IWASS 2021

Proceedings

International Workshop on Autonomous Systems Safety

April 20, 21 and 28, 2021
Online Event

UCLA ENGINEERING
B. John Garrick Institute for the Risk Sciences

NTNU – Trondheim
Norwegian University of Science and Technology

IWASS 2021

Proceedings of the International Workshop on Autonomous Systems Safety 2021

Edited by:

Christoph A. Thieme, Marilia A. Ramos, Ingrid B. Utne, Ali Mosleh

DOI: 10.34948/N33019

October 2021

NTNU
Norwegian University of Science and Technology

The B. John Garrick Institute for the Risk Sciences
UCLA ENGINEERING

Published by: The B. John Garrick Institute for the Risk Sciences, UCLA

www.risksciences.ucla.edu

Preface

The International Workshop for Autonomous System Safety (IWASS) is a joint effort by the B. John Garrick Institute for the Risk Sciences at the University of California Los Angeles (UCLA) and the Norwegian University of Science and Technology (NTNU).

IWASS is a platform for cross-industrial and interdisciplinary exchange of knowledge on autonomous systems' Safety, Reliability, and Security (SRS), by invited experts and observers. The workshop gathers experts from academia, industry, and regulatory agencies to discuss SRS challenges and their potential solutions for SRS of autonomous systems from different perspectives. It complements existing events that are organized around specific types of autonomous systems (e.g., cars, ships, aviation) or particular safety or security-related aspects of such systems (e.g., cyber risk, software reliability, etc.). IWASS distinguishes itself from these events – and complements them – by addressing these topics together in an attempt to focus on proposing solutions for SRS challenges common to different types of autonomous systems.

IWASS 2021 was held online on April 20th, 21st, and 28th 2021. The invitation-only events gathered 49 participants from 39 organizations from around the globe. This report summarizes the presentations and the results of the workshop's discussions. The results outline current challenges and future research directions that need to be addressed to make autonomous systems safe, reliable, and secure in the future.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

Introduction

IWASS 2021 is the second edition of the workshop series on Autonomous System Safety, Reliability, and Security, initiated in 2019. The first IWASS was held in March of 2019 in Trondheim, Norway. The event counted nearly 50 participants from diverse industries, with different backgrounds and from eight countries. The proceedings published by NTNU[1] summarizes the discussions held at the workshop, briefly described in the next section, in addition to six research papers on topics related to autonomous systems safety, reliability, and security.

Initially planned as an in-person event in 2020 at UCLA, IWASS 2021 switched to an online event in 2021 due to the COVID-19 pandemic and related travel restrictions. IWASS 2021 assembled 49 participants with diverse expertise from 39 different organizations and nine countries. The workshop program was distributed over three days. On the first day, five domain experts presented challenges concerning autonomous systems SRS from different perspectives. A keynote presentation opened the second day, followed by four parallel sessions discussing specific topics of autonomous systems. Finally, the findings of each session and the remaining challenges were summarized and discussed on the third day.

The four topics discussed in dedicated breakout sessions included: (1) *Demonstrating Safety of Autonomous Systems; (2) Verification & Validation*, *Risk Acceptance*; *Human on the Loop; (3) The Role of Humans in Autonomous Systems Operations*; *Modeling and Simulation for Understanding Complexity and Cascading Failures*; and (4) *Artificial Intelligence and Data Analytics in Resilient Autonomous Systems*. The discussions of the four sessions are summarized in these proceedings. While solving the issues concerning these topics during a single workshop is not realistic, the findings constitute a path towards the safe development and operation of autonomous systems for researchers, developers, and regulatory agencies.

Demonstrating Safety of Autonomous Systems is not a trivial task, as discussed by the first breakout session participants. Whether the risk associated with these systems is acceptable can be addressed only with the evidence that their performance assessments have been *verified* and *validated*. This group explored the topic from three different perspectives. First, the machine-centric verification and validation, e.g., how the operational codes that govern functionality are developed. While technical specialists are by nature optimistic, carrying out machine-centric verification and validation may be more challenging

[1] Available at: https://bit.ly/2SsPrLd

than one might expect. Second, the human-machine interface verification and validation, namely how "humans-in-the-loop" are represented. A general consensus emerged that the current methods for verification and validation in this realm are inadequate. Social requirements verification and validation present an additional challenge, for instance, how the "rules of the road" are established, what parameters are measured and what requirements are mandated. It is unescapable that autonomous systems will be deployed in an open environment whose actors and conditions may be shifting, sometimes slowly, sometimes rapidly. How can we be sure that societal aspects are adequately reflected in how our models are verified and validated?

Concerning *Human on the Loop*, a consensual perception of the second discussion session participants was that the extent of human involvement in autonomous systems operations and its impact on safety is still not well established, in contrast with the development of software and hardware. For instance, levels of Autonomy (LoAs) tend to oversimply the role of humans in higher LoAs. Yet, while the task load may be reduced in higher LoAs, the tasks may demand significantly higher levels of interaction and effort and be critical for the system safety. The approach to LoAs must thus be revisited for clarifying the human role. An additional discussion point concerns the extent to which the method for analyzing human-system interaction (human reliability analysis, human factors engineering, etc) can be applied to different autonomous systems, given the differences and similarities between them.

Autonomous systems features such as *complexity and possible cascading failures* pose several challenges concerning methods for risk assessment. The third session participants discussed the need for a "framework" with various methods for identifying, analyzing and evaluating different hazards and hazardous events. Such a framework requires qualitative and quantitative methods and approaches and should promote the combination and application of both simulation and more traditional "discrete logic" risk assessment methods. Furthermore, the complexity of autonomous systems could be addressed in risk assessment through the compartmentalization of the systems. A challenge lies in defining the subsystems' boundaries and the correct integration of the sub-models with each other. When choosing the assessment method, it is important to consider the objective and context, i.e., validation or verification. More effort is required to identify suitable methods.

Autonomous systems rely heavily on *Artificial intelligence (AI) and data analytics*, as discussed by the fourth discussion session participants. AI and data analytics can be applied to autonomous systems in two ways: firstly, it can be applied as part of the systems' intelligence, i.e., information processing, decision-

making, or motion control. Secondly, it can be used as part of the verification and safety assurance process of autonomous systems. Whatever the purpose of the application of AI methods is, AI should be used carefully, combining domain knowledge with reliable data. The AI method needs to suit its purpose in an autonomous system and should be combined with other suitable methods for control of the vehicle. When using AI methods an interdisciplinary approach is required.

The issues discussed at IWASS extended beyond the above-mentioned topics, including the need for a multidisciplinary view, international cooperation, and assessment of the impact of possible regional differences. These discussions are summarized below, followed by the summaries of the presentations held at IWASS 2021 and an overview of IWASS 2019.

Further Considerations on Autonomous Systems SRS

The concluding session of IWASS 2021 was dedicated to an open discussion on points concerning autonomous systems SRS that were not directly addressed by the breakout session topics. These considerations are summarized below.

Societal risk acceptance

A feeling of control influences risk acceptance: There is a tendency of accepting a higher risk level when people feel in control of the system or perceive the system to be under control of another. Risk acceptance is also impacted by people's assumptions about the system behavior when confronted with a challenging choice. Some discussion revolved around the "Trolley Problem", where there is only a binary decision between harming one group to save another (e.g., a car passenger or a pedestrian standing in the way). Yet, this type of scenario is not expected to be encountered with a high frequency, and more attention should be directed to daily decisions. Hence, acceptance of a system goes beyond the question of "who to kill" and includes responsibility, reparation, fairness, explainability, social impact, and more. Responsibility is an essential factor when discussing possible accidents and liability. The issue of who should be found responsible in case of an accident – the operator, the manufacturer, the software developers, etc. – is still an open question that can also impact societal acceptance.

Risk acceptance is different for different systems. Accidents involving some systems seem to be more tolerated than other systems. For instance, accidents involving self-driving cars seem to generate minor public outcry compared to airplane accidents

The need for a multidisciplinary view.

Given the impact autonomous systems will have in several aspects of our lives, the development process should include a multi-disciplinary approach to cover ethical and societal aspects. Therefore, discussing interdisciplinary issues and among different fields of application has been highlighted as beneficial to safely advancing autonomous systems development.

Accident behavior

While most research focuses on preventing accidents, accident behavior and follow-up after an accident also need to be investigated. It is still unclear how to ensure that an autonomous system realizes that it has been involved in an accident and then takes the correct actions. For example, an autonomous car that has been involved in a crash should not start moving again as soon as the crash site is being cleaned. Similarly, accident data needs to be collected and analyzed to assess if it is a "behavioral" problem or just a coincidence. In the USA, this is already being done by the National Highway Transportation Safety Authority. In case statistics indicate a trend, recalls may need to be issued.

International cooperation and regional differences

Accident data from different regions need to be combined to assess the safety performance of autonomous systems being operated in these different regions and countries. This task is challenging since little international cooperation on the regulatory levels exists. Furthermore, while accidents with autonomous vehicles in the USA have attracted significant attention in the media, some similar accidents in other countries have received little attention so far.

Assurance and related methods

In general, methods should be scalable and able to treat system elements separably. A simulation approach is likely to be required, but the type of simulation depends on the autonomous system, its development stage, and the types of assurance needed. Currently, the assurance process for many autonomous systems seems to be testing, while (safety) modeling does not seem to be as widely adopted.

Risk communication and regulatory agencies.

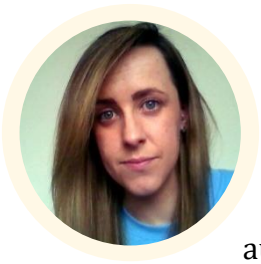
Despite the considerable benefit, conversations between safety experts and regulators are still incipient. The safety community may be perceived as pessimistic; hence safety-related messages need to be formulated in a helpful and enabling way to create awareness.

