.50	11.30	36.55	13.61	47.32	11.98
High	Level 1 SA	.37	.33	.56	.30	.43	.25	.54	.40
	Level 2 SA	.50	.30	.82	.30	.77	.31	.57	.35
	Level 3 SA	.63	.33	.60	.27	.69	.28	.75	.35
	WL	64.31	22.22	61.54	22.05	67.03	23.85	64.91	26.66
	TTC	44.99	16.73	59.75	18.72	53.00	18.14	54.82	20.37



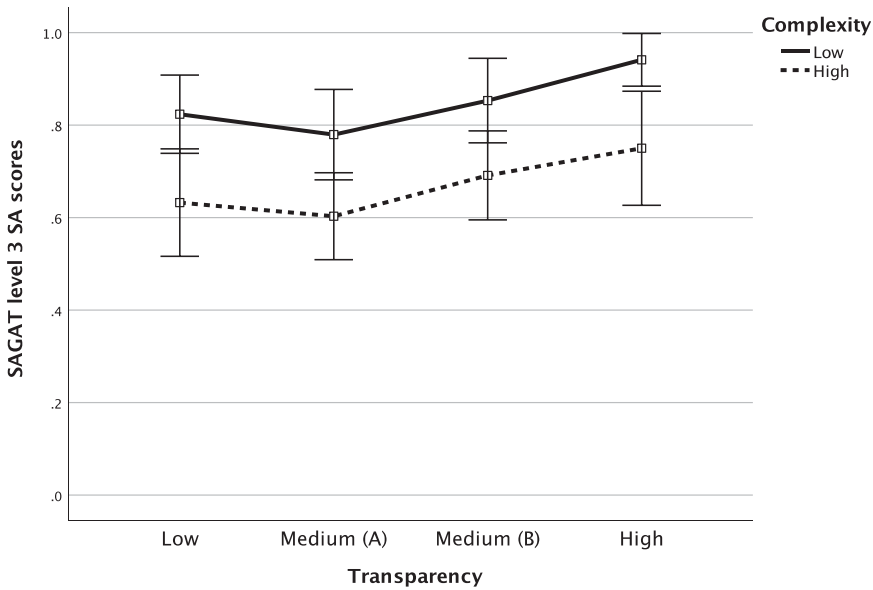
**Figure 6.** Mean scores for level 1 SA as a function of transparency and complexity. Note the error bars represent the 95% confidence interval.



**Figure 7.** Mean scores for level 2 SA as a function of transparency and complexity. Note the error bars represent the 95% confidence interval.

individual dimensions as measured through the NASA-TLX were analysed and showed an effect on the ‘Performance’ sub-dimension ( $F(3, 28) = 7.79, p < .001, \eta_p^2 = .46$ ). Here, the participants reported they were more satisfied with

‘achieving the goals set by the experimenter’ (Hart & Staveland, 1988, p. 30) for the medium (A) transparency level compared to the other levels. Also, a main effect for complexity on mental workload was found ( $F(1, 33) = 21.96,$



**Figure 8.** Mean scores for level 3 SA as a function of transparency and complexity. Note the error bars represent the 95% confidence interval.

$p < .001, \eta_p^2 = .40; M_{low} = 55.29, M_{high} = 64.45$ ). This indicates that participants reported higher levels of workload in the high complexity cases compared to the low complexity cases (see Table 8). Finally, no interaction between complexity and transparency was found (see Table 9 and Figure 9).

**Task Performance**

A main effect for transparency was found for mean TTC ( $F(3, 22) = 24.73, p < .001, \eta_p^2 = .77$ ). The medium (A)-, medium (B)-, and high transparency conditions ( $M_{medium(A)} = 52.62, M_{medium(B)} = 60.12, M_{high} = 60.80$ ) led to increased mean comprehension times compared to the low transparency condition ( $M_{low} = 38.40$ ). Also, the medium (A)- and high transparency conditions resulted in higher mean comprehension times compared to the low- and medium (B) conditions. No difference in TTC was found between the medium (A)- and high transparency conditions (see Table 7). For complexity, a main effect was found on the mean TTC ( $F(1, 24) = 46.65, p < .001, \eta_p^2 = .66; M_{low} = 40.30, M_{high} = 53.14$ ). A high traffic complexity resulted in increased mean comprehension times for the participants (see Table 8). For the interaction between transparency

and complexity no effect was found (see Table 9 and Figure 10).

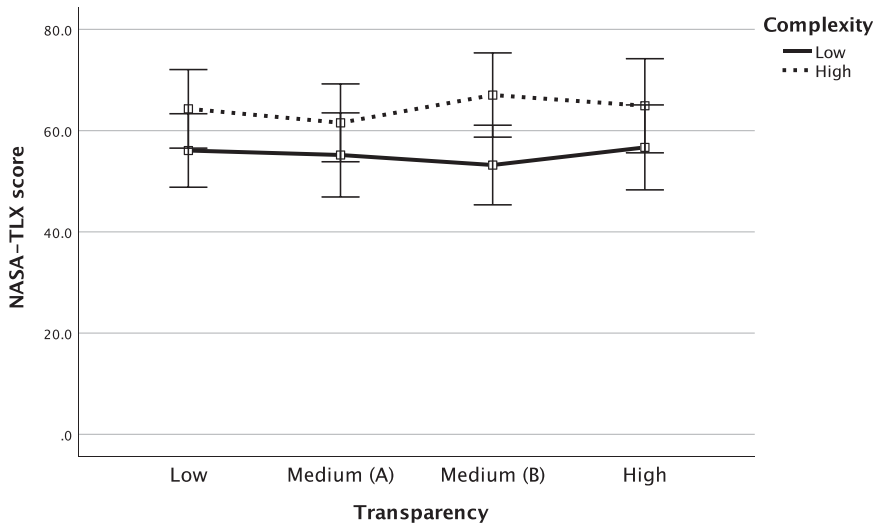
**Preference**

A main effect of transparency was found on the subjective ranking of the transparency levels ( $F(3, 31) = 616.64, p < .001, \eta_p^2 = .98$ ). The medium (A)- and high transparency levels were preferred compared to the low- and medium (B) levels. The low transparency was rated the least preferred, followed by the medium (B) level, and a shared highest preference for the medium (A)- and high transparency level (see Figure 11).

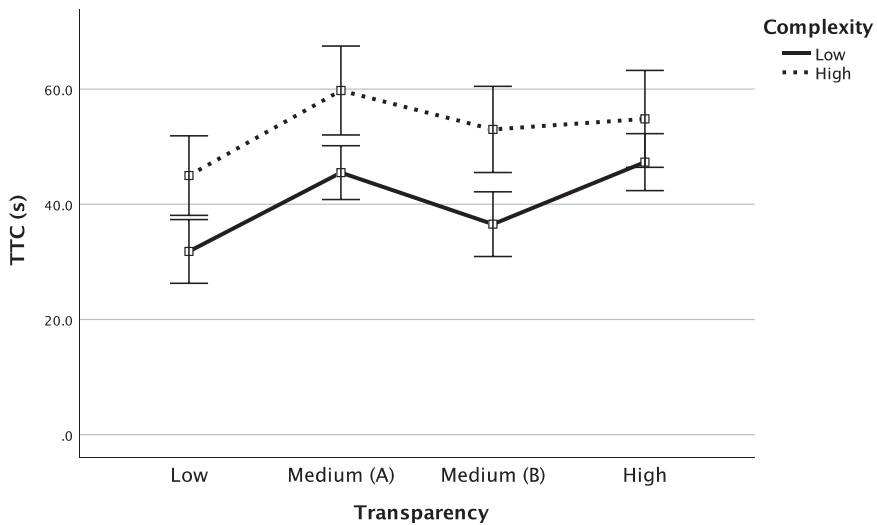
**Results Summary**

To summarise, the results from the experiment showed that SA improved with transparency, indicating that level 1 SA was highest for the high transparency condition, level 2 SA was highest in the medium (A) transparency condition, and level 3 SA was highest in the high transparency condition. For all SA measurements, high complexity traffic situations resulted in reduced levels of SA. Moreover, no significant effect of transparency on mental workload was





**Figure 9.** Mean scores for mental workload as a function of transparency and complexity. Note the error bars represent the 95% confidence interval.

