# IWASS 2023

International Workshop on Autonomous Systems Safety

# Proceedings

September 2-3, 2023
Southampton | UK

Universität Stuttgart

NTNU – Trondheim
Norwegian University of Science and Technology

UCLA ENGINEERING
B. John Garrick Institute for the Risk Sciences

IWASS 2023 International Workshop on Autonomous Systems Safety

# Proceedings of the 4th International Workshop on Autonomous Systems Safety


**Edited by:**

Camila Correa-Jullian, Joachim Grimstad, Spencer August Dugan, Marilia Ramos, Christoph A. Thieme, Andrey Morozov, Ingrid B. Utne, Ali Mosleh



DOI: 10.34948/G4MW2N

ISSN: 2995-8709

December 2023



UCLA ENGINEERING
B. John Garrick Institute for the Risk Sciences

NTNU – Trondheim
Norwegian University of Science and Technology

Universität Stuttgart

Published by:

The B. John Garrick Institute for the Risk Sciences, UCLA
www.risksciences.ucla.edu

Norwegian University of Science and Technology (NTNU)
www.ntnu.edu/imt

University of Stuttgart
https://www.uni-stuttgart.de/

# 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-GIRS), the Norwegian University of Science and Technology (NTNU) and the Institute of Industrial Automation and Software Engineering of the University of Stuttgart.

IWASS is an invitation-only event, designed to be a platform for **cross-industrial** and **interdisciplinary effort** and **knowledge exchange** on autonomous systems' Safety, Reliability, and Security (SRS). The workshop gathers experts from academia, regulatory agencies, and industry to **identify and propose solutions** for common challenges related to SRS of autonomous systems. It complements existing events organized around specific types of autonomous systems (e.g., cars, ships, aviation) or the safety or security-related aspects of such systems (e.g., cyber risk, software reliability). IWASS **envisions a future** where autonomous systems enrich human life while upholding the highest **safety, reliability, and security standards**.

Previous editions of IWASS (*2019* – Trondheim, Norway; *2021* - online; *2022* – Dublin, Ireland) successfully assembled a broad and diverse field of experts from different organizations and countries. **The IWASS proceedings summarize the discussions held during the events and provide a strong foundation concerning autonomous systems SRS**, ranging from risk analysis methods, and cascading failures to "human on the loop" and regulations: 2019[1], 2021[2], 2022[3].

IWASS 2023[4] took place on September 2nd and 3rd in Southampton, United Kingdom, and gathered 39 participants from 30 organizations from around the globe. In addition, a panel session at the European Safety and Reliability Conference (ESREL 2023) discussed the workshop's main conclusions and additional points with a larger audience. This report summarizes IWASS 2023 discussions. It provides an overview of the main points raised by a community of experts on the status of autonomous systems SRS. It also outlines research directions for safer, more reliable, and secure autonomous systems of the future.

---

[1] Proceedings to the 1st International Workshop on Autonomous Systems Safety. Trondheim – Norway, 11-13 March 2019. https://bit.ly/2SsPrLd

[2] Proceedings to the International Workshop on Autonomous Systems Safety 2021. 20, 21 and 28 March 2021. https://www.risksciences.ucla.edu/iwass-2021-proceedings

[3] Proceedings to the International Workshop on Autonomous Systems Safety 2022. Dublin - Ireland, 28 March 2023. https://www.risksciences.ucla.edu/iwass-2022-proceedings

[4] International Workshop on Autonomous Systems Safety 2023. https://www.risksciences.ucla.edu/iwass-2023-home

THIS PAGE INTENTIONALLY LEFT BLANK

# Table of Contents

# Introduction

The International Workshop on Autonomous System Safety (IWASS) 2023 is the fourth edition of the workshop series on Autonomous System Safety, Reliability, and Security initiated in 2019. In the near future, the scope of autonomy is expected to increase across multiple industries, systems, and operations, aiming for safer and more efficient operations. As these systems evolve to higher complexity, so do the challenges in assuring system safety. The aim of IWASS is to congregate experts from industry, academia, and regulations, to foster discussions and explore potential solutions to these many challenges. At IWASS, several discussion groups cover multiple topics about the methods used for modeling, verification, validation and testing of autonomous systems, as well as the challenges brought by increased system complexity, cascading failures, the use of Artificial Intelligence (AI) techniques, and the role of humans in autonomous systems.

The first IWASS (Trondheim, Norway – 2019) counted with participants from eight different countries representing a diversity of industries and expertise. The proceedings published by NTNU summarize the discussions held at the workshop in addition to six research papers on autonomous systems SRS.

Initially planned as an in-person event in Los Angeles in 2020, IWASS 2021 switched to an online event in 2021, due to the COVID-19 pandemic and related travel restrictions. IWASS 2021 assembled a broad and diverse field of experts with participants from 39 different organizations and nine countries. The workshop program was distributed over three days and included domain experts' presentations and discussion sessions, summarized in the proceedings published by UCLA-GIRS.