Summary of the Presentations held at IWASS 2021[2]

Human Factors In The Design Of Autonomous Systems: What Can We Learn From The Boeing 737 Max Accidents?

Claire Blackett, Institute for Energy Technology (IFE), Norway

In 2018 and 2019, two Boeing 737 Max airliners crashed, killing 346 people. Investigations of both crashes revealed failures of sensors related to a newly installed system, the Maneuvering Characteristics Augmentation System (MCAS), which was meant to automatically regulate the pitch of the aircraft nose to avoid stalling, as the root cause of the accidents. Significant organizational, safety culture, and regulatory failures contributed to these tragic events. In my presentation, I will explore what we can learn from the Boeing 737 Max accidents when designing systems of the future, and whether our standard Human Factors methods and best practices are up to the task.



Heterogeneous Verification Of Autonomous Robotic Systems

Marie Farrell, Maynooth University, Ireland

An analysis of the literature has revealed that, as autonomous robotic systems increase in complexity, it will become necessary to employ distinct verification techniques for individual system components in order to ensure the correctness of the entire system. This talk will summarise the approaches that have been used to formally specify and verify autonomous robotic systems as well as the challenges that emerge during this process. This talk will illustrate, via an example of an autonomous rover, how distinct techniques can be used to verify different system components. However, there is currently no holistic framework within which the results from the application of these various techniques can be combined in a meaningful way. This talk will discuss potential ways to link these results and discuss the notion of confidence in overall system verification.

[2] The presentations held at IWASS are available and can be downloaded at: https://www.risksciences.ucla.edu/iwass-presentations

Autonomous Driving Challenges: Toward Scenario-Based Causal Models

Stephen Thomas, Motional, United States

Autonomous Vehicles offer some unique challenges that stretch the limits of traditional safety engineering practices. Most current safety standards and methodologies in the AV industry were not originally intended for application to autonomous vehicles. In this presentation, we discuss the challenges and limitations of current standards and methodologies. We provide a brief overview of a proposed advanced safety analysis framework which addresses these challenges by combining an operational scenario-based approach with advanced causal analysis using Bayesian Networks.

NHTSA Human Factors Research Update

Stacy Balk, National Highway Traffic Safety Administration (NHTSA), United States

The role of human factors research is to provide an understanding of how drivers perform as a system component in the safe operation of vehicles. This role recognizes that driver performance is influenced by many environmental, psychological, and vehicle design factors. The focus of the research is to determine which aspects of vehicle design should be modified to improve driver performance and reduce unsafe behaviors. An additional focus is to evaluate driver's capabilities to benefit from existing or new in-vehicle technologies. An update of ongoing NHTSA's Human Factors Vehicle Safety Research will be provided.

Why Verify Ethical Behaviour?

Marija Slavkovik, University of Bergen, Norway

Machines and software that share the environment with people need to not only accomplish their tasks, but also do so by not violating the norms of behaviour in that environment. Machine ethics studies how to automate moral and common sense reasoning. However, automating behaviour is not sufficient, one also needs to ensure the stakeholders that the intended behaviour is indeed exhibited. Furthermore, one needs to guarantee that unforeseen events do not happen within prescribed use. Can verification help?

Automated Driving Systems Safety

Tim Johnson, National Highway Traffic Safety Administration (NHTSA), United States

The discussion focused on providing an overview of activities underway at the National Highway Traffic Safety Administration in the area of Automated Driving Systems safety.

Summary of IWASS 2019

IWASS 2019 was a three-day workshop and consisted of presentations by subjet matter experts, breakout sessions dedicated to specific topics, and open discussions. Four topics were discussed in the breakout sessions: Autonomous transportation technology: Society and Individuals in the Loop, Safety, Reliability and Security Modelling and Methods for Autonomous Systems, System Verification, Processes and Testing, and Autonomous Systems Intelligence and Decision Support.

The key findings of the workshop discussions are summarized below. For the complete set of take-away messages and breakout session reports, please refer to the IWASS 2019 Proceedings[3].

For developers of autonomous systems and functionalities strategies and methods are needed for safer, reliable, and secure performance to ensure compliance with requirements. For operators, operations need to be executed supported by effective risk control to achieve safe and robust operations. Regulators and authorities need to create new standards and guidelines to perform regulatory and oversight activities efficiently. Systems need to be designed with safety as a key driver to gain customers and the public's confidence. This will require risk identification, risk assessment and modeling, testing and verification and validation of system design and operation.

Trust is a driver for the acceptance of new technologies. Trust is required for the capabilities of the technology and the organizations that operate and regulate the technologies. This requires open and inclusive development processes, including and respecting societal values.

Autonomous systems add to the complexity of a system, which also increases the complexity concerning SRS assessment and assurance. Currently, many of the existing methods are inadequate and lack integrated modeling of hardware, human, and software. Additionally, self-learning systems and data quality for training adds challenges to the SRS assessment. Therefore, new modeling techniques are required that capture interdependencies and connections. Simulations may assist in this assessment and provide input to decision-making during design and operation. Cyber security and software risk

[3] : *Proceedings to the First International Workshop on Autonomous Systems Safety*. Ramos, M.; Thieme, C.; Utne, I.; Mosleh, A. (Editors). 2018. ISBN: 978-82-691120-2-3. Available at https://www.ntnu.edu/documents/139785/1283738018/Proceedings+of+the+1st+IWASS.pdf/dadf6629-ef88-4e48-9c2a-8576c0379da8

are different from traditional security issues and hardware failures. In these cases, past behavior cannot be used to predict future behavior.

Verification and testing will play a vital role in the development of autonomous systems. Regulatory, ethical and societal requirements need to be addressed. However, it is of concern how to derive these requirements. Especially, a self-learning (artificial intelligence based) system needs continuous and integrated verification processes. The results of any verification process should be communicated openly to the public and regulators to build trust.

Finally, the First IWASS proceedings also contains six articles that address some of the above-mentioned challenges. More specifically, Ramos and Thieme present the Human-System interaction method in autonomy (H-SIA) method to integrate software, hardware, and human failures in risk models. Luckuck discusses using several formal methods to engineer safe, trustworthy, and correct autonomous systems. Myklebust et al. discuss safety cases and operational software development in light of existing standards and practices. Ventikos and Louzis address the complex nature of risk assessments of autonomous ships from a systemic perspective. How humans, drivers, and pedestrians interact with autonomous cars is explored by Jafary et al. Basnet et al. turn again to the marine environment by investigating and identifying suitable risk assessment approaches for autonomous marine vessels.

References

Jafary, B. Fiondella, L., Mosleh, A. "A Survey on Autonomous Vehicles Interactions with Human". In: *Proceedings to the First International Workshop on Autonomous Systems Safety*. Ramos, M.; Thieme, C.; Utne, I.; Mosleh, A. (Editors). 2018. ISBN: 978-82-691120-2-3. Pg. 104-111

Luckcuk, M. "Why use Formal Methods for Autonomous Systems?". In: *Proceedings to the First International Workshop on Autonomous Systems Safety*. Ramos, M.; Thieme, C.; Utne, I.; Mosleh, A. (Editors). 2018. ISBN: 978-82-691120-2-3. Pg. 87-94

Ramos, M.; Thieme, C. "Human-System Interaction in Autonomy Method – a Structured Approach to Risk Monitoring". In: *Proceedings to the First International Workshop on Autonomous Systems Safety*. Ramos, M.; Thieme, C.; Utne, I.; Mosleh, A. (Editors). 2018. ISBN: 978-82-691120-2-3. Pg 78-86

Sunil Basnet, S. Valdez Banda, O.; Hirdaris, S. "The Management of Risk in Autonomous Marine Ecosystems - Preliminary Ideas". In: *Proceedings to the First International Workshop on Autonomous Systems Safety*. Ramos, M.; Thieme, C.; Utne, I.; Mosleh, A. (Editors). 2018. ISBN: 978-82-691120-2-3. Pg 112-121

Thor Myklebust, T.; Stålhane, T.; Hanssen, G. K. "Safety Case and DevOps Approach for Autonomous Cars and Ships" In: *Proceedings to the First International*

Workshop on Autonomous Systems Safety. Ramos, M.; Thieme, C.; Utne, I.; Mosleh, A. (Editors). 2018. ISBN: 978-82-691120-2-3. Pg 95-96

Ventikos, N.; Louzis, K. "The Future of Risk in the Context of Autonomous Ship Operation". In: *Proceedings to the First International Workshop on Autonomous Systems Safety*. Ramos, M.; Thieme, C.; Utne, I.; Mosleh, A. (Editors). 2018. ISBN: 978-82-691120-2-3. Pg 97-103

IWASS 2021
Discussion Sessions Reports

The following four reports summarize the discussions held during IWASS 2021 breakout sessions. The reports were written by the session chairs with contributions from the group participants. The discussions were partly guided by points concerning autonomous systems SRS raised by the organizing committee and the participants during the registration process. These points, i.e., "Main challenges with respect to autonomous systems SRS", are presented at the end of the report. Also listed are the names of the group participants.

Demonstrating safety of autonomous systems – verification & validation, risk acceptance

Report of the Discussions of Breakout Session

Authors: Metlay, D. ;Mosleh, A.
Contributors: Carson, S.; Currie-Gregg, N.; Farrell, M.; Feather, M.; Gaither, M.; Glomsrud, J.; Harrison, C.; Hjørungnes, S.; Jones, A.; Lundteigen, M.; Porathe, T.; Rothmund, S.; Torben, T.; Valdez Banda, O.

As autonomous vehicles begin to be widely deployed, questions surrounding the verification and validation of their performance increasingly come to the fore. In particular, concerns about how much risk these innovations pose and whether that risk is acceptable can only be addressed if persuasive evidence has been mustered that their performance assessments have been verified and validated. This group explored topics surrounding verification/validation from three different perspectives.