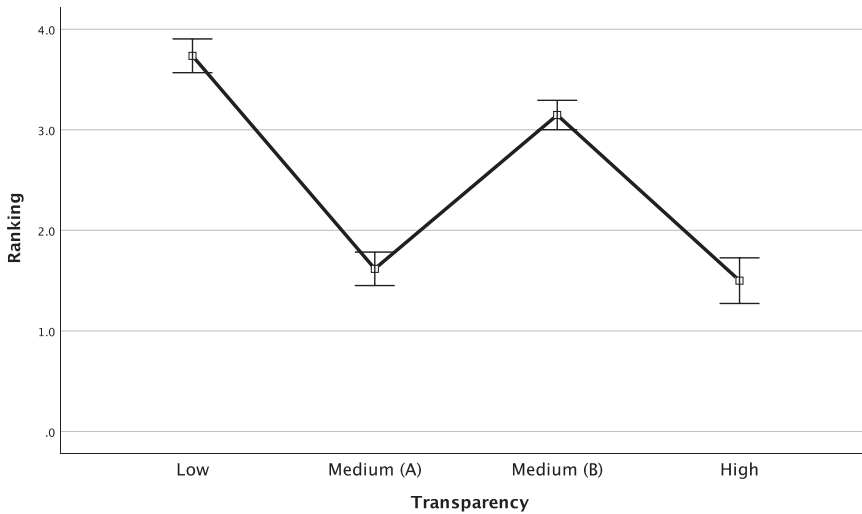


**Figure 10.** Mean scores for TTC as a function of transparency and complexity. Note the error bars represent the 95% confidence interval.

observed, although a significant effect for complexity was found showing that higher traffic complexity resulted in higher perceived mental workload. Furthermore, TTC was highest for the medium (A)- and high level. TTC was also highest for the high complexity traffic situations. Finally, the medium (A)- and high transparency levels were rated as the most preferred by the participants.

### Discussion

This study aimed to investigate the relationship between agent transparency, complexity, and selected human performance variables in a maritime autonomous collision avoidance context. Transparency was predicted to have a positive effect on SA and task performance without affecting mental workload. Complexity was predicted to have a



**Figure 11.** The participants' preferences for the transparency levels. Note that a lower score indicates a higher preference (most preferred = 1 and least preferred score = 4).

negative effect on SA, mental workload, and task performance. Finally, it was predicted that higher transparency levels could mitigate the effect of complexity on SA and task performance. No interaction effect was predicted for mental workload. The hypotheses and corresponding results are summarised in [Table 10](#).

### Situation Awareness

For level 1 SA, the highest SAGAT scores were achieved with the highest level of transparency. In Endsley's definition of SA (1995), level 1 SA is concerned with the perception of elements in their environment and provides the foundation for the higher levels of SA. In this study, it was anticipated that when the system provided information regarding its perception of its environment, that is, 'condition detection' (see [Table 4](#)), this would support level 1 SA. In this level of transparency, the CAGA system depicts which targets it has detected in the short and long range, the type of conflict with all detected targets, uncertainties in the sensor data, and the status of its sensors (see [Table 3](#)). This study anticipated that level 1 SA would be best for transparency levels in which the 'condition detection' information would be presented, that is, the medium (B)- and the high conditions. However, the results indicate that the highest level 1 SA scores were achieved only in the

high transparency condition and not in the medium (B) condition. Furthermore, no significant difference was found between the high- and the medium (A) transparency condition in terms of level 1 SA, indicating similar SAGAT scores. This may indicate that the information depicted in the 'condition analysis' step (e.g. risk objects, intended trajectories, and priorities; absent in the medium (B) transparency condition yet present in the medium (A) condition) may have played a role in achieving improved level 1 SA. Possibly, the additional information regarding collision risk have made the participants more observant of the ship's surrounding traffic and thus better able to achieve level 1 SA.

For level 2 SA, the highest level of SA was achieved with the medium (A) level of transparency regardless of complexity level. Again, this is as hypothesized as it is at this level the system's analysis is depicted on the HMI and made available to the supervisor, for example, depicting risk objects, risk priorities, intended trajectories, conflict type, and safe speed parameters (see [Table 3](#) and [Table 4](#)). However, the same level of level 2 SA was also achieved for the medium (B) level of transparency compared to the medium (A) level. In the medium (B) level of transparency, the CAGA system depicts which objects were detected in the short and long range, target type, relative motion, status and uncertainties of sensor data, that is no

**Table 10.** Summary of Predictions and Results Regarding the Effect of Transparency and Complexity on Situation Awareness, Mental Workload, and Task Performance.

Measure	Impact of Transparency	Results Match Prediction?	Impact of Complexity	Results Match Prediction?	Interaction	Results Match Prediction?
SA	Improved SA with increased transparency	Level of SA: 1: Yes 2: Yes 3: Yes	Reduced SA with high complexity	Level of SA: 1: Yes 2: Yes 3: Yes	Increased transparency may negate effect of high complexity	Level of SA: 1: No 2: Yes 3: No
Mental workload	No effect predicted	Yes	Increased mental workload with high complexity	Yes	No interaction predicted	Yes
Task performance	Improved task performance with increased transparency	No	Reduced task performance with high complexity	Yes	Increased transparency may negate effect of high complexity	No

analytical information, yet participants were able to achieve equally high level 2 scores compared to the medium (A) level, where the system's analytical information was readily available. For example, at the medium (A) level of transparency, the system depicts which objects it sees as posing a collision danger by extrapolating the objects' current vector and highlighting the level of risk using specific symbology and colours. This way, participants could directly perceive the outcomes of the system's risk analysis process and use this information to understand the system's interpretation of the traffic situation. In addition to the medium (B) results, what is somewhat unexpected is that the same level of SA was not achieved in the high level of transparency condition. As the high transparency level includes all information from the medium (A) transparency level, that is, also the analytical information (see Table 3 and Table 4), one could reasonably expect that participants would score equally well on level 2 SA for both the medium (A)- and high transparency conditions. As this is not the case, one explanation may be that the additional information about the system's detection and sensor information, as shown in the high transparency case (see Figure 5), may have distracted the participants in establishing an understanding of the system's analysis.

Finally, for level 3 SA, the highest level was achieved with the highest transparency level. No differences were observed between the low and medium (A) transparency levels. To support level 3 SA, the system provided the future state prediction of own ship and target objects. The future state of own ship, that is, its future track and speed (see Table 4), was depicted for each level of transparency. The future state of target ships was depicted for the medium (A)- and high transparency levels but not for the other levels. As such, it would follow that either all transparency levels scored equal on level 3 SA, or that the medium (A)- and high transparency levels scored equal. However, given that only the high transparency level resulted in the highest level 3 SA scores makes this finding somewhat challenging to interpret. One explanation is that the high transparency level provided the complete picture of the system's interpretation of the traffic situation: its decision and future actions, its analysis, and its object detections, including sensors states. Possibly, providing participants with a complete information overview allowed them to understand own ship's future state more adequately, considering that they now have a more comprehensive information basis to build this on. In addition, based on the full picture, participants may be better

able to reason towards the correct answer when answering the SAGAT.

For traffic complexity, SAGAT scores were lower for the high complexity traffic situations indicating it was more challenging to achieve a similar level of SA in the high complexity cases compared to the low complexity ones. This finding is consistent with earlier observations where increased number of objects presented to a supervisor, including their interactions, increases the number of goals and decisions to be made which, given the limitations of human information processing capabilities, will have an effect on how well SA can be achieved (Endsley, 1995). In terms of interactions between transparency and complexity, an effect was found for level 2 SA pointing towards a positive contribution of the depiction of the system's reasoning, for example, risk objects, intended trajectories, and priorities, as present in the medium (A) transparency level, for high complexity cases.