The third IWASS was organized in 2022 as an in-person workshop before the European Safety and Reliability Conference (ESREL) in Dublin, Ireland. Thirty participants from ten countries attended the workshop, engaged in cross-disciplinary discussions, and explored solutions related to fundamental challenges associated with SRS of autonomous systems. The workshop presentations and main discussions were summarized in the proceedings jointly published by UCLA-GIRS, NTNU, and the Technological University of Dublin (TU Dublin).

At IWASS 2022, the final message embodied a call to increase efforts in five different areas:

1. Including risk specialists in the design process of autonomous systems, aiming for user-centered technologies that allow human intervention, when needed, in a timely and safe manner.
2. Determining an adequate level of safety for autonomous systems. While various solutions have been proposed, including the implementation of safety envelopes and constraints, the challenge of defining acceptable risk levels persists.

3. Advancing the methods for risk assessment and Validation and Verification (V&V) processes of autonomous systems. In particular, enabling data-driven technologies presents opportunities to develop more representative simulations and scenario generations.
4. Characterizing risks associated with different Levels of Autonomy. It is crucial to examine autonomous functions in various operational modes, including transitions between different levels of autonomy and shared control with human operators.
5. Reducing the gap between academic research and experience from industry and regulators. Developing adequate safety standards and best practices may become crucial for autonomous system safety certification.

To continue and expand these discussions, the fourth IWASS took place in Southampton, United Kingdom, prior to the 33rd ESREL conference. IWASS 2023 was attended by thirty-nine in-person and online participants from eight countries and thirty organizations from academia, industry, and government. With the increasing use of safety cases for autonomous systems safety, IWASS 2023 discussions revolved around safety assurance processes, the challenges in constructing defensible safety cases, the role of academia, industry, applied researchers, and policymakers have in the discussion surrounding the use of safety cases as credible safety assurance frameworks. To support these discussions, a white paper was prepared before the workshop and distributed to all participants providing an overview of the history, motivation, and structure of the safety case (See Appendix[5]). A panel presented at ESREL 2023 showcased the main topics and questions discussed at IWASS to a broader audience.

The following sections of this document provide a summary of IWASS 2023. This includes a summary of the presentations given at the workshop by our invited speakers. This is followed by an overview of the discussions held during the workshop and the panel at ESREL. The discussion points were supplemented with information from the white paper and references where applicable. These findings provide a strong foundation to correctly approach discussions regarding autonomous SRS, as well as a path toward the safe development and operation of autonomous systems for researchers, developers, and regulatory agencies.

[5] The white paper titled 'The Safety Case for Autonomous Systems: An Overview" can also be found online at IWASS 2023 site.

# Presentations held at IWASS 2023

IWASS 2023 showcased three invited speakers to introduce the topic of safety cases, complementing the themes presented in the white paper and setting the context for the rest of the workshop. The speakers discussed the use of safety cases for autonomous and software-intensive system safety assurance processes, as well as examples of the current regulatory and legislative gaps in this area.

**Dr.-Ing Rasmus Adler**, Program-Manager Autonomous Systems, Fraunhofer-IESE, Germany

## *Safety assurance cases for Autonomous Systems from the perspective of applied science*

Dealing with safety engineering uncertainty is a significant challenge for the assurance of autonomous systems. Measures for conventional software or systems are not directly applicable or unreasonable for autonomous systems. The question of how to argue system safety for such systems is therefore explored through the lens of assurance cases. This presentation reviews the use, advantages, and ongoing research of assurance cases. The discussion concluded with relevant open issues, including issues of argument strength and sufficient safety.

iese.fraunhofer.de/en/innovation_trends/autonomous-systems rasmus.adler@iese.fraunhofer.de

**Mollie Cohen D'Agostino**, Executive Director, Mobility Science Automation and Inclusion Center (MOSAIC), Institute of Transportation Studies at UC Davis, USA

## *Mobility System Science Project: Federal and State AV Safety Standards*

Automated vehicles (AVs) may offer substantial societal benefits, including enhanced mobility, reduced traffic congestion, and improved fuel efficiency. However, the current policy landscape is unprepared for AVs, leading to legislative gaps and the potential exacerbation of transportation issues. Effective governance and collaboration between public and private sectors are essential to develop data-driven policies that ensure safe AV deployment. To achieve diverse societal goals such as reduced congestion, equitable travel options, lower emissions, and sustainable funding, federal policies must empower local and state governments while preserving their authority.