Machine-centric verification and validation: How operational codes that govern functionality are developed

This type of verification and validation is the most familiar to technical specialists. However, what is especially problematic is the concern: Who decides what needs to be verified or validated? On the one hand, developers are motivated by self-interest; the cost of failure can be extraordinarily high, both in terms of human casualties or ecological damage and missed opportunities and "wasted" resources. For example, performance malfunctions in an autonomous car can lead to injuries and death. Breakdowns in an autonomous helicopter on the surface of Mars can endanger a costly and long-planned mission.

On the other hand, public demands for governmental intervention are likely to increase over time. The ubiquitous presence of social media means that challenges for autonomous vehicles will sooner or later be uncovered and publicized. The greater the difficulties, the more likely they will stimulate calls to rein in the errant technology. Already a small number of states have adopted rules that restrain the deployment of autonomous vehicles. It does not take much imagination to anticipate that others, as well as the federal government, will also become involved.

As an autonomous system engages in increasingly complex and differentiated environments, innovative and creative methods for verification and validation may need to be constructed. In particular, it is unclear whether it will be possible to continue to rely exclusively on incremental learning. This traditional approach for adjusting verification and validation methods appears to be burdened in at least three respects.

First, "wheat must be separated from chaff". Artificial intelligence techniques require a substantial amount of input data. Programs must be devised to interrogate that data so that reliable inferences can be drawn. Second, the situation is further complicated by the potentially long time-constant of feedback between the revisions and any determination whether the changes have had their expected effect. Third, any changes to methods for validation and verification must be implemented by organizations populated with human beings. Conflicts among those in charge can arise for the following reasons. (1) Interpretations of the artificial intelligence may differ. Those processes can open the door to disagreements about the conclusion's soundness. (2) Debates may be sparked over the costs and benefits of making any change. It is not unheard of for organizations to cover up errors rather than rectifying them. (3) Getting agreement among the wide spectrum of stakeholders—including implementers, manufactures, and possibly regulators—may be hard to secure. The modification of the 737MAX software illustrates well the obstacles standing in the face of attempting to improve the verification and validation methods.

In sum, while technical specialists are by nature optimistic, carrying out machine-centric verification and validation may be more problematic than one might expect.

Human-machine interface verification and validation: How "humans-in-the-loop" are represented

Every participant in the breakout session recognized that significant challenges would have to be surmounted in this area. However, overall, most of them skewed positively. They judged that with sufficient time and resources, the difficulties could be overcome. Nonetheless, a general consensus emerged that the current methods for verification and validation in this realm are inadequate. Among the issues that must be resolved are the following:

- Additional work needs to be done to quantify system requirements. Typically very little examination has been given in advance to what level of performance should be demanded from the autonomous system. This gap

may have important implications especially as autonomous system development moves away from scripting what needs to be done to simply specifying the goal the technology has to achieve.

- For any new technology, you can never conceive of the problems that lie ahead, but once the technology is developed, it may be too late. Participants reacted to this dilemma differently. Some remarked: "We don't have to solve all problems and issues in order to bring systems to society. Systems are introduced even though we know that there are risks." This point of view may implicitly assume (1) that individuals' (stakeholders') risk profiles are similar; (2) that risk evaluations are persuasive; and (3) that incremental learning is timely and effective. [See the discussion above for a more detailed discussion of this final point.]
- Substantial social psychological research suggests that judgments under uncertainty may be suspect. Studies exploring heuristics and cognitive biases demonstrate convincingly that people simplify complex decisions. These short-cuts may be appropriate under certain circumstances and not others. Decision-aiding methodologies may be useful in mitigating the effects of simplification. But it is not at all evident that they can fully compensate.

Social requirements verification and validation: How the "rules of the road" are established; what parameters are measured and what requirements are mandated

When we use the term "system," it implies the existence of boundaries. These may be either implicit (often times) or explicit (less frequently). It is, however, unescapable that these systems will be deployed in an open environment whose actors and conditions may be shifting, sometimes slowly or sometimes rapidly. How can we be sure that societal aspects are adequately reflected in how our models are verified and validated?

Participants noted that models would never be complete. The central question is then how the choice is made about which factors can be "safely" eliminated? This question is relevant for both developers and, at least in some cases, for the regulators as well.

Two considerations seem to be pertinent in seeking answers. First, those decision-makers may respond by asserting: "We have reduced the risk to "an acceptable level." Bundled into the claim, however, seems to be the presumption that their risk calculations are reliable. Yet, given the challenges involved even in verification/validation perspectives considered above, how confident should one be in the performance assessments? To be sure, some techniques could be used

to calculate risk distributions and to "quantify" uncertainty. But whether the result is **acceptable** depends critically on whether the decision-makers are risk-tolerant or risk-averse. (It is particularly instructive to study when the decision-makers acceptability levels differ.)

Second, the behavior of those decision-makers may or may not be trustworthy. Developers in particular have built-in preferences that may be at work as they define the system's scope. To begin with, they all too frequently possess an unrealistically high regard for their wisdom and insightfulness. As one participant aptly put it: "There is some arrogance. They know what they are doing and how to do it." This overconfidence is often directed at the uninitiated, whose views can then be discounted or completely dismissed. (Only occasionally will the regulators' technical incapacity clearly stand out as it did the case of the 737MAX software.)

Further, developers and sometimes regulators tend to emphasize the benefits of an autonomous system and neglect to explore the negative downstream impacts fully. For developers, this pattern of behavior is to be expected; autonomous system owners are reluctant to seriously question the utility of their efforts. For regulators, the phenomenon of "capture" may erode their credibility. Sustained contacts with developers almost naturally lead the officials to view the world through developers' eyes and sympathize with the developers' perspectives.

Despite the challenges to verification and validation at the societal level, there is noteworthy too little funding available to address those concerns.

An intellectually deep agenda lies ahead

The breakout group participants were cautiously optimistic about the prospects for wide-scale deployment of vehicles at an autonomous Level 4 and perhaps at a Level 5 (see Table 1 at the following report, "Human on the Loop"). Notwithstanding this point of view, no one underestimated the difficulty in realizing those goals in the short term.

This optimism seems to be derived ultimately in a "faith" that incremental deployment choices guided by methods and models such as risk-benefit calculations are a prudent way to proceed. Although disciplinary predispositions may color that view, it is at least consistent with how technologies have been implemented in the past.

Main challenges of autonomous systems SRS: Safety demonstration, verification & validation, risk acceptance

- What are the similar challenges between the different autonomous systems, operations and industries with respect to verification and validation?
- How to achieve trust in autonomous systems? How is this related to verification and validation processes? And to social and cultural aspects in a society? What factors will influence public trust of autonomous systems?
- How can verification and the corresponding test scope for autonomous systems be managed dynamically?
- Should risk acceptance be higher, similar or lower for autonomous systems? Do autonomous systems and operations need to be "as safe as" or safer than other types of systems? Should "acceptable risk" change with the level of autonomy (LoA)?
- What is as safe as a conventional system?
- How is verification and validation linked to the safe performance of autonomous systems?
- How can machine learning and AI be used for verification of autonomous systems?
- If a digital twin is to be used for verification, how can we be sure that the twin actually behaves similarly to the real system so that risks are detected?
- How can AI methods be used to demonstrate safety and compliance?
- Where are the regulatory gaps for adoption of autonomous systems?
- What must risk models for software intensive and AI-based systems convey to convince users, regulators and standardization bodies?
- How are the creation and economic incentives behind algorithms affecting safety?
- How to handle the state space problem when validating an autonomous system?
- What data is needed to assess the performance of autonomous systems?
- How to map and balance the needs and requirements from a behavioral and operative perspective?
- Do our current methods fully and adequately explore the changing role of the human in an increasingly automated and autonomous world

- How can we be sure that the models are complete and even discuss it, if the organizational issues are not taken into account or completely outside of the box?

Group Participants

Ali Mosleh – *Session Chair*
University of California, Los Angeles, USA

Nancy Currie-Gregg
Texas A&M University, USA

Martin Feather
Jet Propulsion Laboratory, California Institute of Technology, USA

Jon Arne Glomsrud
DNV, Norway

Chris Harrison
Rail Safety and Standards Board, United Kingdom

Alun Jones
Norwegian University of Science and Technology , Norway

Thomas Porathe
Norwegian University of Science and Technology, Norway

Tobias Torben
Norwegian University of Science and Technology, Norway

Daniel Metlay – *Session Chair*
University of California, Los Angeles, USA

Marie Farrell
Maynooth University, Ireland

Michael Gaither
Texas A&M University System, USA

Siri Granum Carson
Norwegian University of Science and Technology , Norway

Siv R. Hjørungnes
Kongsberg Maritime, Norway

Mary Ann Lundteigen
Norwegian University of Science and Technology, Norway

Sverre Rothmund
Norwegian University of Science and Technology, Norway

Osiris Valdez Banda
Aalto University, Finland

Human on the loop – role of humans in the autonomous system

Report of the Discussions of Breakout Session

Authors: Ramos, M.; Ma, J.
Contributors: Balk, S.; Bindingsbø, J.; Blackett, C.; Bremens, J.; Gaither, M.; Hoem, Å.; Johansen, T.; Johnsen, S. O.; Massaiu, S.; Price, J.; Rødseth, Ø.; Slavkovic, M.; St. Clair, A.; Ventikos, N.; Wróbel, K.