Comparing our results to similar studies in which the relationship between transparency and SA was investigated, we find comparable results. For example, Roth et al. (2020) found improvements in SAGAT scores when participants were evaluating agent-generated proposals in an unmanned-manned helicopter teaming operation. In their study, level 3 SA was most improved in the high transparency condition compared to the low condition. Chen et al. (2014b, 2015) found improvements in SA when participants were supervising unmanned aerial vehicles in a search operation, and Selkowitz et al. (2017) reported improved SAGAT scores when monitoring an autonomous robot for level 2 and 3 SA, but not for level 1. However, some studies failed to identify a relationship between transparency and SA for supervision (Skraaning & Jamieson, 2021; Experiment 3) and monitoring tasks (Pokam et al., 2019; Selkowitz, Lakhmani, Chen, & Boyce, 2015; Wright et al., 2020). Overall, these studies point towards an overall neutral to positive relationship between transparency and SA, and this study has strengthened these findings.

### *Mental Workload*

No effect of transparency on mental workload was found. For complexity, increased workload scores

were found for all high complexity traffic situations, but there was no interaction effect with transparency.

Still, for one sub-dimension of the NASA-TLX scale: 'Performance' a significant relationship between transparency and mental workload was found. Here, participants rated their own performance in relation to the experimental task as better for the medium (A) transparency level compared to the other transparency levels. In other words, as the experimental task was to understand the traffic situation and the system's handling of it, participants felt they achieved this best in the medium (A) transparency condition. Possibly, participants felt they had sufficient information in the medium (A) condition and therefore felt they were able to meet the goals of the experiment.

When comparing these results to similar studies where participants were tasked with monitoring an autonomous agent only, limited effects of transparency on mental workload were also reported (e.g. Du et al., 2019; Selkowitz et al., 2015, 2017; Wright et al., 2020). A study by Panganiban et al. (2020) found a reduction in mental workload as measured through the NASA-TLX when an autonomous agent communicated its intentions to support the participant in its task execution. Conversely, a study by Selkowitz et al. (2017) reported an increase in eye-fixation duration, a measure of visual search and mental processing (Di Nocera, Camilli, & Terenzi, 2007; Harris, Glover, & Spady, 1986), when monitoring an autonomous robot's display for its actions.

In studies where participants took the role as a supervisor of an autonomous agent, mostly reductions in workload were found (T. Chen et al., 2014b, e.g. 2015; Skraaning & Jamieson, 2021; Experiment 1 and 2), although an increase (Guznov et al., 2020) and no effect (Skraaning & Jamieson, 2021; Experiment 3) were also reported. Finally, in studies where participants were asked to respond to system-generated proposals, no effect on mental workload was reported (e.g. Bhaskara et al., 2021; Loft et al., 2023; Mercado et al., 2016; Roth et al., 2020; Stowers et al., 2020).

This may imply that the relationship between transparency and mental workload depends on the type of task and role given to the participant (van de Merwe, et al, 2024a). In this experiment, participants did not interact with the autonomous

CAGA system as they were only asked to perceive and comprehend its information. Although several of the studies mentioned above found a relationship between transparency and mental workload, 17 out of 23 indicators, as reported in the study by [van de Merwe, et al, 2024a](#) did not. This experiment's result does not change the overall conclusion that adding information that supports transparency to an HMI has a limited effect on mental workload.

### Task Performance

The results indicate participants take more time in building up a mental picture in the medium (A)- and high transparency conditions and less time in the low- and medium (B) transparency conditions. Participants consistently took more time to comprehend the traffic situation in the medium (A)- and high levels compared to the low transparency level. This was the case for both the low- and high complexity conditions indicating an equal effect of traffic complexity regardless of transparency level. The results were inconsistent with the hypothesis that the cognitive processes associated with developing a mental picture of the traffic situation would be supported when much of the information needed was readily available on the HMI for the higher transparency cases. It was also hypothesized that this effect would be stronger for the high complexity condition than the low complexity condition, but this was not the case.

Earlier studies have shown inconsistent effects for time-related performance measures associated with transparency. A recent study investigating the impact of transparency on decision risk in human-agent teams measured the time it took for participants to choose between two options suggested by a recommender system ([Loft et al., 2023](#)). No differences between various levels of transparency and decision time were found, except for an interaction between decision time and decision risk indicating that transparency alleviated the negative effect of increased risk on response time. A study performed by [Skraaning and Jamieson \(2021\)](#) found reduced response times to events in a nuclear control room simulation study. Here, control room operators were tasked with controlling a simulated nuclear power plant and handle small to large system upsets, including taking corrective

action. A reduction in response time to system upsets were found in the transparency condition indicating a better task performance when information that supports transparency was integrated in the primary task HMI. Conversely, a study by [Stowers et al. \(2020\)](#) found an increase in response time with increased levels of transparency. In this study, participants were tasked with monitoring and controlling multiple unmanned vehicles and evaluate plans for these provided by an intelligent agent. Here, the addition of information that supports transparency in the form of basic projection and uncertainty information significantly increased response time, albeit with a small effect size. Finally, [Wright et al. \(2020\)](#) found no difference in the time participants took to identify and assess events when monitoring an autonomous robot.

In our study, response time was driven by the instruction for the participants to 'continue to the next step when you feel you have built up a sufficient understanding of the traffic situation', that is, the time needed for comprehension. In contrast with the aforementioned studies, in which participants were asked to evaluate plans, respond to events, or monitor autonomous agents, this study asked participants to build a mental representation of the traffic situation only. Considering that there were no significant differences in TTC between the medium (A)- and high transparency conditions and that both showed significantly higher TTCs than the low- and medium (B) conditions indicates that the analytical information contributed to the participants' time needed to comprehend the traffic situations. Conversely, this also implies that the addition of the system's detection information did not contribute to the participants' TTC.

Considering [Table 3](#) and [Table 4](#), the information presented in the condition analysis step, represented in the medium (A)- and high transparency conditions, depicts elements primarily concerned with collision risk, for example, objects that pose a risk, risk object priority, conflict type, and their predicted course and speed. This information is essential in understanding the CAGA system's risk determination and is the primary basis for interpreting the reasoning behind its avoidance actions. The information in the condition detection step, represented in the low- and medium (B) transparency conditions, primarily

consists of elements depicting what the ship has detected, for example, objects in the short and long range, object type and size, and basic classification of relative motion. That is, whereas the analytical information is specific to objects posing a risk, the detection information covers all objects irrespective of risk.

In this experiment, the participants, all experienced navigators, took the role of a supervisor of a ship equipped with a CAGA system with the task to observe and understand the system's depicted solutions to traffic conflict situations. Since the system's analysis and avoidance actions are the most safety critical information to understand, participants may have taken additional time to evaluate the analytical information provided by the CAGA system, as presented in the medium (A)- and high transparency conditions, because they wanted to understand the situation as accurately as possible. The correlational results between TTC and SA support this assumption as participants with higher TTC values also have higher level 1 SA and level 3 SA scores. In other words, those that spent more time observing, interpreting, and understanding the traffic situations also scored better on the SAGAT. Similar results have been reported in eye-tracking studies where increased focus on critical information elements was correlated with improved SA (van de Merwe, et al, 2012). Alternatively, participants in the medium (A)- and high transparency conditions may also have taken more time to analyse the traffic situations because they were comparing CAGA's analysis with their own. That is, rather than taking the system's interpretation of the traffic situation at face value, the participants may have performed their own analysis first to ensure they were equipped with sufficient knowledge to be able to scrutinise the systems. Also, given that the CAGA system's analytical information was not depicted in the low- and medium (B) transparency conditions, the TTC was less than the medium (A) and high transparency conditions because there was less critical information to evaluate and compare. Similar observations have been reported when operators are required to evaluate recommendations and need to compare these to system information and other information sources (Endsley, 2017). As such, considering the potential role of humans in the ship autonomy context where a

thorough understanding of the CAGA system's performance is essential for supervisory performance (van de Merwe et al., 2024b), this finding demonstrates the importance of addressing the type of information in developing transparent agents and not only the amount.