This presentation reviewed currently proposed policies to identify the primary gaps for the implementation of AVs. Future policies should prioritize AV safety, promote data sharing with privacy safeguards, and support interoperability.

mosaic.ucdavis.edu/people/mollie-cohen-dagostino mdagostino@ucdavis.edu

**Thor Myklebust**, Senior Researcher, SINTEF Digital, Norway

## *Future challenges, pitfalls, and opportunities when using a safety case approach for software-intensive systems*

The presentation explored the future landscape of safety case approaches for software-intensive systems, shedding light on challenges, potential pitfalls, and opportunities. As software increasingly underpins critical systems, understanding and presenting its safety is paramount. The presentation included the results of interviews with 36 experts from 18 companies involved in the development of safety cases. The results explored current best practices as well as emerging challenges and complexities, such as evolving regulations and rapid technological advancements, while highlighting potential pitfalls, including the challenge of documentation, insufficient communication, and argumentation. In the face of these challenges, the presentation identified opportunities for improved safety assurance, emphasizing agile development. This work aimed at equipping practitioners with insights to navigate safety cases effectively for the assurance of AI-enabled systems.

https://www.sintef.no/en/all-employees/employee/thor.myklebust thor.myklebust@sintef.no

# IWASS 2023
# Discussion Summary

UCLA ENGINEERING
B. John Garrick Institute for the Risk Sciences
NTNU – Trondheim
Norwegian University of
Science and Technology
Universität Stuttgart

## Levels of Autonomy and Automation

Autonomy can be defined as a system's or subsystem's own ability of integrated sensing, perceiving, analyzing, communicating, planning, decision-making, and acting, to achieve its goals as assigned by its human operator(s) through a designed human-machine interface (HMI).[6] The diversity of emergent autonomous systems across different industries brings up questions about their precise definition and the potential implications on safety assurance processes.

The concept of autonomy encompasses a wide spectrum of capabilities depending on the industry and area of application. Multiple taxonomies have been introduced to characterize a system's degree of autonomy, mainly depending on the division of tasks between the human and the system[7]. Each taxonomy relies on different criteria to determine which agent, human or autonomous, is responsible for the planning, decision-making, action implementation, and supervision tasks[8]. Some taxonomies use the terms automation and autonomy interchangeably, others offering a clear distinction between them. For instance, Sheridan and Verplank defined the Levels of Automation (LoA) on a scale from 1 to 10, ranging from complete manual control to fully autonomous control[9]. Future taxonomies evolved according to the needs of the different industries, across manufacturing, avionics, marine and land transportation, among others. More recently, the SAE J3016[10] categorized vehicle automation capabilities under six levels, from Level 0 to Level 5, depending on the combination of driving support and automated driving features.

Rather than present a precise definition of these systems, we explore the unique characteristics of autonomous systems and their challenges for safety assurance. Exploring these characteristics can provide guidance on whether additional safety goals and requirements are necessary for ensuring safety. Autonomous systems can perform functions that are not possible for traditional cyber-socio-technical systems, increasing the difficulty of transferring traditional functional safety concepts, metrics, and goals to systems in which autonomy plays an important role[11]. The internal processing and decision-making tasks can be highly inscrutable, increasing the difficulty for human designers, operators, and users to understand or explain the reasons for the system's

---

[6] This definition is based on NIST (2008), but adjusted for autonomous systems and operations, both manned and unmanned. This means that unmanned systems are a "subcategory" of autonomous systems.

[7] Vagia, M., Transeth, A.A. and Fjerdingen, S.A. (2016) 'A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed?', Applied Ergonomics, 53, pp. 190–202. Available at: https://doi.org/10.1016/j.apergo.2015.09.013.

[8] Parasuraman, R., Sheridan, T., Wickens, C., 2000. A model for types and levels of human interaction with automation. IEEE Trans. Syst. Man Cybern. Part A Syst. Humans 30.

[9] Sheridan TB, Verplank W. Human and Computer Control of Undersea Teleoperators. Cambridge, 1978.

[10] SAE International, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles", SAE Standard J3016, 2021.

[11] R. Adler and M. Klaes, "Assurance Cases as Foundation Stone for Auditing AI-Enabled and Autonomous Systems: Workshop Results and Political Recommendations for Action from the ExamAI Project", *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 13520 LNCS, pp. 283–300, 2022, doi: 10.1007/978-3-031-18158-0_21.

specific actions or decisions. This lack of interpretability or *explainability* implies that regular methods for verifying system functionalities and performance may not be sufficiently robust to demonstrate safety.

The potential scope of autonomous operations may be justification enough to demand new efforts to define and construct specific safety goals and criteria. For example, as many autonomous systems are envisioned to operate under a wide range of operational conditions, the specified "safe" behavior under nominal conditions may become unsafe under other unforeseen scenarios[12]. How can system designers demonstrate that the autonomous agent can determine what is the safest course of action under uncertain conditions?

[12] S. Ballingall, M. Sarvi, and P. Sweatman, "Safety assurance for automated driving systems that can adapt using machine learning: A qualitative interview study", J. Safety Res., vol. 84, pp. 243–250, 2023, doi: 10.1016/j.jsr.2022.10.024.

