Potential of explanations in enhancing trust in autonomous vessels - a systematic literature review.

Rohit Ranjan[1,2] (rohit_ranjan@kgpian.iitkgp.ac.in)

Ketki Kulkarni[1,3] (ketki.kulkarni@hanken.fi)

Mashrura Musharraf[1] (mashrura.musharraf@aalto.fi)

[1 - Aalto University; 2 - Indian Institute of Technology Kharagpur; 3 - HUMLOG Institute, Supply Chain Management and Social Responsibility, Hanken School of Economics]

Introduction

The development of autonomous vessel systems presents a complex socio-technical challenge where AI and humans must coexist and cooperate. A crucial aspect of successful deployment of these systems is ensuring trust in the AI powered autonomy. Our research aims to explore the potential of explanations in enhancing trust in autonomous vessels. While investigation of the notion of explainability and its role in increasing end-user trust is still at the elementary level for intelligent ships, it has already been identified as a key requirement for successful adoption of self-driving cars and highly automated vehicles (HAV) in general. We conduct a systematic literature review to investigate how the impact of explainability on trust has been studied in the domain of autonomous vehicles, what types, timings and modes of explanations contribute to building trust, and provide a framework for experimental methodologies to measure this impact. By bridging theoretical propositions with empirical validation, this study contributes to a deeper understanding of how explainability can effectively enhance trust in the context of autonomous vessels, fostering the development of trustworthy autonomous vessel systems.

Materials and methods

While there has been work proposing different designs of explainability for the goal of trustworthiness of autonomous vessel systems, there is a lack of validation of explanation effects on real-life trust building and also a lack of empirical methods for evaluation of this correlation. Our research is motivated by the need for validated measures that demonstrate the real-life impact of explainability on trust in autonomous systems. In this context, we seek to answer the question of whether explanations can enhance trust in vessel automation by drawing upon the more mature field of autonomous vehicles to define the correlation of explainability and trust.

Through a systematic literature review encompassing recent studies on explainability and trust in autonomous vehicles, we analyze the metrics and experimental settings employed to measure the impact, and the effects of explanation types, timings and mode. PRISMA guidelines were followed for the review and 25 articles were included. The review is done on popular scientific databases Scopus and Web of Science and takes into account all relevant peer-reviewed conference and journal articles published in 2015-2023.

Results

Our review offers a rigorous assessment of the relationship between explainability and trust in autonomous vehicles. It delves into the diverse types of explanations and their temporal aspects, providing valuable insights into how these factors contribute to trust-building. We identify the diverse experimental methodologies used to assess the impact of explainability on trust, encompassing surveys, questionnaires, and behavioral data collection. The review highlights

various correlated human factors that can act as surrogates for trust, and factors like autonomy levels, age, and demographics, which influence the impact of explanations on trust formation. We explore theoretical propositions regarding the constituents of effective explanations, shedding light on the mechanisms through which explanations enhance trust.

Implications on sustainable maritime operation

By identifying the types and aspects of explanations that contribute to building trust, this study informs the design and development of trustworthy autonomous vessels. The findings can guide stakeholders in making informed decisions regarding the integration of explanations, enhancing the safety, reliability, and efficiency of autonomous vessels. Additionally, the research outcomes can influence policy-making and regulatory frameworks, ensuring transparency and trustworthiness in autonomous vessel operations. Ultimately, this research contributes to the advancement of sustainable maritime operations by fostering trust, acceptance, and responsible deployment of autonomous vessel technologies.