### *Practical Considerations*

The results of this study imply that transparency has value as a design principle for designing CAGA systems given the positive results for SA. In addition, the qualitative feedback from the navigators about which of the levels of transparency they prefer clearly indicates a positive attitude towards HMIs depicting the system's analytical information at minimum. Conversely, these results also clearly indicate which of the transparency levels were not preferred. For example, the low transparency level, that is, where the system only showed its decisions and planned actions, was the least preferred. In addition, the medium (B) transparency level, that is, where the system's analytical information was not depicted, ranked just slightly better than the low level. Clearly, our participants preferred to have information about the system's analytical information in addition to its decisions and planned actions, as indicated by the shared highest ranking of the medium (A) and high transparency levels. Nevertheless, there is no clear result pointing towards the optimal level of transparency across our dependent variables. This means that, when designing for transparency, it may be challenging to decide on which level to implement. Possibly, a more demand-driven transparency, that is, where users adjust the level of transparency depending on the task and context, can be used to provide control to the supervisor over the amount of system information presented. A study by Vered et al. (2020) demonstrated that such an approach could avoid the downsides of presenting transparency information whilst maintaining its benefits. For example, when applied to autonomous shipping, supervisors may only depict a low level of transparency in situations with little to no traffic whilst 'dialling up' the level of transparency for situations that require closer supervision. This way, this approach may improve comprehension times compared to the sequential transparency

approach as used in our study. However, a potential risk associated with this approach is the potential for choosing an inappropriate transparency level and thereby overseeing important information. Furthermore, this approach allows for potentially large variation in how information is presented on the HMI and the possibility for confusion regarding which level is active. Although an iterative and human-centred design process should address these concerns when developing HMIs, future studies should investigate these risks further.

### Limitations and Future Work

This experiment adjusted the transparency of a CAGA system for which information was overlaid onto static radar images. Our approach assumed that future operators of autonomous ships may need to divide their attention between multiple ships and/or tasks and may not continuously monitor a single ship. Therefore, when a ship requires attention, the supervisor may be 'dropped-into' the specifics of the operational traffic situation. Our study hypothesized that transparency facilitates this sense-making process needed to quickly build SA. However, despite significant effort put into making the traffic situations as realistic as possible, real-world situations are, of course, dynamic. As such, in dynamic situations supervisors would be able to build a mental representation of the developing traffic situation over time. Although this study provided insights into the effects of transparency on human performance variables in a maritime collision avoidance setting, future research should focus on the application of transparency implementation in dynamic settings, for example, by using real-time simulation facilities.

In this experiment, the CAGA system provided information about its perceptions, analysis, and future intentions regarding a traffic situation to the participants. Participants were only required to answer SA queries about the traffic situation and the system's proposed handling of it. Through the development of the traffic situations and the transparency levels, significant effort was put into ensuring that the system provided sound conflict resolutions such that disagreements between the participants' solution to a situation and the

system's solution were kept at a minimum and would not confound the results (van de Merwe et al., 2023a). As such, this experiment did not study the effects of incorrect resolutions or solutions that which the supervisor disagreed with. However, given the body of knowledge available about the potential pitfalls for humans in supervising automation (Endsley, 2017; Onnasch et al., 2014; Strauch, 2018), future work should elaborate on the effect of transparency on the supervisor's ability to detect and resolve performance deviations, especially when performing under concurrent task demands, such as supervising multiple autonomous ships (Burmeister et al., 2014; Gegoff et al., 2023; Porathe, 2014; Tataschiere et al., 2023).

### Conclusions

This study highlighted the relationships between agent transparency and human performance variables, SA, mental workload, and task performance. Our overall findings point towards improvements in all levels of SA as a consequence of transparency, albeit that different levels of transparency affect different levels of SA. In addition, this study found that more time was needed to create a mental representation of the situation when the system's reasoning was depicted. Interestingly, no significant correlations between mental workload and SA, and mental workload and TTC were found. Given the relationship between task performance, SA, and mental workload (Wickens, Hollands, Banbury, & Parasuraman, 2013), these findings indicate an effort-performance trade-off where participants with increased SA scores also used more time to comprehend the traffic situations, albeit without increased mental workload ratings. Moreover, this study showed clear and consistent effects of complexity on both SA scores, workload ratings, and TTC, consistent with predictions from earlier models (e.g. Endsley, 1995, 2017). No interaction effects between transparency and complexity were found, except for level 2 SA, where transparency negated the effect of traffic complexity. Finally, the medium (A)- and high transparency levels were also the most preferred by the participants.

To summarise, as agent transparency is frequently operationalised through an HMI, our results imply that agent transparency has merits as a

design philosophy when developing highly automated systems that require human supervision (e.g. see MITRE, 2018 for guidance). However, implementing transparency 'is as much an art as it is a science' given the risk of visual clutter and potential distractions caused by additional information (Wickens, 2018, p. 39). Also, the exact operationalisation of transparency depends on the domain it is applied to and the function allocation between humans and systems (Holder, Huang, Chiou, Jeon, & Lyons, 2021). Although, there is limited evidence-based guidance available for designers to develop transparent agents (Jamieson et al., 2022), this study demonstrated that, by basing the transparency design on a structured human-centred design approach, the purported effects of clutter and information overload were kept to a minimum whilst achieving improvements in SA. Hence, given supervisors have sufficient time available to process the additional transparency information, improved levels of SA may be achieved without burdening supervisors with additional mental workload. As such, if effort is made to integrate information supporting transparency in the primary task interface, human performance benefits can be expected.

### Acknowledgements

The authors would like to express their sincere gratitude to the navigators for their participation in the experiment and workshops. We would also like to express our sincere gratitude to Koen Houweling, MSc. for his contribution in developing the traffic situations and the transparency illustrations. Finally, we would like to thank the anonymous reviewers for their significant contributions and reflections on the work.

### Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is sponsored by the Research Council of Norway, project nr. 311365 and 327903.

### ORCID iD

Koen van de Merwe  <https://orcid.org/0000-0002-0168-872X>

### Supplemental Material

Supplemental material for this article is available online.

### References

- Alsos, O. A., Hodne, P., Skåden, O. K., & Porathe, T. (2022). Maritime autonomous surface ships: Automation transparency for nearby vessels. *Journal of Physics: Conference Series*, 2311(1), 012027. <https://doi.org/10.1088/1742-6596/2311/1/012027>
- ASKO. (2022). *Verdens første batterielektriske autonome sjødroner har ankommet Norge!*. ASKO. Retrieved May 10, 2022, from <https://asko.no/nyhetsarkiv/verdens-forste-autonome-sjodroner-har-ankommet-norge/>
- Bhaskara, A., Duong, L., Brooks, J., Li, R., McInerney, R., Skinner, M., Pongracic, H., & Loft, S. (2021). Effect of automation transparency in the management of multiple unmanned vehicles. *Applied Ergonomics*, 90, 103243. <https://doi.org/10.1016/j.apergo.2020.103243>
- Bhaskara, A., Skinner, M., & Loft, S. (2020). Agent transparency: A review of current theory and evidence. *IEEE Transactions on Human-Machine Systems*, 50(3), 215–224. <https://doi.org/10.1109/THMS.2020.2965529>
- Burmeister, H.-C., Bruhn, W., Rødseth, Ø. J., & Porathe, T. (2014). Autonomous unmanned merchant vessel and its contribution towards the e-navigation implementation: The MUNIN perspective. *International Journal of E-Navigation and Maritime Economy*, 1, 1–13. <https://doi.org/10.1016/j.enavi.2014.12.002>
- Chen, J. Y. C., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. J. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, 19(3), 259–282. <https://doi.org/10.1080/1463922X.2017.1315750>
- Chen, J. Y. C., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. J. (2014a). *Situation awareness-based agent transparency (No. ARL-TR-6905)*. U.S. Army Research Laboratory. Aberdeen Proving Ground. <https://doi.org/10.21236/ADA600351>
- Chen, T., Campbell, D. A., Gonzalez, F., & Coppin, G. (2014b). The effect of autonomy transparency in