# Challenges in Autonomous System Safety Assurance

As industry seeks to increase the adoption of autonomous systems, persistent challenges continue to confound both industry and regulatory stakeholders. Compelling answers for whether current safety assurance processes are sufficient at governmental, media, and public level remain elusive. For example, when it comes to demonstrating the safety of a highly automated vehicle, is conducting a test drive spanning one billion miles efficient, necessary, or sufficient? In cases where traditional risk assessments or standard analyses fall short, what alternative approaches can we rely on? Under which circumstances do conventional risk assessment and safety assurance methods prove inadequate? Are there distinctive characteristics inherent to autonomous systems that require unique considerations? What specific certification requirements will regulatory authorities set for autonomous systems? And how will these requirements differ from those for complex systems involving multiple hardware, software, and human-interactable components? These are just a small subset of the questions that arise during discussions about the safety of autonomous systems.

The wide variety of autonomous system designs may prove a significant challenge to providing robust definitions and applicability criteria of safety standards at an industry level. For instance, an important aspect often overlooked in automated function and autonomous operation design is the role of human operators, crew, or users of the autonomous system. Depending on the Level of Automation, some autonomous systems are designed to interact with humans at all times; others might require human interventions under certain specific scenarios, while other systems are intended to operate with minimum human supervision. In this sense, autonomous systems can either extend, complement, or partially replace the human elements within complex system operation – but not necessarily replace it completely. This not only highlights the need to actively include human factors and human reliability analysis when discussing autonomous systems, but also the resulting system design, connectivity requirements, and overall safety goals.

The unique characteristics of autonomous systems and the scope of their operations can substantially increase the challenges to perform comprehensive and robust risk assessments and safety assurance certifications. Two aspects are particularly heavily discussed regarding autonomous systems: the open-world nature of their potential applications and the reliance on AI and data-driven Machine Learning (ML) techniques for safety-related decision-making processes. A profound discussion is needed to determine whether the current risk identification, modeling, and evaluation methods used are effectively able to adequately represent, simulate, and evaluate autonomous system functions and interactions. Similarly, there is a need for transparent dialogue around the current safety goals and metrics employed by industry to design and demonstrate system safety.

A major challenge in safety assurance of these types of complex systems is that depending on the use case, there may be no single safe state for the system to fallback to in the case of an emergency. Instead, the potential safest state would depend on the operational context and the state of the hardware, software, and human elements of the system. The safety goals of a system need to be connected to the specific operational conditions. For instance, the designed safety mechanisms may produce unintended consequences under environments in which these have not been adequately trained, tested, or validated, i.e., a reliable system is not necessarily a safe one. To reduce the inherent uncertainty of open-world operations, autonomous systems rely on both design-time and runtime safety mechanisms. From a design standpoint, this refers to the construction of safety envelopes constraining the operation of the system to known and validated safe states. In this context, the concept of an operational safety envelope is fundamental to constrain system operation under tested scenario conditions or environments that satisfy a specific safety goal. The safety envelope serves as a starting point to address the challenges of autonomous system operation in open, real-world environments with multiple hardware, software, and human subsystem interactions. During operation, the safety envelope includes safety measures (such as a safety system and/or human supervisor) designed to enforce the safety envelope and intervene in, restrict, or interrupt system operation to reach a safe state, either preventing or mitigating safety hazards.

However, restricting the system's operation under certain nominal operational conditions also emphasizes the need to have robust and effective safety mechanisms in place to act when the safety envelope is breached. In this regard, safety assurance of autonomous systems should not only be concerned with nominal system operation within established safety boundaries and the mechanisms that enforce it. Rather, to determine how safe a system is, we will need to know how the system behaves not only in normal operation, but also in a constrained, restricted, or failed state, what unintended errors and consequences may exist, and how to determine if the system's fail-safe or fail-functional states are safe enough. We may need more sophisticated methods for testing, simulating, verifying and validating both the system operation and the safety envelope enforcement mechanisms.

# The Safety Case for Autonomous Systems

Safety cases are a construct that serves as a framework combining claims, arguments, and the supporting evidence, justifications, and assumptions about the system's safety. They serve multiple purposes, such as safety certification and providing a structure for risk management or risk communication tasks. Yet, there are also multiple conflicting viewpoints on the use of safety cases. What is the exact purpose of a safety case, related to autonomous systems safety? How does it relate to safety standards? What are the roles of the respective regulatory authorities?

The goal of a safety case is to present a comprehensive defense of a system's safety. Though the concept has been described in several ways, it is formally defined as "*a structured argument, supported by a body of evidence that provides a compelling, comprehensible, and valid case that a system is safe for a given application in a given environment.*"[13] Safety cases should be clear, comprehensive, compelling, and defensible. Fundamentally, it is recognized that the safety case framework should include risk assessment and risk mitigation plan development[14]. In general, safety cases not only require developers to provide evidence on regulation compliance, but also on application-specific safety and risk targets. This aims to surpass traditional prescriptive approaches, as well as providing an alternative means to incorporate evidence which may not be compatible with classical risk assessment methods. While the concept of safety cases has been central to the regulation of multiple safety-critical systems, including nuclear, railway, oil and gas, automotive, industrial automation, and aerospace, compliance of industry-specific safety standards and best practices are significantly more widely adopted today. Recently, interest in using safety cases has increased given their potential to address the challenging safety assurance of autonomous and software-intensive systems[15].