Several aspects of software and hardware for autonomous systems are well established or developing at a fast pace. Yet, the research on how humans will be included in these systems is still developing. Human-system interaction can be analyzed through different perspectives and has many aspects (Figure 1). For instance, Human Factors Engineering or Ergonomics mainly analyzes the design of the system and its impact on human performance, Human Reliability Analysis models human-system interaction with a focus on human error and its influencing factors, often quantitively. Systems may be analyzed during different life stages, such as the design phase, the operation stage, or often after an incident. Further, while often applied to system operators, methods for assessing and/or modeling human-system interaction can also target system users, such as passengers of an autonomous bus, or people external to the system, such as pedestrians interacting with an autonomous vehicle. The discussion summarized in this report focuses on system operators.

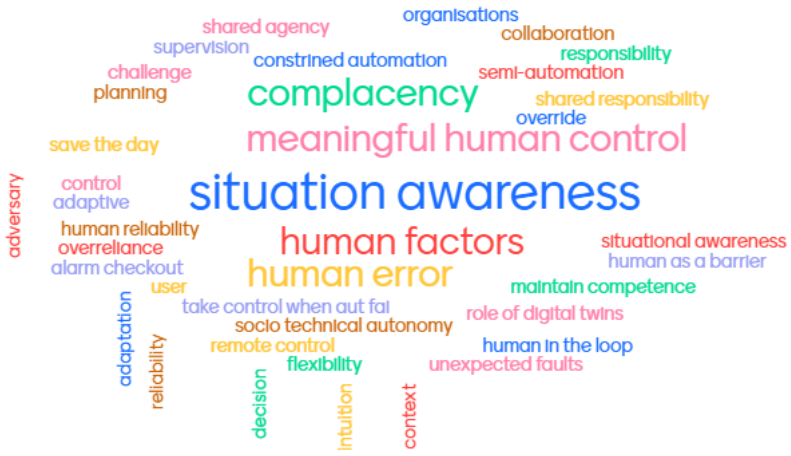


Figure 1: Breakout session participants' keywords for human role in autonomous systems

By autonomy we mean a system that is non-deterministic in that it has a freedom to make choices, and by automated we mean a system that is more deterministic in that it will do exactly what it is programmed to do. Automation can increase from a system that is human-controlled to a system that is fully autonomous, without human intervention. The Level of Automation (LoA) has been used to describe the degree of automation from the lowest level (i.e. total human control) to full autonomy. Different number of levels has been used to describe the responsibility between the humans and the automated system, as mentioned by Kaber (2018).

Human operators may adopt different roles in autonomous systems, such as monitoring and supervision, remote control, or onboard control. While certain expectations are built around how humans should act in these roles (e.g., timely control for troubleshooting), many systems are not designed to ensure that these expectations are met. For instance, if the system is designed such that a human operator should take over control in case it runs into an unexpected, it should take into account the time needed for humans to not only detect an alert but also assess the situation, make a decision, and take action. Recent incidents with self-driving cars illustrate how the reliance on a safety driver to take over control when needed does not always match the conditions given to the driver: in a collision between a developmental automated driving system and pedestrian in 2018, the system provided an auditory alert only 0.2 seconds before the collision (National Transportation Safety Board 2019). The human takeover time varies from 2 to 26 sec., (Eriksson A & Stanton N A, 2017) challenging the design of autonomous systems to enable rapid (in-time) human intervention.

This problem is also interesting for uncrewed ships where the ship's high value makes the concept of a remote control center (RCC) very relevant as the additional relative cost of the RCC is smaller than for, e.g., cars. This creates the possibility for developing new forms of cooperation between ship automation and remote ship operators. However, this area is still being researched, and conclusive results on how this cooperation should be organized and described are still in the draft stage (Rødseth, 2021; Ramos et al. 2019).

The extent to which humans will (or should) be part of autonomous systems' operations depends on many factors that surpass technical constraints. For instance, regulations may demand that human operators should be onboard of the physical system, manufacturers may rely on a human controller for liability reasons, or society may feel more comfortable interacting with the system knowing that it is being monitored by a human. Humans' tasks on autonomous systems operations are also dependent on a realistic assessment of humans' abilities and limitations: How fast can humans react? What are humans'

limitations in supervision and control? In which tasks do humans excel? How can we ensure that humans will have meaningful control? There are no fast or easy responses to these questions. Not only may they vary with individuals, but they also depend on external factors. For instance, the probability of a fast reaction increases with continuous monitoring. How is effective monitoring affected by one's trust in the system performance? The example of self-driving cars points to a low vigilance that may result from, among other factors, an "overselling" of these cars' autonomous capabilities by the manufacturers or the news, leading to an overtrust in the system (automation complacency).

The expectations on human tasks and possible influencing factors differ among autonomous systems. Appropriate training, for instance, is considered a crucial factor for ensuring safety[4]. Humans must be trained not only for regular operation, but also for recognizing hazardous situations and how to act to prevent or mitigate incidents. Some systems are operated by a highly trained crew, such as dynamic positioning vessels. One can expect operators to receive special training for monitoring or remote controlling the system once it operates with a higher level of autonomy (LoA)[5]. Autonomous cars, in contrast, can be driven by drivers that obtained their license with no special training for interacting with these vehicles' technologies.

Given the differences between autonomous systems and to which extent they will include humans in their operations, can "human on the loop" be discussed using a common framework for all systems? As a first step, the meaning of "human on the loop" must be defined. Secondly, the LoA under which the system will operate – and its impact on human operators - must be understood. The following sub-sections briefly address these points, followed by proposals of research directions on the human role in autonomous systems' safety.

What do we mean by "Human on the Loop"?

While used interchangeably by some authors, it is important to differentiate between the terms "human in the loop", "human on the loop" and "human in control". According to the Ethics Guidelines for Trustworthy AI by the European Commission (European Commission 2019), Human in the Loop refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. Human on the Loop refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. Human in Control refers to the capability to oversee the overall activity of the AI system (including its broader economic,

[4] It should be noted that for many systems, in particular those operating in a dynamic environment, training cannot cover ALL possible situations that may arise.

[5] The efficiency of this training is to be assessed.

societal, legal, and ethical impact) and the ability to decide when and how to use the system in any particular situation. It should be noted that these definitions are not entirely in line with the "out of the loop (OOTL) performance problems" as defined in the human factors literature before the introduction of highly automated or autonomous systems, and their adoption may need clarification to avoid confusion.

Levels of Autonomy (LoA): What are the Implications in terms of Humans Roles?

Autonomous systems are often classified according to their LoA. The LoA indicates the tasks that are responsibilities of the system, such as providing assistance to the operator or implementing actions without the need for operator approval. Many LoA tables have been proposed, either for specific systems or for more generic applications. While some tables also specify the role of humans for each of the LoA (e.g., Table 1), these are often simplified and are not sufficient to evaluate the human-system interactions. Furthermore, they may lead to a misunderstanding of the associated challenges. For instance, higher LoAs are generally defined by the technical system being responsible for the majority of the tasks - until reaching full autonomy, in which the system would operate without human supervision or intervention.

At first glance, given that humans are responsible for fewer tasks, the human-system interaction may seem less demanding, and human tasks may seem less critical to the system performance in higher LoAs. Nevertheless, a lower task load does not imply a lower task complexity. The situation assessment and action following a long period of a monitoring role, the understanding of the state of a complex system, and the assessment of a dynamic environment and surrounding systems may demand a significant level of interaction and effort. Furthermore, these tasks may be critical to system safety, in particular in case of system failure.

Another way to illustrate LoA is related to the need for an operator to be available at all times, although possibly only seldom engaged, as Table 1 implies for all but level 5. This is illustrated in Figure 2, where C0 to C3 indicates the human level of watchkeeping: 3 – continuous, 2 – away for shorter periods (some minutes), 1 – away for longer periods (tens of minutes), and 0 – not attending at all. A0 to A3 indicates the automation system's ability to control the process: 0 – always need human assistance, 1- can operate without human support, but is not able to calculate for how long, 2- can operate without human support for longer periods and can calculate the duration of the period, and 3- can operate completely without human assistance (Rødseth et al. 2021).

Table 1: Pilot's and aircraft's main tasks in different levels of autonomy (Johnsen and Evjemo, 2019)

LoA level	Human Pilot	Aircraft System and control
1:No automation	All operations	Warns Protect
2:Limited assist; Auto throttle	Drives In-the-loop	Guides Assist
3:Assist, Tactical; Supervised	On-the loop Pilot monitors all time	Manage movement within defined limits
4:Automated Assist Strategic	Out-of-loop Asked by system	Flies, but may give back control
5: Autonomous	Completely out-of-loop	Flies with graceful degradation

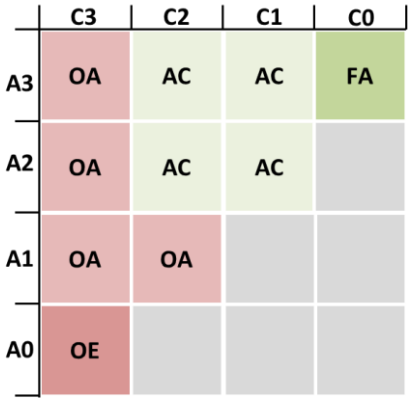


Figure 2: LoA defined as the human's need to be available

This creates four types of LoA: OE – Operator Exclusive, where an operator is needed at all times; OA – Operator Assisted, where an operator can use own judgment as for long they can be away; AC – Autonomous Control where an operator can leave control station and will be alerted in time for safely retaking control; and FA – Full Autonomy, where no operator is attending at all. The unlabeled squares are impossible as operator attendance does not match automation capabilities (Rødseth 2021b). This type of classification says something about how long the operator can be away from controls, but does not say anything about how automation assists the human.

The problem of LoA also gets more complicated when the system under consideration contains several more or less independent processes. An example at hand is again ships, where automation and humans must cooperate to control, e.g., stability and water ingress, cargo conditions, fire safety, and more, in addition to the more obvious lookout and navigation functions. Likewise, a ship on a 40-day voyage from China to Europe will encounter a wide range of different operational conditions, which may require different LoA for the different processes and phases of the operation.