- human-robot interactions: A preliminary study on operator cognitive workload and situation awareness in multiple heterogeneous uav management. In *Proceedings of Australasian conference on robotics and automation 2014*. Australian Robotics and Automation Association. Retrieved from. <https://www.araa.asn.au/acra/acra2014/papers/pap166.pdf>
- Chen, T., Campbell, D. A., Gonzalez, L. F., & Coppin, G. (2015). Increasing Autonomy Transparency through capability communication in multiple heterogeneous UAV management. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 2434–2439. IEEE. <https://doi.org/10.1109/IROS.2015.7353707>
- Cummings, M. L., & Guerlain, S. (2007). Developing operator capacity estimates for supervisory control of autonomous vehicles. *Human Factors*, 49(1), 1–15. <https://doi.org/10.1518/001872007779598109>
- Di Nocera, F., Camilli, M., & Terenzi, M. (2007). A random glance at the flight deck: Pilots' scanning strategies and the real-time assessment of mental workload. *Journal of Cognitive Engineering and Decision Making*, 1(3), 271–285. <https://doi.org/10.1518/155534307X255627>
- Doom, E. V., Rusák, Z., & Horváth, I. (2017). A situation awareness analysis scheme to identify deficiencies of complex man-machine interactions. *International Journal of Information Technology and Management*, 16(1), 53–72. <https://doi.org/10.1504/IJITM.2017.080958>
- Doshi-Velez, F., & Kim, B. (2017). Towards A rigorous science of interpretable machine learning. *arXiv: 1702.08608 [Cs, Stat]*. Retrieved from. <https://arxiv.org/abs/1702.08608>
- Du, N., Haspiel, J., Zhang, Q., Tilbury, D., Pradhan, A. K., Yang, X. J., & Robert, L. P. (2019). Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transportation Research Part C: Emerging Technologies*, 104, 428–442. <https://doi.org/10.1016/j.trc.2019.05.025>
- Eccles, D. W., & Arsal, G. (2017). The think aloud method: What is it and how do I use it? *Qualitative Research in Sport, Exercise and Health*, 9(4), 514–531. <https://doi.org/10.1080/2159676X.2017.1331501>
- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Error in Aviation*, 37(1), 217–249. <https://doi.org/10.4324/9781315092898-13>
- Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. *Human Factors*, 59(1), 5–27. <https://doi.org/10.1177/0018720816681350>
- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Error in Aviation*, 37(1), 217–249. <https://doi.org/10.4324/9781315092898-13>
- Endsley, M. R. (2023). Supporting human-AI teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior*, 140, 107574. <https://doi.org/10.1016/j.chb.2022.107574>
- Endsley, M. R., Bolté, B., & Jones, D. G. (2003). *Designing for situation awareness: An approach to user-centered design*: Taylor and Francis.
- Endsley, M. R., & Garland, D. J. (Eds.), (2000). *Situation awareness: Analysis and measurement*. Lawrence Erlbaum Associates.
- Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(2), 381–394. <https://doi.org/10.1518/001872095779064555>
- Ezenyilimba, A., Wong, M., Hehr, A., Demir, M., Wolff, A., Chiou, E., & Cooke, N. (2023). Impact of transparency and explanations on trust and situation awareness in human-robot teams. *Journal of Cognitive Engineering and Decision Making*, 17(1), 75–93. <https://doi.org/10.1177/15553434221136358>
- Gawron, V. J. (2019). *Human performance and situation awareness measures* (3rd ed.). CRC Press/Taylor and Francis Group.
- Gegoff, I., Tatasciore, M., Bowden, V., McCarley, J., & Loft, S. (2023). Transparent automated advice to mitigate the impact of variation in automation reliability. *Human Factors*, 0(0), 00187208231196738. <https://doi.org/10.1177/00187208231196738>
- Guznov, S., Lyons, J. B., Pfahler, M., Heironimus, A., Woolley, M., Friedman, J., & Neimeier, A. (2020). Robot transparency and team orientation effects on human-robot teaming. *International Journal of Human-Computer Interaction*, 36(7), 650–660. <https://doi.org/10.1080/10447318.2019.1676519>
- Harris, R. L., Glover, B. J., & Spady, A. A. (1986). Analytical techniques of pilot scanning behavior and their application. Retrieved from. <https://ntrs.nasa.gov/citations/19860018448>
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical

- and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in psychology* (pp. 139–183). North-Holland. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Holder, E., Huang, L., Chiou, E., Jeon, M., & Lyons, J. B. (2021). Designing for Bi-directional transparency in human-AI-robot-teaming. *Proceedings of the Human Factors and Ergonomics Society - Annual Meeting*, 65(1), 57–61. <https://doi.org/10.1177/1071181321651052>
- IEC. (2022). *IEC 62288:2022 Maritime navigation and radiocommunication equipment and systems*. International Electrotechnical Commission.
- IMO. (1977). *Convention of the international regulations for preventing collisions at sea (COLREGS)*. International Maritime Organisation.
- IMO. (2018). *Maritime Safety Committee (MSC)*. IMO, 100th session, 3-7 December 2018. Retrieved October 23, 2020, from International Maritime Organisation website. <https://www.imo.org/en/MediaCentre/MeetingSummaries/Pages/MSC-100th-session.aspx>
- ISO. (2019). *ISO 9241-210:2019 Ergonomics of human-system interaction—Part 210: Human-centred design for interactive systems*. International Organization for Standardization.
- Jamieson, G. A., Skraaning, G., & Joe, J. (2022). The B737 MAX 8 accidents as operational experiences with automation transparency. *IEEE Transactions on Human-Machine Systems*, 52(4), 794–797. <https://doi.org/10.1109/THMS.2022.3164774>
- Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019). Automation transparency: Implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics*, 62(3), 345–360. <https://doi.org/10.1080/00140139.2018.1547842>
- Lipton, Z. C. (2017). The mythos of model interpretability. *arXiv:1606.03490 [Cs, Stat]*. Retrieved from <https://arxiv.org/abs/1606.03490>
- Loft, S., Bhaskara, A., Lock, B. A., Skinner, M., Brooks, J., Li, R., & Bell, J. (2023). The impact of transparency and decision risk on human-automation teaming outcomes. *Human Factors*, 65(5), 846–861. <https://doi.org/10.1177/00187208211033445>
- Loft, S., Bhaskara, A., Lock, B. A., Skinner, M., Brooks, J., Li, R., & Bell, J. (2023). The impact of transparency and decision risk on human-automation teaming outcomes. *Human Factors*, 65(5), 846–861. <https://doi.org/10.1177/00187208211033445>
- Loïck, S., Guérin, C., Rauffet, P., Chauvin, C., & Éric, M. (2023). *How humans comply with a (potentially) faulty robot: Effects of multidimensional transparency*. (p. 1–10). *IEEE Transactions on Human-Machine Systems*. <https://doi.org/10.1109/THMS.2023.3273773>
- Lyons, J. B. (2013). Being transparent about transparency: A model for human-robot interaction. In *2013 AAAI spring symposium series*. Stanford University.
- Massterly. (2023). A snapshot of some of the projects we are involved in. Retrieved August 5, 2023, from <https://www.massterly.com/news-1>
- Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human-agent teaming for multi-UxV management. *Human Factors*, 58(3), 401–415. <https://doi.org/10.1177/0018720815621206>
- Metzger, U., & Parasuraman, R. (1999). Free flight and the air traffic controller: Active control versus passive monitoring. *Proceedings of the Human Factors and Ergonomics Society - Annual Meeting*, 43(1), 1–5. <https://doi.org/10.1177/154193129904300101>
- MITRE. (2018). *Human-machine teaming systems engineering guide (No. MP180941)* (p. 68). MITRE Corporation. Retrieved from MITRE Corporation website: <https://www.mitre.org/publications/technical-papers/human-machine-teaming-systems-engineering-guide>
- Moacdieh, N., & Sarter, N. (2015a). Data density and poor organization: Analyzing the performance and Attentional effects of two aspects of display clutter. *Proceedings of the Human Factors and Ergonomics Society - Annual Meeting*, 59(1), 1336–1340. <https://doi.org/10.1177/1541931215591221>
- Moacdieh, N., & Sarter, N. (2015b). Display clutter: A review of definitions and measurement techniques. *Human Factors*, 57(1), 61–100. <https://doi.org/10.1177/0018720814541145>
- Moacdieh, N., & Sarter, N. (2017). The effects of data density, display organization, and stress on search performance: An eye tracking study of clutter. *IEEE Transactions on Human-Machine Systems*, 47(6), 886–895. <https://doi.org/10.1109/THMS.2017.2717899>
- National Academies of Sciences. (2022). *Engineering and medicine. Human-AI teaming: State of the art and research needs*: The National Academies Press. <https://doi.org/10.17226/26355>