The historical development of safety cases has been usually tied to severe industrial accidents since the 1960s. Safety cases were introduced as tools to comply with legislative modifications introduced to avoid future losses. Work on the conceptual basis of safety cases was formally established in the 1990s by Kelly, McDermid, Bishop & Bloomfield[16,17,18]. Structured approaches to develop and present safety arguments have

[13] Defence Standard 00-56 "Part 1: Safety Management Requirements for Defence Systems – Requirements", *(UK MoD Def Stan 00-56 Part 1, Issue 7, 2017).*

[14] T. Kelly, "Safety Cases", Handbook of Safety Principles, 10.1016/B978-1-4377-3524-6.00006-X, Ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2018, pp. 361–385. doi: 10.1002/9781119443070.ch16.

[15] T. Myklebust and T. Stålhane, "The Agile Safety Case", Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-319-70265-0.

[16] I. Habli, R. Alexander, and R. Hawkins, "Safety cases: an impending crisis", Safety-Critical Syst. Symp., 2020, [Online]. Available: https://eprints.whiterose.ac.uk/169183/

[17] R. Bloomfield and P. Bishop, "Safety and Assurance Cases: Past, Present and Possible Future – an Adelard Perspective", in Making Systems Safer, London: Springer London, 2010, pp. 51–67. doi: 10.1007/978-1-84996-086-1_4.

[18] T. P. Kelly, "Arguing safety-a systematic approach to safety case management", DPhil Thesis York University, Department of Computer Science Report YCST, 1999.

received significant attention from researchers. Currently, most safety cases are based on the use of two notations and their derivatives[19,20]: Claims, arguments, and evidence (CAE)[21,22] and Goal Structuring Notation (GSN)[23]. Additional details about these models can be found in the Appendix.

Safety cases are typically created by a team of engineers, scientists, and other experts providing system, software, human factors, risk, and standard compliance perspectives. Given the highly specialized nature of standard compliance, it is usually expected that the system's operators have their own internal safety management team or committee in charge of developing and overseeing the implementation of safety-related policies. The safety case is usually presented in the form of a report, describing the assumptions made about the system's functions and boundaries, the methods employed to assess risk, a justification of how the evidence was collected, and what deductions may be extracted from the evidence. The purpose of the report is to explicitly present the safety argument, i.e., demonstrate that the process or system meets the required regulations, the hazards have been comprehensively identified and mitigated, that key safety responsibilities have been defined, and that the level of residual risk is acceptable.

From a safety certification perspective, while regulators should determine what is required for safety assurance, the system's developers are responsible for providing the safety case. Classical methods employed in risk assessment and management provide sound support to develop defensible safety cases. Given the wide variety of industries, applications, and use cases involving complex systems, specific methods have been developed addressing system characteristics and hazards involved in their operation. Multiple verification and validation methods have been developed to address black-box behavior of software-intensive systems. However, in the case of autonomous systems, it is recognized that at the time we might not have the appropriate methods or expertise to create defensible safety arguments. In this regard, the safety case framework provides flexibility to include different sources of evidence as opposed to only compliance with industry-specific safety standards.

Proponents of safety cases argue that this approach to system safety can lead to significant benefits during system design and operation, including increased safety, improved risk management, a systematic method to record residual risk, and an effective risk communication tool to demonstrate compliance to regulatory requirements.

---

[19] T. Kelly, "Safety Cases", Handbook of Safety Principles, 10.1016/B978-1-4377-3524-6.00006-X, Ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2018, pp. 361–385. doi: 10.1002/9781119443070.ch16.

[20] V. Sklyar and V. Kharchenko, "Assurance Case For Safety And Security Implementation: A Survey Of Applications", Int. J. Comput., vol. 19, no. 4, pp. 610–619, Dec. 2020, doi: 10.47839/ijc.19.4.1995.

[21] R. Bloomfield and P. Bishop, "Safety and Assurance Cases: Past, Present and Possible Future – an Adelard Perspective", in Making Systems Safer, London: Springer London, 2010, pp. 51–67. doi: 10.1007/978-1-84996-086-1_4.

[22] R. E. Bloomfield and K. Netkachova, "Building Blocks for Assurance Cases", International Symposium on Software Reliability Engineering (ISSRE), 2014. [Online]. Available: http://openaccess.city.ac.uk/5121/

[23] The Assurance Case Working Group (ACWG), "Goal Structuring Notation Community Standard Version 3 The Assurance Case Working Group (ACWG)", 2021, [Online]. Available: https://scsc.uk/scsc-141C

However, it has also been recognized that there are many barriers to developing representative safety cases. Many open questions remain on how to address the safety case' complexity, the uncertainty, variability and reproducibility of safety claims and evidence, and the resources and expertise required to develop and assess the validity of the safety arguments. Currently, there are no prescribed specific methodological requirements to develop the safety arguments. The inherent flexibility and the lack of guidelines on what constitutes evidence, appropriateness of methods, and safety case upkeep are some of the main challenges from the perspective of both safety case developers and reviewers. There is a need for critical review of the current risk assessment tools, as well as quality and completeness criteria to be established so that the evidence and methods supporting safety cases, and risk assessments in general, are capable of adequately demonstrating a system's safety.