Adding other dimensions to the LoAs definition may be beneficial for better understanding humans' tasks and avoiding misconceptions, e.g., cognitive efforts of the tasks and responsibilities and environmental factors (Figure 3).

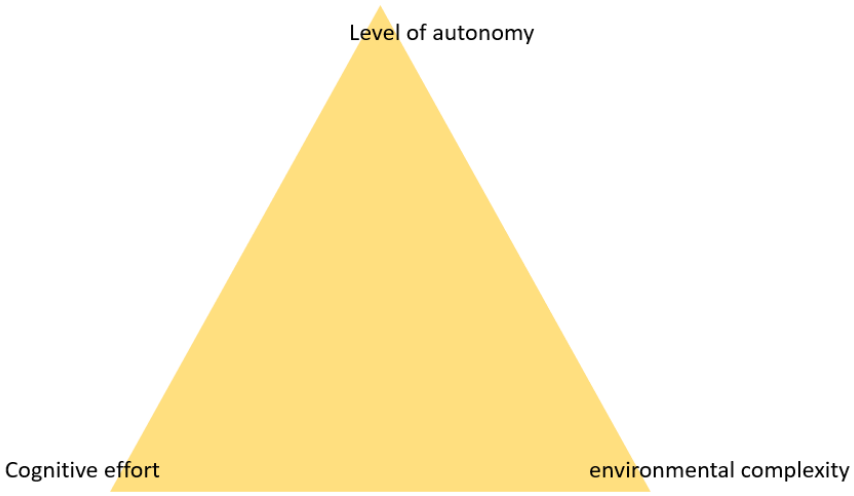


Figure 3: Illustration of LoA dimensions

Different Autonomous Systems, Same Framework?

Research on human factors applied to automated systems is not new, and many frameworks and methods exist. Yet, many of these methods and ideas may be outdated and not relevant. Moreover, autonomous systems are more complex than highly automated ones, resulting in specific human factors issues unique to fully automated systems. Several issues can be common to different autonomous systems in general, such as automation complacency or boredom due to long monitoring time. Yet, distinct operational aspects of specific modes may lead to unique issues. For instance, the "lack of feel" of an operator remotely controlling a vessel from a control center onshore does not apply to a self-driving car in which the human is onboard the vehicle. Additional aspects include training, as mentioned in the introduction, and available time for reaction. While the available time for avoiding an imminent collision in a car may be one second only, operators of an autonomous ship would typically have a much longer horizon for acting.

This problem is also very much linked to the many different "definitions" of autonomy. On the one hand, there are opinions that machine autonomy can only be achieved when the machine is entirely independent of a human, which in most cases will render it useless for practical use, as it cannot be instructed. On the other hand, automatic guided vehicles operating in a semi-controlled

environment with rudimentary anti-collision systems are also called autonomous.

Observing the implementation of different automated systems in different modes and different areas, one concludes that automation is not perfect; the sensors can fail or misunderstand the environment. The consequences of these failures are dependent on the use-case (i.e., operational design domain). If the operations are simple and there is redundancy or focus on high-reliability operations, the probability of failure can be low. In addition, if we manage to ensure meaningful human control and a forgiving operational domain, then the consequences of automation failure will be reduced. However, a basis for successful automated systems is understanding user requirements and building on meaningful human control (see Johnsen et al. (2020).

Given the differences between the various autonomous systems concerning LoAs, onboard or remote operation, and time constraints, do these systems need to be analyzed through different human factors/human reliability frameworks? Or can they be analyzed through the same framework, adapted for reflecting their unique aspects? There are indications supporting the latter. The community needs to identify systematic differences in how systems are being implemented, the different environments, the different times for action, and the type of people who use these systems (background, training, etc.).

Research Directions

The discussions on the human role in autonomous systems and human on the loop summarized in this report can form a foundation for future research on the topic. It is clear that autonomous systems design must address how humans will interact with the system and that this is crucial for reaching a safe operation. Further, while taxonomies need to be concise and straightforward, they are critical since they are used for developing policies and a shared understanding of the system. For instance, the term "cyber-physical systems" (CPS), often used to refer to autonomous systems, can be misleading for excluding the human element – still involved depending on the LoA. Terms such as cyber-physical-human systems (CPHS) or liveware may be more representative. Similarly, how the community is approaching LoAs must be re-discussed for clarifying human roles.

Applying the same method/framework for analyzing human-system interaction across different autonomous systems may be possible. However, the differences and similarities between these systems mean in terms of human interaction modeling and methods must be further investigated. Finally, the field needs more interdisciplinary research and discussions on terminologies.

While the discussions summarized above have a primary focus on transportation systems, it should be noted that the considerations on the human role, human-system interaction and human error must be extended to other autonomous systems such as industrial processes and factories.

References

Eriksson A & Stanton N A "Takeover time in highly automated vehicles: noncritical transitions to and from manual control". *Human factors* 59(4), 689-705. 2017

European Commission. "Ethics Guidelines for Trustworthy AI." Brussels. 2019

Johnsen, S. O.; Evjemo, T. E. . "State of the art of unmanned aircraft transport systems in industry related to risks, vulnerabilities and improvement of safety". In *Proceedings of the 29th European Safety and Reliability Conference (ESREL)*. 22–26 September 2019 Hannover, Germany. 2019

Johnsen, S. O., Holen, S., Aalberg, A. L., Bjørkevoll, K. S., Evjemo, T. E., Johansen, G., & Porathe, T. "Automation and autonomous systems: Human-centred design in drilling and well." . SINTEF Digital. 2020

Kaber, D. B. "Issues in human–automation interaction modeling: Presumptive aspects of frameworks of types and levels of automation*". Journal of Cognitive Engineering and Decision Making*, 12(1), 7-24. 2018

National Transportation Safety Board. 2019. "Collision between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, HWY18MH010, Tempe, Arizona," 1–5. https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf.

Ramos, M., Utne, I. B., Mosleh, A. "Collision avoidance on maritime autonomous surface ships: operators' tasks and human failure events". *Safey Science, 116* . Pg. 33-44. 2019

Rødseth Ø.J. "Constrained Autonomy for a Better Human–Automation Interface", In *Sensemaking in Safety Critical and Complex Situation*s, CRC Press, Boca Raton & Oxton, ISBN 978-1-003-00381-6. 2021

Rødseth, Ø.J., Wennersberg, L.A.L. and Nordahl, H. "Towards approval of autonomous ship systems by their operational envelope". *Journal of Marine Science and Technology.* 2021

IWASS 2021

Main challenges with respect to autonomous systems SRS: Human on the loop

- Are our current human factors methods compatible with these new technologies and the new pace of technology?
- Do our current methods fully and adequately explore the changing role of the human in an increasingly automated and autonomous world
- Do autonomous systems really reduce risk compared to conventional systems? How do risks change with higher autonomy levels?
- Are humans as back-up for automation at all a good idea? Operations envelopes for the automation system and good criteria seem to be important.
- What are the challenges and solutions related to shared autonomy? Shifting autonomy and responsibilities?
- How should autonomous systems communicate in operation? How should they communicate with non-autonomous systems? How to validate and verify systems with respect to the interaction with the operators/users?
- How should autonomous systems communicate safety to users and third-party stakeholders?
- What skills of human operators/supervisors are needed?
- How will/should human operators be included in the operation and decision making of autonomous systems?
- How should complex decision-making systems interact with humans operators/supervisors/ users?
- How will/should autonomous systems interact with conventional systems and/or human operators?
- What will be the role of the human operator during normal operation and in cases of emergency situations/failure of the autonomous system?
- How can security training for users and operators be ensured?

IWASS 2021

Group Participants

Jiaqi Ma – *Session Chair*
University of California, Los Angeles, USA

Marilia Ramos – *Session Chair*
University of California, Los Angeles, USA

Stacy Balk
National Highway Transportation Safety Authority, USA

Arne Ulrik Bindingsbø
Equinor, Norway

Claire Blackett
Institute for Energy Technology, Norway

Jens Einar Bremens
Norwegian University of Science and Technology, Noway

Åse Hoem
Norwegian University of Science and Technology, Noway

Thomas Johansen
Norwegian University of Science and Technology, Norway

Stig Ole Johnsen
SINTEF Digital, Norway

Salvatore Massaiu
Institute for Energy Technology, Norway

Jana Price
National Transportation Safety Board, USA

Ørnulf Jan Rødseth
SINTEF Ocean, Norway

Michael Gaither
Texas A&M University

Marija Slavkovic
University of Bergen, Norway

Asun Lera St. Clair
DNV, Norway

Nikolaos P. Ventikos
National Technical University of Athens, Greece

Krzysztof Wróbel
Gdynia Maritime University, Poland

Modeling and simulation for understanding complexity and cascading failures

Report of the Discussions of Breakout Session

Authors: Utne, I. B.; Aldemir, T.
Contributors: Andrews, J.; Bye, A.; Guarro, S.; Haugen, S.; Kim, H.; Louzis, K.; Montewka, J.; Parhizkar, T.; PeterA.; Rao, A.; Sankaran, K.; Thomas, S.; Titlestad, K.; Yang, R.

General methodology

Risk assessment of autonomous systems is a challenging task and therefore requires a "framework" consisting of various types of methods to identify, analyze and evaluate different hazards and hazardous events. This framework should consist of both qualitative and quantitative methods and approaches. Quantitative risk metrics, for example, are needed for autonomous systems to know how much safety is exceeded.

In general, there is a need for bridging and exploring the gap between simulation and discrete logic methods. Binary logic trees are crude and cannot cope with the complexity of an autonomous system. Exploring methods that try to bridge the gap between traditional binary models and simulation is important. A simulation cannot process all the different ways a scenario can develop into. This territory hasn't been explored, especially in logic methods. In some cases, the Dynamic flowgraph methodology (DFM) may be feasible. Bayesian Networks (BN) could be used to "couple the gap".