- NYK. (2022). *NYK group companies participate in trial to simulate the actual operation of fully autonomous ship*. NYK. Retrieved August 25, 2022, from [https://www.nyk.com/english/news/2022/20220303\\_02.html](https://www.nyk.com/english/news/2022/20220303_02.html)
- Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human Factors*, 56(3), 476–488. <https://doi.org/10.1177/0018720813501549>
- Ososky, S., Sanders, T., Jentsch, F., Hancock, P., & Chen, J. Y. C. (2014). Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In R. E. Karlssen, D. W. Gage, C. M. Shoemaker, & G. R. Gerhart (Eds.), *Proceedings volume 9084: Unmanned systems Technology XVI*. SPIE Defense + Security. <https://doi.org/10.1117/12.2050622>
- Panganiban, A. R., Matthews, G., & Long, M. D. (2020). Transparency in autonomous teammates: Intention to support as teaming information. *Journal of Cognitive Engineering and Decision Making*, 14(2), 174–190. <https://doi.org/10.1177/1555343419881563>
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans: A Publication of the IEEE Systems, Man, and Cybernetics Society*, 30(3), 286–297. <https://doi.org/10.1109/3468.844354>
- Pokam, R., Debernard, S., Chauvin, C., & Langlois, S. (2019). Principles of transparency for autonomous vehicles: First results of an experiment with an augmented reality human–machine interface. *Cognition, Technology and Work*, 21(4), 643–656. <https://doi.org/10.1007/s10111-019-00552-9>
- Porathe, T. (2014). Remote monitoring and control of unmanned vessels –the MUNIN shore control Centre. In *Proceedings of the 13th international conference on computer applications and information Technology in the maritime industries (COMPIT '14)* (pp. 460–467). May.
- Porathe, T. (2021). Human-automation interaction for autonomous ships: Decision support for remote operators. *TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation*, 15(3), 511–515. <https://doi.org/10.12716/1001.15.03.03>
- Porathe, T., Fjortoft, K., & Bratbergsengen, I. L. (2020). Human Factors, autonomous ships and constrained coastal navigation. *IOP Conference Series: Materials Science and Engineering*, 929(1), 012007. <https://doi.org/10.1088/1757-899X/929/1/012007>
- Psychology Software Tools, Inc. (2023). *E-Prime 3.0*. Psychology Software Tools, Inc. Retrieved from <https://support.pstnet.com/>
- Ramos, M. A., Utne, I. B., & Mosleh, A. (2019). Collision avoidance on maritime autonomous surface ships: Operators' tasks and human failure events. *Safety Science*, 116, 33–44. <https://doi.org/10.1016/j.ssci.2019.02.038>
- Roth, G., Schulte, A., Schmitt, F., & Brand, Y. (2020). Transparency for a workload-adaptive cognitive agent in a manned–unmanned teaming application. *IEEE Transactions on Human-Machine Systems*, 50(3), 225–233. <https://doi.org/10.1109/THMS.2019.2914667>
- Russell, S. J., & Norvig, P. (2022). *Artificial intelligence: A modern approach* (4th ed., global edition): Pearson.
- Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, 29(4), 260–278. <https://doi.org/10.1080/12460125.2020.1819094>
- Selkowitz, A. R., Lakhmani, S. G., & Chen, J. Y. C. (2017). Using agent transparency to support situation awareness of the Autonomous Squad Member. *Cognitive Systems Research*, 46, 13–25. <https://doi.org/10.1016/j.cogsys.2017.02.003>
- Selkowitz, A. R., Lakhmani, S. G., Chen, J. Y. C., & Boyce, M. (2015). The effects of agent transparency on human interaction with an autonomous robotic agent. *Proceedings of the Human Factors and Ergonomics Society - Annual Meeting*, 59(1), 806–810. <https://doi.org/10.1177/1541931215591246>
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4), 591–611. <https://doi.org/10.2307/2333709>
- Simic, V., & Alsos, O. A. (2023). Automation transparency: Designing an external HMI for autonomous passenger ferries in urban waterways. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference (1145–1158)*. ACM. <https://doi.org/10.1145/3563657.3596130>
- Skraaning, G., & Jamieson, G. A. (2021). Human performance benefits of the automation transparency design principle: Validation and variation. *Human*