# Constructing Safety Cases

The safety case framework potentially offers a comprehensive and flexible means of demonstrating the safety of a system, employing the whole spectrum of techniques from quantitative risk assessment to sound engineering principles, and serving as an effective communication framework with regulatory bodies. Nonetheless, many issues regarding their construction and validation need to be addressed prior to playing a key role in autonomous system safety assurance processes.

## *Challenges and Differences in Industries*

One identified challenge is the absence of clear guidelines for formulating, structuring, and implementing safety cases. Some questions arise, such as: How should we incorporate additional evidence? When is it necessary to seek supplementary evidence? What methodologies should be considered appropriate in different contexts?

In the context of railways, the challenge appears more defined and structured, i.e., the traffic is tracked, often enclosed by physical barriers or tunnels, and includes traffic operators that oversee the traffic in the system. For autonomous road vehicles, the environment is considerably more unstructured. For systems with a high degree of human-system interaction, humans are sometimes modeled as simple input-output components. In the context of autonomy, perhaps the best examples are human-supervised autonomous systems, such as autonomous cars with safety drivers. Even in cases where the safety drivers are rigorously trained, a thorough Human Reliability Analysis (HRA) has been conducted, and measures taken to mitigate risks brought on by known human performance factors such as fatigue and distraction; humans can very much be considered a grey box, sometimes acting in unpredictable, undesirable, or even unexplainable ways and even being unable to reproduce the decision-making process after the fact. In autonomous systems containing AI/ML components, a parallel can be drawn. Despite the AI/ML being trained in a structured way, on a known curriculum, with a neural network designed to withstand known faults, these components can still act in unpredictable, undesirable, or even unexplainable ways and can therefore be considered grey boxes. Consequently, our approach should focus on identifying the distinctive characteristics inherent to autonomous systems. However, the question remains. How do we address the challenges that these grey boxes create?

In the maritime sector, it may be possible to constrain the operational scope to certain areas where we have a grasp of the potential unknowns. However, the issue is twofold: We lack a comprehensive understanding of all the system's challenges, and even if we did know all the challenges, we may not have the requisite methodologies to address them. Safety cases may offer a structured approach to tackling this issue. When complete safety assurance is unattainable, we can restrict a system's operation to within a defined safety envelope.

Another challenge is validation of self-learning autonomous systems. There are numerous studies that investigate the human learning process, and how cognitive

processes are affected by certain factors such as stress. In contrast, we currently lack a comprehensive understanding of AI/ML on system safety. One avenue could be to limit the degrees of freedom for learning-based systems by imposing rules and constraints. This approach mirrors the constraints we impose on human learning processes through training, procedures, and simulations. However, many of these autonomous systems will still interact with the open world, prompting the question whether limiting their scope or degrees of freedom is sufficient to convince stakeholders of their safety. In the case of autonomous cars, will using dedicated lanes or roads for autonomous vehicles help convince the relevant stakeholders that the vehicle is safe?

Embracing new methodologies or achieving broad adoption of safety cases poses a significant challenge in change management. Organizations accustomed to a particular approach need time to transition to a new one. In established industries with a well-established safety track record, there's potential for safety cases to serve as a versatile framework for integrating best practices and novel methods. Likewise, in emerging industries, the safety case can offer a flexible structure for incorporating best practices from established industries and seamlessly integrating them with innovative approaches.

## *Completeness and confidence in safety cases*

A critical aspect to address in complex, intelligent, and autonomous systems is how to assure the system's safety when operating under an open and dynamically changing environment. This implies that sufficient knowledge of the system's expected behavior and potential operational hazards should be demonstrated, as well as that the appropriate preventive and mitigative actions have been designed and implemented accordingly. Further, this would require that the development of the safety case be closely tied to the design process of the system.

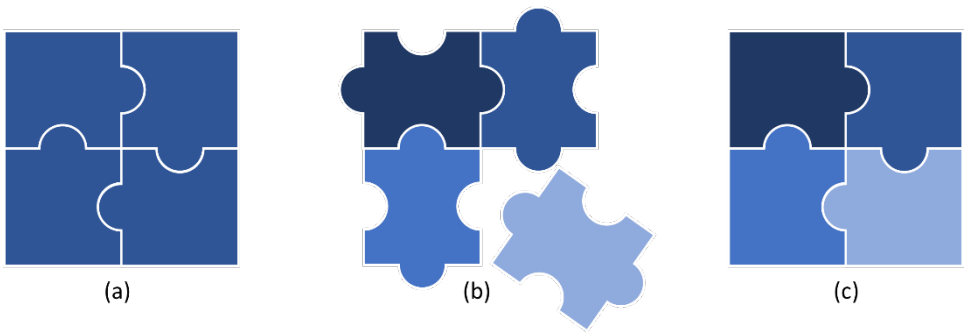



Figure 1: Safety case as a puzzle-analogy: (a) Safety case in adherence to a standard; (b) Safety case where the argument for the appropriateness of each element is addressed in the safety case; (c) Safety case where in addition to (b) the argument for the completeness of the safety case is addressed.