To some extent, dynamic probabilistic risk assessment (DPRA) methods have been implemented for autonomous systems. Simulation is partly used in Probabilistic Risk assessment (PRA) and Human Reliability Analysis (HRA). It is important to know which methods can be used for what purpose, for example, related to human behavior, because we may have to make a lot of assumptions and definitions of limitations/boundaries. We are not able to handle humans and nature to an acceptable extent in simulation, so should dynamic simulation be used for human interaction or only for the technical system?

The Hybrid Causal Logic (HCL) provides one example of an existing framework, combining scenario-based models and risk assessment methods. HCL may be connected to a «safety case approach» used to check if a design works or not and how much safety is required. Simulation related to HCL could be linked

to the implemented Event Sequence Diagrams (ESD), where the ESDs are considered an in-between step to be converted into an executable simulation.

Autonomous systems may be complex and challenging to analyze and model. Compartmentalization may be a way forward for decomposing complex systems. Then different models and methods can be used for each subsystem – in a hybrid approach for different problems and parts. A challenge is that methods that may work at a lower system level might struggle when attempting to implement the results at a higher level, due to scalability or prioritization issues. Defining system boundaries and interactions are crucial.

A lot of work related to DPRA has been done in the nuclear industry, and at institutions, such as Idaho National Lab, Ohio State University and University of Maryland, but the methods have not been transferred to other applications such as autonomous systems, except in isolated cases, and infusion in other areas is slow. For example, the automotive and maritime industries rely more on failure mode and effect analysis (FMEA) than on fault trees. A challenge is that the nuclear plant is relatively static and stationary, which is different from operating autonomous vehicles. Also, in many industrial environments, a lot of uncertainties arise when trying to increase the operational envelope.

An important question is related to what type of models are better suited as means of validating autonomous systems requirements (this is what "validation" is to systems engineers, i.e., demonstrating that functional requirements are correct); and what types are useful to show that a system design does indeed satisfy requirements ("verification"). More efforts should also be put into this area for clarification and further development.

Software

Incorporating software failures is the most challenging part of risk modeling. Hardware is easier to model, as its performance and impact from the environment can be predicted. However, we cannot predict how software behaves since it is not comparable to hardware failure for which we could collect failure data. Hence, it is challenging to incorporate software into the «traditional» risk and reliability models developed for hardware systems.

Probability assignment to software is tricky. A software fault can be present in the code or not, when a software specification is not coded or implemented correctly, and therefore it's not stochastic, as such, so it is difficult to assign probabilities. Faults can also be caused by an incorrect or incomplete set of specifications. This means that the software does what it is designed to do, but the designed to do is "wrong".

It is not always the software itself that causes faults and failures. The hardware platform and memory can also cause a fault, which may seem like a software stochastic or random type of failure but is actually a random failure of the underlying electronics and its interaction with the software.

Some would claim that software faults are a quality issue rather than a reliability issue. Others state that software quality relates to defects in the code, whereas software reliability relates to system failures that occur during execution. In general, there has been a disagreement for several decades on how to model software.

Software depends on hardware, and hardware can change with impact from an environment, which adds complexity. Whether a fault is triggered (triggering event) or not by operation could be considered stochastic. In addition, the effect of a fault may also be considered stochastic, since it may depend on environmental conditions.

Every time a software patch is released, re-analysis is needed, but it is impossible to go through millions of lines of code to check. Software may become so complex it might have to be treated as stochastics, also because the hardware surrounding is becoming so complex. Because of growth in software complexity and its multiple subtle interactions with the underlying computer and network environments, there are types of failures of highly complex software that are not deterministically identifiable and happen in a "pseudo-random" fashion. Perhaps we need separate approaches for dealing with the deterministic software logic specification errors or omissions, and the pseudo-random failures resulting from poorly understood and "unpredictable" interaction of software with its "environment." This may also mean that our traditional methods may become more relevant for software.

It is difficult to compartmentalize software, since the software performance may be reduced or functions may be changed.

Cybersecurity

Cyber-security is an essential part of system safety. Safety is a requirement for security.

Unknown unknowns are a challenge in all types of risk assessment, but cyber-security is a domain where this is highly relevant and important. From a cyber-security perspective, a challenge is that when you try to take the human into the loop, the modeling becomes too challenging to handle. From a cybersecurity perspective there is an infinite number of failure modes or zero, because people are so creative. There are too many ways to crash software. To

limit the number of failure modes for software (or ways software may fail), we may put more focus on their effect.

A failure can be caused by human error, but some errors are undetected. It is hard to include everything in a model, because there are too many variables in nature and humans. For comparison: You don't deal with human reliability by going through every node in your brain. A similar approach may have to be soon used for software and cybersecurity as well.

Since we cannot handle technology and consequences through existing risk assessment methods, more qualitative methods may be the only solution, or better solution.

It may be questioned whether we know the limitations of simulations. For what kind of scenarios do we not have models? This is a general problem. Generally, a lot of behavior can be monitored and simulated, but may become very hard to explain. In some cases, the behavior «breaks» out of the models, for example, the Alpha-zero chess engine behaving unpredictably and a lot of chess players giving up explaining the behavior. For the Boeing 737-max incident, simulations and models were not able to predict the particular problem.

AI and machine learning

The automotive industry is the biggest human killer in the history of vehicles, and similar operational data of autonomous vehicles will not be accepted. Hence, it is necessary to prove that an autonomous vehicle is safer and better than a human-driven vehicle before the public will trust them.

AI has potential to reduce the risk, but it is hard to control the evolution of AI and regulate it appropriately. Neural networks, for example, can potentially perform better than humans, for instance, in terms of identifying the presence of cancer cells. In the future, AI may become better than humans, for example, in determining whether pedestrians want to cross the road or not. Still, many different interactions and variables could make it difficult to predict for the AI. Further, neural networks, for example, are hard to verify.

Currently, designers of vehicles are given large neural networks spitting out a prediction whether passengers will cross in front of the car and then they are asked to make the vehicles behave safely. A problem with machine learning is that if you haven't covered a particular area, you don't know what will happen. In addition, probabilities are changing in machine learning, and code is changing over time.

There are no stochastic failures in AI because of its binary nature, which is a positive thing. On the other hand, sometimes AI failure is more like random

failure, as slightly different data can cause totally different results. The failure probability of "trained" autonomous system elements operating in different environments could be obtained by fusing neural networks into a BN and simulating different failures. One way to potentially improve system safety is to have a good system architecture that enables decomposition, and as previously mentioned – compartmentalization.

Main challenges with respect to autonomous systems SRS: Modeling and simulation, cascading failures

- A challenge with hazard identification and risk analysis of novel systems is to identify everything that can go wrong. How can we deal with the unknown unknowns?
- How to capture the large spectrum of potential risk scenarios and contexts?
- What methods are feasible for analyzing and modeling cascading failures in systems? Between hardware, software, humanware etc.? What are important requirements to such methods?
- How can we be sure that the models are complete and even discuss it, if the organizational issues are not taken into account or completely outside of the box?
- How is software risk different from software reliability?
- Are software failures deterministic or probabilistic?
- What are the challenges related to simulation in probabilistic risk assessment? Can simulation increase the risks?
- Uncertainty in sensor data is a challenge. How should this uncertainty be handled in risk assessments of autonomous systems?
- How can risk assessments and risk models of autonomous systems take shared control and "adaptive autonomy" sufficiently into account in the identification of hazards and the analysis of risk?
- How can vulnerabilities in the software and communication systems of autonomous systems be reduced to mitigate cyber-attacks and security problems?
- What must risk models for software intensive and AI-based systems convey to convince users, regulators and standardization bodies?
- How to handle the state space problem when validating an autonomous system?
- What data is needed to assess the performance of autonomous systems?

Group Participants

Ingrid B. Utne – *Session Chair*
Norwegian University of Science and Technology, Norway

Tunc Aldemir - *Session Chair*
Ohio State University, USA

John Andrews
University of Nottingham, United Kingdom

Andreas Bye
Institute for Energy Technology, Norway

Sergio Guarro
ASCA Inc., USA

Stein Haugen
Norwegian University of Science and Technology, Norway

Hyungju Kim
University of South-Eastern Norway

Konstantinos Louzis
National Technical University of Athens, Greece

Jakub Montewka
Gdynia University/World Maritime University, Poland

Tarannom Parhizkar
University of California Los Angeles, USA

Anto Peter
Teradyne, USA

Alan Rao
United States Department of Transportation OST-R/Volpe Center, USA

Karthik Sankaran
University of California Los Angeles, USA

Stephen Thomas
University of Maryland/Motional, USA

Kenneth Titlestad
Sopra Steria, Norway

Ruochen Yang
Norwegian University of Science and Technology, Norway

Artificial Intelligence and Data Analytics in Resilient Autonomous Systems

Report of the Discussions of Breakout Session

Authors: Thieme, C. A.; Morozov, A; Blindheim, S.; Maidana, R.
Contributors: Hu, Y.; Katsikas, S.; Mengshoel, O.; Smidts, C.; Torben, S.

Artificial intelligence (AI) and data analytics methods can be applied to autonomous systems in two ways; (i) as part of the autonomous system, and (ii) as part of the safety, reliability and security (SRS) assurance efforts. Two questions arise from this perspective that built the foundation of the discussion in this breakout session; 1) How can AI systems used in autonomous systems be made safe, and 2) how can AI methods be applied to support assurance efforts of autonomous systems. These relationships are depicted in Figure 4, developed during the breakout session with the participants.