- Factors*, 63(3), 379–401. <https://doi.org/10.1177/0018720819887252>
- Stowers, K., Kasdaglis, N., Rupp, M. A., Newton, O. B., Chen, J. Y. C., & Barnes, M. J. (2020). The IMPACT of agent transparency on human performance. *IEEE Transactions on Human-Machine Systems*, 50(3), 245–253. <https://doi.org/10.1109/THMS.2020.2978041>
- Strauch, B. (2018). Ironies of automation: Still unresolved after all these years. *IEEE Transactions on Human-Machine Systems*, 48(5), 419–433. <https://doi.org/10.1109/THMS.2017.2732506>
- Tataciore, M., Bowden, V., & Loft, S. (2023). Do concurrent task demands impact the benefit of automation transparency? *Applied Ergonomics*, 110, 104022. <https://doi.org/10.1016/j.apergo.2023.104022>
- van de Merwe, K., Mallam, S., Engelhardtsen, Ø., & Nazir, S. (2023a). Operationalising automation transparency for maritime collision avoidance. *TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation*, 17(2), 333–339. <https://doi.org/10.12716/1001.17.02.09>
- van de Merwe, K., Mallam, S., Engelhardtsen, Ø., & Nazir, S. (2023b). Towards an approach to define transparency requirements for maritime collision avoidance. *Proceedings of the Human Factors and Ergonomics Society - Annual Meeting*, 67(1), 483–488. <https://doi.org/10.1177/21695067231192862>
- van de Merwe, K., Mallam, S., & Nazir, S. (2024a). Agent transparency, situation awareness, mental workload, and operator performance: A systematic literature review. *Human Factors*, 66(1), 180–208. <https://doi.org/10.1177/00187208221077804>
- van de Merwe, K., Mallam, S., Nazir, S., & Engelhardtsen, Ø. (2024b). Supporting human supervision in autonomous collision avoidance through agent transparency. *Safety Science*, 169, Article 106329. <https://doi.org/10.1016/j.ssci.2023.106329>
- van de Merwe, K., Oprins, E., Eriksson, F., & van der Plaats, A. (2012). The influence of automation support on performance, workload, and situation awareness of air traffic controllers. *The International Journal of Aviation Psychology*, 22(2), 120–143. <https://doi.org/10.1080/10508414.2012.663241>
- van Doorn, E., Horváth, I., & Rusák, Z. (2021). Effects of coherent, integrated, and context-dependent adaptable user interfaces on operators' situation awareness, performance, and workload. *Cognition, Technology and Work*, 23(3), 403–418. <https://doi.org/10.1007/s10111-020-00642-z>
- Vered, M., Howe, P., Miller, T., Sonenberg, L., & Velloso, E. (2020). Demand-driven transparency for monitoring intelligent agents. *IEEE Transactions on Human-Machine Systems*, 50(3), 264–275. <https://doi.org/10.1109/THMS.2020.2988859>
- Weaver, B. W., & DeLucia, P. R. (2020). *A systematic review and meta-analysis of takeover performance during conditionally automated driving*. Human Factors, Advance online publication. <https://doi.org/10.1177/0018720820976476>
- Westin, C., Borst, C., & Hilburn, B. (2015). Strategic conformance: Overcoming acceptance issues of decision aiding automation? *IEEE Transactions on Human-Machine Systems*, 46(1), 41–52. <https://doi.org/10.1109/THMS.2015.2482480>
- Wickens, C. D. (2018). Automation stages and levels, 20 Years after. *Journal of Cognitive Engineering and Decision Making*, 12(1), 35–41. <https://doi.org/10.1177/1555343417727438>
- Wickens, C. D., & Carswell, C. M. (2021). Information processing. In G. Salvendy & W. Karwowski (Eds.), *Handbook of human Factors and ergonomics* (5th ed., p. 1603). John Wiley and Sons.
- Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2013). *Engineering psychology and human performance* (4th ed.). Pearson.
- Wohleber, R. W., Stowers, K., Barnes, M., & Chen, J. Y. C. (2023). Agent transparency in mixed-initiative multi-UxV control: How should intelligent agent collaborators speak their minds? *Computers in Human Behavior*, 148, 107866. <https://doi.org/10.1016/j.chb.2023.107866>
- Wright, J. L., Chen, J. Y. C., & Lakhmani, S. G. (2020). Agent transparency and reliability in human–robot interaction: The influence on user confidence and perceived reliability. *IEEE Transactions on Human-Machine Systems*, 50(3), 254–263. <https://doi.org/10.1109/THMS.2019.2925717>
- Yara International (2022). *Crown Prince and youths christen world's first emission-free container ship*. Yara International. Retrieved May 16, 2022, from <https://www.yara.com/corporate-releases/crown-prince-and-youths-christen-worlds-first-emission-free-container-ship/>
- Koen van de Merwe is a principal researcher at DNV Group R&D at Høvik, Norway. He received

his MSc. in Cognitive Psychology in 2004 and an MSc. in Industrial Ecology in 2006 from Leiden University, The Netherlands and he is currently pursuing his PhD in Nautical Operations at the University of South-Eastern Norway.

Steven Mallam is an Associate Professor of Maritime Human Factors at the Fisheries and Marine Institute, Memorial University of Newfoundland, Canada, and at the Faculty of Technology Natural Sciences and Maritime Sciences at The University of South-Eastern Norway. He received his PhD in Human Factors in 2016 from Chalmers University of Technology, Sweden.

Salman Nazir is a Professor in Training and Assessment at Department of Maritime Operations at the University of South-Eastern Norway. He received his PhD in Industrial Chemistry and Chemical Engineering from Politecnico di Milano, Italy in 2014.

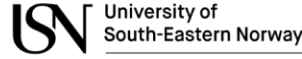
Øystein Engelhardtson is group lead of the Ship Autonomy section at DNV Group R&D at Høvik, Norway. He is trained as a commanding officer in the Royal Norwegian Navy. He received his MSc. in Cybernetics at the Norwegian University of Science and Technology in 2007.



## **Appendix J – Statements of co-authorship**



Norwegian University of  
Science and Technology



UiT / THE ARCTIC UNIVERSITY  
OF NORWAY

## NATIONAL JOINT PHD PROGRAMME IN NAUTICAL OPERATIONS

### List of papers to be included in the PhD thesis and author contribution statements

*This form must be signed by the PhD candidate, the main supervisor (where he/she is a co-author), and at least the other two most central authors. The corresponding author must be among them.*

<b>PhD candidate</b>	Koen van de Merwe
<b>Home institution</b>	USN
<b>Main supervisor</b>	Salman Nazir
<b>Authors</b>	Koen van de Merwe, Steven Mallam & Salman Nazir
<b>Title</b>	Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review
<b>Journal</b>	Human Factors, 66(1), 180–208
<b>URL</b>	<a href="https://doi.org/10.1177/00187208221077804">https://doi.org/10.1177/00187208221077804</a>
<b>Scientific level in the <a href="#">Norwegian Register for scientific journals, series, and publishers</a></b>	1

Author	Contribution <sup>1</sup>
Koen van de Merwe	Conceptualization, methodology, validation, formal analysis, investigation, resources, data curation, writing (original draft), writing (review & editing), visualization, project administration, funding acquisition
Steven Mallam	Supervision, conceptualization, methodology, writing (review & editing)
Salman Nazir	Supervision, conceptualization, methodology, writing (review & editing)

Initials	Name	Signature
K.M.	Koen van de Merwe	van de Merwe, Koen  Digitally signed by van de Merwe, Koen Date: 2024.01.29 11:06:00 +01'00'
S.M.	Steven Mallam	
S.N.	Salman Nazir	

**With my signature I confirm that the contributions are as described above and I consent to this article, where I am a co-author, being part of the PhD thesis of the PhD candidate.**

<sup>1</sup> Allen, L., O'Connell, A., & Kiermer, V. (2019). How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship. *Learned Publishing*, 32(1), 71–74. <https://doi.org/10.1002/leap.1210>





UiT / THE ARCTIC UNIVERSITY OF NORWAY



Norwegian University of Science and Technology



University of South-Eastern Norway



Western Norway University of Applied Sciences



## NATIONAL JOINT PHD PROGRAMME IN NAUTICAL OPERATIONS

### List of papers to be included in the PhD thesis and author contribution statements

This form must be signed by the PhD candidate, the main supervisor (where he/she is a co-author), and at least the other two most central authors. The corresponding author must be among them.

<b>PhD candidate</b>	Koen van de Merwe
<b>Home institution</b>	USN
<b>Main supervisor</b>	Salman Nazir
<b>Authors</b>	Koen van de Merwe, Steven Mallam, Øystein Engelhardtzen & Salman Nazir
<b>Title</b>	Supporting human supervision in autonomous collision avoidance through agent transparency
<b>Journal</b>	Safety Science, 169, 13
<b>URL</b>	<a href="https://doi.org/10.1016/j.ssci.2023.106329">https://doi.org/10.1016/j.ssci.2023.106329</a>
<b>Scientific level in the <a href="#">Norwegian Register for scientific journals, series, and publishers</a></b>	2

Author	Contribution <sup>1</sup>
Koen van de Merwe	Conceptualization, methodology, validation, formal analysis, investigation, resources, data curation, writing (original draft), writing (review & editing), visualization, project administration, funding acquisition
Steven Mallam	Supervision, conceptualization, methodology, writing (review & editing)
Øystein Engelhardtzen	Supervision, conceptualization, methodology, writing (review & editing)
Salman Nazir	Supervision, conceptualization, methodology, writing (review & editing)

Initials	Name	Signature
K.M.	Koen van de Merwe	van de Merwe, Koen  Digitally signed by van de Merwe, Koen Date: 2024.01.29 11:06:43 +01'00'
S.M.	Steven Mallam	
Ø.E.	Øystein Engelhardtzen	Engelhardtzen, Øystein  Digitally signed by Engelhardtzen, Øystein Date: 2024.01.29 14:37:33 +01'00'
S.N.	Salman Nazir	

With my signature I confirm that the contributions are as described above and I consent to this article, where I am a co-author, being part of the PhD thesis of the PhD candidate.