In some applications, it may be sufficient to demonstrate that a specific standard has been adhered to, in order to argue that the safety case is complete; however, in many cases, no such standard exists. In such cases, a potential approach to achieve safety case completeness is to rely on a diverse set of risk identification, modeling, and assessment methodologies applied at different stages of the safety case. However, there is a lack of

clear guidelines as to what kind of evidence, methods, and uncertainties can be incorporated into the safety case. Therefore, the appropriateness of the methods and levels of details required to build the safety cases need to be addressed to account for the complexity of autonomous systems. Guidelines for defining selection criterion accounting for method diversity, validity, advantages, and limitations must be defined depending on the specific industry and use case. Additionally, new approaches, methods, data collection, and validation practices may be required to address system safety during design time and runtime, respectively.

Another aspect that can contribute to the completeness of the safety case is the diversity of experts involved in the data collection, analysis, safety argument construction and validation. Consequently, when defining a team to develop a safety case, criteria for the members' diversity, competence, and qualifications should be established, including expertise on aspects such as methods employed, industry-specific requirements, and applicable areas of risk assessment (e.g., hardware, software, human, environmental). It may be that the required diversity of expertise and perspectives evolves throughout the safety case development process. For instance, the team could also include experts on the methods applied in the safety cases outside from the system's specific domain, acting as independent moderators of the safety assurance process.

The uncertainty of claims, assumptions, and evidence should be stated, disclosed, and quantitatively addressed in the safety case when possible. Both the known unknowns and unknown unknowns should be stated qualitatively and quantitatively if possible. It is important to address that uncertainty has many dimensions. On one hand, there is the strength of knowledge about the system's behavior and functionality given its evolving nature in a dynamic environment. On the other hand, there is the strength of knowledge about the safety case itself, i.e., how claims, arguments, and evidence are constructed and presented in a coherent and verifiable format. It is fundamental to understand the uncertainty of the system and communicate clearly what degree of uncertainty exists regarding safety claims. Knowledge and models for uncertainty related to the system's functional and operational behavior, the collected evidence, assumptions, arguments, and claims are required for decision-making at both design and runtime stages. The system model used in the design and construction of the safety case will never be a full representation of the system's operation in a real-world environment. Remaining uncertainty about operational environments for autonomous systems and the system behavior itself implies that field engineering feedback and lifecycle upgrade are essential.

### *Use of simulation data to account for real-world data limitations*

The use of simulation techniques, involving different combinations of synthetic data, environments, and system behavior, among other aspects, can play an important role in the development of autonomous systems. In the case of emerging technologies, the use of real-world data to test and validate system safety is highly desired. However, this data collection is frequently costly and time consuming. At the early stages of system

design, the rush to collect operational data may also raise safety concerns if the appropriate controls are not put in place to ensure sufficient system safety and maturity.

In the absence of real-world data to account for low frequency scenarios and emergency situations, simulation offers a starting point for system validation. Formal system verification and validation methods should continue to rely on real-world data. A common conception is that system development may initially rely on simulation and then transition to real-world data as it is collected at later stages of development, testing and operation. The role and relevance of simulated and real-world data is expected to vary from system to system.

The use of simulated data, environments, and system behaviors also raises several challenges regarding completeness, degree of fidelity to real-world conditions, and validity as evidence to build safety arguments. There is a strong need to determine what methods and tools are adequate to validate the use of synthetic data, particularly to claim completeness or sufficient coverage of high severity edge cases.

A potential solution to the issues arising from the use of simulated data is the use of standardized benchmark datasets to achieve more transparency and confidence in the development, validation, and testing process of autonomous systems. These benchmark datasets and associated performance metrics would cover the quality and variety of various sources and types of information, including selected scenarios, environmental sensor data, control parameters, and other input data relevant to test the system's safety. Following discussions should address which entities (industry, regulatory, etc.) are responsible for developing and validating representative benchmark datasets, as well as the use of standardized tests to assess the system's safety.

### *The use of AI/ML functions in autonomous systems*

Depending on the application, it is expected that many autonomous system functions will rely on ML and other AI techniques. In this regard, the specific functionality of AI/ML in the system should be clearly defined, stated, and addressed in the risk assessment. Two primary challenges are frequently discussed regarding the risk assessment of autonomous systems that incorporate AI/ML functionality:

- They may drift from their original operation over time, especially if self-learning techniques are used.
- They may be responsible for decision-making within the system, but these decisions may be unpredictable or unexplainable.

These issues raise additional questions regarding system verification and *explainability*, and whether the safety case approach can be used to counter these challenges.