On the left side of the figure, we see typical components of an autonomous system, for instance, a self-driving car, a small unmanned arieal vehicle (UAV), or a big autonomous ship. Here we highlight the software that implements an AI-based algorithm, e.g., a Convolutional Neural Network (CNN) that receives images from the camera of a car recognizing and classifying road signs. It is to be noted that the human element was added to the figure. As humans are the final decision-makers in most cases, their influence weighs heavily on autonomous operation. Humans may also interfere with autonomous systems as rogue or adversarial agents. On the right side, we see the classical methods for SRS assurance that can be applied to an autonomous system. Here we also highlight AI techniques that can be exploited to improve the performance and precision of the SRS methods.

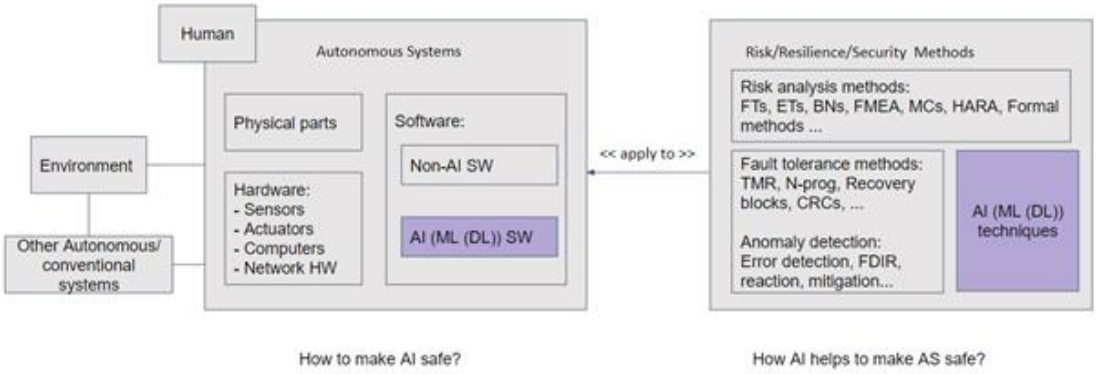


Figure 4 Autonomous systems and their relationship with AI methods, developed during the breakout session.

How to make Artificial Intelligence in autonomous systems safe?

To answer the question "How to make AI safe?", it is necessary to identify what makes software employing artificial intelligence methods different from conventional (non-AI) software, and where this software is employed. Conventional software is often claimed to be deterministic, predictable, and explainable, i.e., that for a certain input one can determine what the output will be and explain why it is so. If faulty inputs are given (e.g. due to a faulty sensor), it is possible to determine what the expected output from the conventional software will be. For software with AI methods, this output is in some cases unpredictable, since it is not known how the AI's decisions are made, for example for deep neural networks and other "Blackbox approaches". This leads to uncertainty and increased difficulty when testing autonomous systems with AI, since deriving conclusions from limited edge cases may not hold true over the whole state space of input combinations.

Regarding the application of AI methods in the software of autonomous systems, several applications are imaginable. AI may be used to generally analyze the environment, for example through image detection or evaluation of other system sensors to extract information. AI could also be used to detect abnormal system, software or environmental conditions that require a change of the current actions. Furthermore, AI can be used to control an autonomous system or parts of its actuators. AI can additionally be part of the entertainment system or human-machine interface of an autonomous system. These application areas show that AI may be applied in different layers of autonomous systems, and hence SRS efforts need to be tailored to the level of application. The interaction between the different layers of software are also of interest for SRS assessment as they may interfere with each other, potentially leading to undesired consequences. There are current research efforts to address these interactions - e.g., using an overall "supervisor", which monitors the software stack as a whole

and enacts emergency behaviors when it detects these potentially dangerous interactions.

When AI software is used to make decisions regarding (emergency) control of an autonomous systems, the AI methods should have the following properties:

- Explainability: Why was this decision or interpretation made?
- Transparency: Is the decision made or interpretation made possible to make?
- Interpretability: What does the decision/interpretation mean for the system, its future state and its environment?

These are also important elements that form trust of users and operators of AI-based systems. They may be more inclined to accept decisions and risks associated with the decisions if they understand the reasoning and have the feeling of control or being on the loop. The role of humans in an autonomous system in general needs to be discussed for each system. The human's role will determine the information that the system needs to display. Important questions that arise from this discussion are; How similar should the reasoning of an AI-based system be to human decision-making? What information should be presented to the user? If we could perfectly replicate human behavior in an AI-based system, would we be able to trust this program fully?

Reliability and resilience to disturbances of an AI-based autonomous system are important concepts with respect to trust and the three concepts mentioned above. Concerning the assurance of reliability and resilience, one issue is the assessment of hardware failure effects on the software, and vice versa. The assessment of the interactions among the physical and software components is far from trivial, due to the state space that needs to be covered. In addition, interactions among different hardware components and among different software components need to be taken into account. One software component may be robust to a hardware component failure, while another software component may not, leading to conflicting information and a resulting fault.

Bit-flips in CPU, RAM, network components, or sensors are common examples of a hardware fault propagating to the software level and leading to so-called silent data corruption. This happens in space systems because of the cosmic radiation as well as on the on-ground systems because of fluctuating voltage, heat, or induction. These effects are well-studied for non-AI software. Similar analysis has to be applied to deep-learning (DL)-based components because different neural network architectures exhibit different resilience levels to such kinds of faults. Thus, for AI-based systems, in particular DL systems, the

architectures' robustness against hardware faults should be assessed. One approach is to carry out extensive fault injection experiments. Another approach to solving this challenge could be "formal methods", AI methods and AI-based systems that help assuring the correctness of the system in different situations. Also, supervisors are needed that monitor the system and assist in resolving these situations.

Security of AI-based autonomous systems is another key aspect of discussion. No clear definitions exist yet on what constitutes an attack on an AI-based system. For example, a drawing on the road made by kids, may be interpreted by an autonomous car as an obstacle and thus the car stops. Will this constitute an attack?

An additional aspect regarding the security of AI-based systems is the detection of attacks. Image recognition algorithms have been shown to be prone to tempering, where just a few parts of a traffic sign have been altered, leading to misinterpretation of the sign. More generally, even small changes in the data input may significantly affect the results of the AI algorithm. How will it be possible to detect these types of attacks, when the system is certain about its interpretation? This is just a starting point of the largely unsolved problem of how an autonomous system's security against, among others, tampering can be assessed.

How to use artificial intelligence to make autonomous systems safe?

Several possibilities are imaginable regarding how AI methods can contribute to SRS assessment of (autonomous) systems. Among these are:

- Reading system code and technical documentation, e.g. structural and behavioral diagrams, (deep learning) for existing systems as a basis to automatically produce SRS models, e.g. fault trees.
- Extraction of statistical data for use in SRS models. i.e., conditional probabilities for Bayesian Belief Networks. SRS models require a lot of statistical input data, e.g. the failure probabilities of specific components or failure scenarios. The more complex and precise an SRS model is, the more data it requires.
- Text analysis to extract risk influencing factors and scenarios for risk assessment.
- Optimization of SRS analytical and simulation models. AI can help to create more computationally efficient SRS models or reduce available models to more compact equivalents.

- Optimization and planning of maintenance of (autonomous) systems. DL-based predictive maintenance.
- Detection of anomalies during operation to inform (automatic) decision-making. DL-based anomaly and error detection.

One challenge when applying AI methods for SRS efforts with respect to quantification is the difference in accuracy/uncertainty. Typically, for an AI algorithm to be considered to give good predictions, 99 % accuracy or confidence in the results is needed; however, this depends highly on the application. Yet, in risk and reliability analysis, probabilities and certainties are needed that are several magnitudes lower when calculating the risk level.

The detection of tampering or intrusion of a threat agent into a system with the help of AI is being explored. However, since AI methods rely on data, the detection can only capture scenarios that have been in the dataset. Hence, the multitude of possible attacks and limitations of the datasets make the current AI algorithms for security incident detection not useful.

Conclusion and future research directions

Concluding from the discussion; AI should be used intelligently, combining expertise with reliable data. AI methods should not be applied just for the sake of applying AI to build an autonomous system. AI methods need to be combined with suitable methods that are tailored to the case of the autonomous system being developed.

When applying AI methods, ethical considerations have to be addressed. These considerations should not start and end with the often cited "Trolley-Problem". Ethical considerations have to address critical issues, such as consent of third-party people, responsibilities and the socio-economic impact of the introduction of the autonomous systems. Therefore, broadly interdisciplinary work is necessary to develop ethical AI-based systems holistically. It was noted that the current research funding schemes are not sufficient to truly allow for a holistic approach.

Finally, following from the discussion, several questions should be addressed by the research community to solve the challenges associated with AI-based autonomous systems concerning safety, reliability, resilience, security:

- What are or should be safety-critical and security-critical AI applications in autonomous systems?
    - Can classical risk, resilience, and security analysis methods be applied to make AI-based systems safe?

- What are the differences between conventional (non-AI) and AI software from a safety point of view?
- How robust are different DL architectures against hardware faults?
- What makes humans trust other humans, and how can AI be trusted?
- How do AI-based systems interact with the environment, humans, and a combination of autonomous systems and non-autonomous systems?
- How can AI be used to generate risk, resilience, and security models?
- How can AI be employed to optimize classical risk, reliability, security models and their evaluations?
- Can AI be used to generate data to feed probabilistic models?
- How can AI-based anomaly detection and decision-making support safety and security of autonomous systems?