<sup>1</sup> Allen, L., O'Connell, A., & Kiermer, V. (2019). How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship. *Learned Publishing*, 32(1), 71–74. <https://doi.org/10.1002/leap.1210>



UiT / THE ARCTIC UNIVERSITY OF NORWAY



Norwegian University of Science and Technology



University of South-Eastern Norway



Western Norway University of Applied Sciences





## NATIONAL JOINT PHD PROGRAMME IN NAUTICAL OPERATIONS

### List of papers to be included in the PhD thesis and author contribution statements

This form must be signed by the PhD candidate, the main supervisor (where he/she is a co-author), and at least the other two most central authors. The corresponding author must be among them.

<b>PhD candidate</b>	Koen van de Merwe
<b>Home institution</b>	USN
<b>Main supervisor</b>	Salman Nazir
<b>Authors</b>	Koen van de Merwe, Steven Mallam, Øystein Engelhardtzen & Salman Nazir
<b>Title</b>	Towards an approach to define transparency requirements for maritime collision avoidance
<b>Journal</b>	Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 67(1), 483–488
<b>URL</b>	<a href="https://doi.org/10.1177/21695067231192862">https://doi.org/10.1177/21695067231192862</a>
<b>Scientific level in the <a href="#">Norwegian Register for scientific journals, series, and publishers</a></b>	1

Author	Contribution <sup>1</sup>
Koen van de Merwe	Conceptualization, methodology, validation, formal analysis, investigation, resources, data curation, writing (original draft), writing (review & editing), visualization, project administration, funding acquisition
Steven Mallam	Supervision, conceptualization, methodology, writing (review & editing)
Øystein Engelhardtzen	Supervision, conceptualization, methodology, writing (review & editing)
Salman Nazir	Supervision, conceptualization, methodology, writing (review & editing)

Initials	Name	Signature
K.M.	Koen van de Merwe	van de Merwe, Koen  Digitally signed by van de Merwe, Koen Date: 2024.01.29 11:05:11 +01'00'
S.M.	Steven Mallam	
Ø.E.	Øystein Engelhardtzen	Engelhardtzen, Øystein  Digitally signed by Engelhardtzen, Øystein Date: 2024.01.29 14:38:22 +01'00'
S.N.	Salman Nazir	

With my signature I confirm that the contributions are as described above and I consent to this article, where I am a co-author, being part of the PhD thesis of the PhD candidate.

<sup>1</sup> Allen, L., O'Connell, A., & Kiermer, V. (2019). How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship. *Learned Publishing*, 32(1), 71–74. <https://doi.org/10.1002/leap.1210>



UiT / THE ARCTIC UNIVERSITY OF NORWAY



Norwegian University of Science and Technology



University of South-Eastern Norway



Western Norway University of Applied Sciences

## NATIONAL JOINT PHD PROGRAMME IN NAUTICAL OPERATIONS

### List of papers to be included in the PhD thesis and author contribution statements

This form must be signed by the PhD candidate, the main supervisor (where he/she is a co-author), and at least the other two most central authors. The corresponding author must be among them.

<b>PhD candidate</b>	Koen van de Merwe
<b>Home institution</b>	USN
<b>Main supervisor</b>	Salman Nazir
<b>Authors</b>	Koen van de Merwe, Steven Mallam, Øystein Engelhardtzen & Salman Nazir
<b>Title</b>	Operationalising Automation Transparency for Maritime Collision Avoidance
<b>Journal</b>	TransNav, International Journal on Marine Navigation and Safety of Sea Transportation, 17(2)
<b>URL</b>	<a href="https://doi.org/10.12716/1001.17.02.09">https://doi.org/10.12716/1001.17.02.09</a>
<b>Scientific level in the <a href="#">Norwegian Register for scientific journals, series, and publishers</a></b>	1

Author	Contribution <sup>1</sup>
Koen van de Merwe	Conceptualization, methodology, validation, formal analysis, investigation, resources, data curation, writing (original draft), writing (review & editing), visualization, project administration, funding acquisition
Steven Mallam	Supervision, conceptualization, methodology, writing (review & editing)
Øystein Engelhardtzen	Supervision, conceptualization, methodology, writing (review & editing)
Salman Nazir	Supervision, conceptualization, methodology, writing (review & editing)

Initials	Name	Signature
K.M.	Koen van de Merwe	van de Merwe, Koen Digitally signed by van de Merwe, Koen Date: 2024.01.29 11:08:11 +01'00'
S.M.	Steven Mallam	
Ø.E.	Øystein Engelhardtzen	Engelhardtzen, Øystein Digitally signed by Engelhardtzen, Øystein Date: 2024.01.29 14:38:59 +01'00'
S.N.	Salman Nazir	

With my signature I confirm that the contributions are as described above and I consent to this article, where I am a co-author, being part of the PhD thesis of the PhD candidate.

<sup>1</sup> Allen, L., O'Connell, A., & Kiermer, V. (2019). How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship. *Learned Publishing*, 32(1), 71–74. <https://doi.org/10.1002/leap.1210>



UiT / THE ARCTIC UNIVERSITY OF NORWAY



Norwegian University of Science and Technology



University of South-Eastern Norway



Western Norway University of Applied Sciences





## NATIONAL JOINT PHD PROGRAMME IN NAUTICAL OPERATIONS

### List of papers to be included in the PhD thesis and author contribution statements

This form must be signed by the PhD candidate, the main supervisor (where he/she is a co-author), and at least the other two most central authors. The corresponding author must be among them.

<b>PhD candidate</b>	Koen van de Merwe
<b>Home institution</b>	USN
<b>Main supervisor</b>	Salman Nazir
<b>Authors</b>	Koen van de Merwe, Steven Mallam, Salman Nazir & Øystein Engelhardtzen
<b>Title</b>	The Influence of Agent Transparency and Complexity on Situation Awareness, Mental Workload, and Task Performance
<b>Journal</b>	Journal of Cognitive Engineering and Decision Making, 18(2), 156-184
<b>URL</b>	<a href="https://doi.org/10.1177/15553434241240553">https://doi.org/10.1177/15553434241240553</a>
<b>Scientific level in the Norwegian Register for scientific journals, series, and publishers</b>	1

Author	Contribution <sup>1</sup>
Koen van de Merwe	Conceptualization, methodology, validation, formal analysis, investigation, resources, data curation, writing (original draft), writing (review & editing), visualization, project administration, funding acquisition
Steven Mallam	Supervision, conceptualization, methodology, writing (review & editing)
Salman Nazir	Supervision, conceptualization, methodology, writing (review & editing)
Øystein Engelhardtzen	Supervision, conceptualization, methodology, writing (review & editing)

Initials	Name	Signature
K.M.	Koen van de Merwe	van de Merwe, Koen  Digitally signed by van de Merwe, Koen Date: 2024.04.26 09:48:01 +02'00'
S.M.	Steven Mallam	Steven Mallam  2024/05/10
S.N.	Salman Nazir	Salman Nazir  Date:2024.05.06 02:48:01 +02'00'
Ø.E.	Øystein Engelhardtzen	Engelhardtzen, Øystein  Digitally signed by Engelhardtzen, Øystein Date: 2024.05.02 16:03:00 +02'00'

With my signature I confirm that the contributions are as described above and I consent to this article, where I am a co-author, being part of the PhD thesis of the PhD candidate.

<sup>1</sup> Allen, L., O'Connell, A., & Kiermer, V. (2019). How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship. *Learned Publishing*, 32(1), 71–74. <https://doi.org/10.1002/leap.1210>