Main challenges with respect to autonomous systems SRS: Artificial Intelligence and Data Analytics in Resilient Autonomous Systems

- How to make AI Safe?
- What is the relationship between autonomy and AI? Common software and AI software?
- What data is needed to assess the performance of autonomous systems?
- How AI can help to make autonomous systems safe?
- How can AI methods be used to demonstrate safety and compliance? How can risk assessment and modeling be coupled to AI?
- Is AI accuracy low if we compare with the probability of failure?
- How can we design resilience into autonomous systems by use of AI?
- Resilient robots are needed for increased levels of autonomy. Suggested attributes of resilient robots are robustness, redundancy and resourcefulness. How can AI be used to achieve these?
- What machine learning approaches are most feasible for providing input to risk assessment?
- Improved intelligence and online decision-making capabilities are needed in autonomous systems. Risk assessments performed and utilized by robots means that risks have to be detected and quantified. What risks are not possible to identify by an autonomous system?
- What risks cannot be quantified? How does this impact safety of autonomous systems?
- Can risk really be translated into cost functions?
- Who is responsible for decisions made by the autonomous system (in case of accidents)? How can dependable situational awareness systems be build?
- Are AI systems suitable for making ethical decisions – what is necessary for ethical decision-making?
- How is risk assessment and situation awareness linked for an autonomous system?
- How are the creation and economic incentives behind algorithms affecting safety?
- How to map and balance the needs and requirements from a behavioral and operative perspective?

- How should complex decision-making systems interact with human operators/supervisors/users?

Group Participants

Andrey Morozov - *Session Chair*
University of Stuttgart, Germany

Christoph A. Thieme - *Session Chair*
Norwegian University of Science and Technology, Norway

Simon Blindheim
Norwegian University of Science and Technology, Norway

Yunwei Hu
Amazon, USA

Sokratis Katsikas
Norwegian University of Science and Technology, Norway

Renan Maidana
Norwegian University of Science and Technology, Norway

Ole Jakob Mengshoel
Norwegian University of Science and Technology, Norway

Carol Smidts
Ohio State University, USA

Sverre Torben
Kongsberg Maritime, Norway

Organizing Committee

Christoph A. Thieme, PhD – NTNU

Dr. Christoph Thieme obtained his PhD in Marine Technology from NTNU. He has experience with risk analysis and modelling of autonomous marine systems. Until recently he was a postdoctoral research fellow at NTNU in the UNLOCK project, working on risk assessment methods development and applications on autonomous control systems. His main research interest is the software and control systems' effect on the risk level of autonomous systems.

www.christophthieme.com christoph.thieme@ntnu.no

Marilia Ramos, PhD - UCLA

Dr. Marilia Ramos is a Research Scientist at the B. John Garrick Institute for the Risk Sciences, UCLA. She has a PhD in Chemical Engineering from the Federal University of Pernambuco, Brazil. Her expertise is on Risk Analysis and Human Reliability, and she has extensive experience in projects concerning the oil and gas field. Currently, she applies risk and reliability analysis to complex systems, including autonomous systems, the process industry, and the energy field. Her main research interest is on the human-software-hardware interaction in such systems.

www.mariliaramos.net marilia.ramos@ucla.edu

Ingrid B. Utne, PhD – NTNU

Dr. Ingrid Bouwer Utne is a Professor at Department of Marine Technology, NTNU, with expertise on safety and risk assessment of marine systems. Utne is the Co-Director of the research center SFI Autoship. She is the principal investigator and project manager of research/industry projects associated with the Center of Excellence on Autonomous Marine Operations and Systems (NTNU AMOS). Her research focuses on supervisory risk control, bridging risk management and engineering cybernetics enhancing the safety and intelligence of autonomous systems.

ntnu.edu/employees/ingrid.b.utne ingrid.b.utne@ntnu.no

Ali Mosleh, PhD – UCLA

Dr. Ali Mosleh is Distinguished University Professor and holder of the Knight Endowed Chair in Engineering at UCLA, where he is also the director of the Institute for the Risk Sciences. He conducts research on methods for probabilistic risk analysis and reliability of complex systems and has made many contributions in diverse fields of theory and application. He was elected to the US National Academy of Engineering in 2010 and is a Fellow of the Society for Risk Analysis, and the American Nuclear Society. Prof. Mosleh is the recipient of many scientific achievement awards.

riskscience.ucla.edu/institute-director mosleh@ucla.edu

Organizers and Sponsors

Department of Marine Technology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

The Department of Marine Technology at NTNU provides world-class education and research for engineering systems in the marine environment. The focus is on methods and techniques for sustainable development and operation of ship technology, fisheries and aquaculture technology, oil and gas extraction at sea, offshore renewable energy, and marine robotics for mapping and monitoring the ocean. The Department hosts an excellent research group working on safety and risk management of marine and maritime systems. The Centre of Excellence Autonomous Marine Operations and Systems (NTNU AMOS) is also located at the Department.

The Norwegian University of Science and Technology in Trondheim (NTNU) is the largest technical university in Norway.

The B. John Garrick Institute for the Risk Sciences, University of California, Los Angeles, USA

The B. John Garrick Institute for the Risk Sciences has declared its mission to be the advancement and application of the risk sciences to save lives, protect the environment and improve system performance. The purpose of the Garrick Institute is for the research, development, and application of technology for (1) quantifying the risk of the most serious threats to society to better enable their prevention, reduce their likelihood of occurrence or limit their consequences and (2) improving system performance with respect to reliability and safety. The institute is hosted at the Department of Engineering at the University of California Los Angeles (UCLA).

DNV

DNV is a global quality assurance and risk management company. DNV provides classification, technical assurance, software and independent expert advisory services to several industries. Combining technical, digital and operational expertise, risk methodology and in-depth industry knowledge, DNV GL assists its customers in decisions and actions with trust and confidence. With origins stretching back to 1864 and operations in more than 100 countries. DNV are dedicated to helping customers make the world safer, smarter and greener.

Kongsberg Maritime

Kongsberg Maritime (KM) is a leading supplier of offshore and marine energy solutions, deck machinery and automation systems. In addition, KM provides services related to complex system integration, and vessel design. KM is a leader in marine ship intelligence, automation and autonomy and is a part of the Kongsberg Group.

Research Council of Norway

The Research Council of Norway serves as the chief advisory body for the government authorities on research policy issues, and distributes roughly nine billion Norwegian kroners to research and innovation activities each year. The Research Council of Norway co-financed the IWASS workshop through the MAROFF knowledge-building project for industry ORCAS (Project number 280655) and the FRINATEK project UNLOCK (Project number 274441).

IWASS Participants

Organizing Committee

Ali Mosleh	*The B. John Garrick Institute for the Risk Sciences, University of California Los Angeles (UCLA)*
Christoph A. Thieme	*Norwegian University of Science and Technology (NTNU)*
Ingrid B. Utne	*NTNU*
Marilia A. Ramos	*The B. John Garrick Institute for the Risk Sciences, UCLA*

Speakers

Claire Blackett	*Institute for Energy Technology*
Marie Farrell	*Maynooth University*
Marija Slakvovik	*University of Bergen*
Stacy Balk	*National Highway Traffic Safety Administration*
Stephen Thomas	*University of Maryland/Motional*
Tim Johnson	*National Highway Traffic Safety Administration*

Participants

Alan Rao	*United States Department of Transportation OST-R/Volpe Center*
Alexandros Koimtzoglou	*National Technical University of Athens*
Andreas Bye	*Institute for Energy Technology*
Andrey Morozov	*University of Stuttgart*
Anto Peter	*Teradyne*
Arne Ulrik Bindingsbø	*Equinor*
Åsa Hoem	*NTNU*
Asgeir, J. Sørensen	*NTNU*
Asun Lera St. Clair	*DNV*
Carol Smidts	*Ohio State University*
Chris Harrison	*Rail Safety and Standards Board*
Daniel Metlay	*The B. John Garrick Institute for the Risk Sciences, UCLA*
Erik Aleksander Veitch	*NTNU*

Hyungju Kim	*University of South-Eastern Norway*
Jakub Montewka	*Gdynia University/ World Maritime University*
Jiaqi Ma	*The B. John Garrick Institute for the Risk Sciences, UCLA*
John Andrews	*University of Nottingham*
Jon Arne Glomsrud	*DNV*
Kenneth Titlestad	*Sopra Steria*
Konstantinos Louzis	*National Technical University of Athens*
Krzysztof Wróbel	*Gdynia Maritime University*
Lance Fiondella	*University of Massachusetts Dartmouth*
Martin Feather	*Jet Propulsion Laboratory, California Institute of Technology*
Mary Ann Lundteigen	*NTNU*
Michael Gaither	*Texas A&M*
Michael Parsons	*University of York*
Myron Hecht	*Aerospace Corporation*
Nancy Currie-Gregg	*Texas A&M*
Niels Nijdam	*University of Geneva*
Nikolaos P. Ventikos	*National Technical University of Athens*
Ole Jakob Mengshoel	*NTNU*
Ørnulf Jan Rødseth	*SINTEF Ocean*
Osiris Valdez Banda	*Aalto University*
Salvatore Massaiu	*Institute for Energy Technology*
Sergio Guarro	*ASCA Inc.*
Siri Granum Carson	*NTNU*
Siv Randi Hjørungnes	*Kongsberg Maritime*
Sokratis Katsikas	*NTNU*
Stein Haugen	*NTNU*
Stig Ole Johnsen	*SINTEF digital*
Sverre Torben	*Kongsberg Maritime*
Tarannom Parhizkar	*The B. John Garrick Institute for the Risk Sciences, UCLA*
Thomas Porathe	*NTNU*
Tunc Aldemir	*Ohio State University*
Xue Yang	*Dalian Maritime University*
Yan-Fu Li	*Tsinghua university*
Yunwei Hu	*Amazon*
Zhang Di	*Wuhan University of Technology*

Published by:
The B. John Garrick Institute for the Risk Sciences,
University of California Los Angeles
404 Westwood Plaza, Engineering VI
Los Angeles – 90095, California
www.risksciences.ucla.edu

NTNU
Norwegian University of
Science and Technology

The B. John Garrick Institute for the Risk Sciences
UCLA ENGINEERING