In response to the first challenge, efforts should be directed to develop tools to facilitate the evaluation of the system's safety throughout its lifecycle. A baseline safety case should be developed for a fixed, specific version of the system's software (including

training data, initialization and model parameters, etc.). Future updates may be incorporated into the systems currently operating in the field only after extensive testing and change analysis, as changes may have invalidated the baseline safety case. Whether a safety case should be required for each software version update will likely depend on both the criticality of the system and the complexity of the system's operation.

Increasing the system's *explainability* may be a potential counter to the second challenge regarding system decision-making. However, there are challenges with *explainability*, especially for data-driven AI models. Current research suggests *explainability* is hampered by the complexity of the model. Systems employing deep learning are typically more accurate yet are also more difficult to explain[24]. There is an additional challenge with verification of the decision-making ability. It is possible that the system may encounter scenarios for which it is not equipped or trained properly to make the safe decision. In the maritime sector the Collision Regulations (COLREGS) provide rules for behavior when ships meet, but these are in some situations open to interpretation by the crew, e.g., the term *good seamanship*. Can you validate that an AI/ML algorithm displays *good seamanship*? Or is there a need for clear concise guidelines to aid in the development of safe autonomous vessels?

The limitations of ML methods should be clearly addressed and considered when constructing the safety case. In this regard, concerns of function dependence and unintended interactions should consider the particular characteristics of the employed ML methods, in addition to the concerns regarding training, testing, and validation datasets discussed previously. Further, there is a need to develop methods and metrics for runtime monitoring of AI/ML performance to address retraining and concept drift concerns. The idea of a negative safety case has been proposed and discussed for systems that use AI/ML as a method to avoid confirmation bias[25]. Instead of claiming system safety, a negative safety case argues and presents evidence for unsafe system operation. Negative safety cases could be compiled and presented by a third party for emerging technologies alongside the presentation of a traditional safety case.

## The role of regulatory authorities

The extent of the role of regulatory authorities regarding safety cases is currently debated across industries either searching to adopt safety cases as safety certification process or exploring methods to update their processes in accordance to increased system

[24] Saranya A., Subhashini R., A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends, Decision Analytics Journal, Volume 7, 2023

[25] N. Leveson, "The Use of Safety Cases in Certification and Regulation", ESD Work. Pap. Ser., no. November, pp. 1–12, 2011, [Online]. Available: https://dspace.mit.edu/bitstream/handle/1721.1/102833/esd-wp-2011-13.pdf

complexity. Hence, the nature of the relationship between industries and their corresponding regulatory authorities is also a topic which requires future discussions.

While the industry is responsible for demonstrating the system's safety, regulatory authorities are tasked with the responsibility of assessing the validity of the safety claims, for which they need access to information, tools, expertise, and experience. It is recognized that at the time there may be a lack of expertise from the regulatory environment tasked with verifying the validity of the presented safety cases. The evaluation of a safety case is a complex task that requires expertise in a range of disciplines, including engineering, risk analysis, and human factors. Therefore, while regulatory entities may define a high-level framework for the construction and update of the safety case, the appropriate level of detail and expertise required to define safety goals is unclear.

For certification purposes it may be feasible for regulatory authorities to determine a minimum safety benchmark. This could include the use of specific datasets, scenarios, validation metrics, and public tests. However, the use of standardized tests also raises concerns regarding the potential gap between test performance and real-world system performance. In the United States, one part of the regulative process is public hearings where interested parties and the public can present experts, evidence, and interests. This could be an interesting approach in other areas as well, especially for emerging technologies where regulators may lack the pre-requisite expertise but may need the opportunity to question relevant experts in the domain.

When it comes to autonomous systems across sectors and nations, regulatory approaches vary, and their impact can be significant. One aspect is the variation in the level of government regulation. In certain countries, automotive regulations are managed at the state level, whereas the aviation, rail, and maritime sectors are governed by national-level regulations. An example is California's early adaptation of autonomous vehicles regulations, which transformed the State into a hub for real-world testing of autonomous vehicles; and as a result, experienced several high-profile autonomous driving related accidents. Interestingly, social acceptance of accidents also differs across sectors, with automotive accidents often met with greater tolerance. Social acceptance often informs regulative safety acceptance criteria; thus, the question is: Should autonomous systems be held to a higher safety standard than the best-case scenario for human-operated counterparts? This remains an open question to regulators, however, in the maritime sector, autonomous vessels must demonstrate a safety performance at least as good as their non-autonomous counterparts. Without concrete definitions of autonomous systems safety expectations, the challenge to develop appropriate regulations may surpass the technical challenge to demonstrate system safety.

It is evident that the social acceptance of autonomous systems and, consequently, the regulations, may change following accidents involving these systems. To address this evolving landscape, regulators should proactively establish and standardize the collection of "standard data" from autonomous systems during their operation. This data can then be used as a mechanism whereby regulators can improve safety acceptance criteria

continuously over time and not just in incremental steps posterior to an accident and subsequent investigation. Today it is a challenge for investigators to get sufficient data from the manufacturers, and even to determine whether an autonomous vehicle was operating autonomously or under human control at the time of an accident. One could also ask, for emerging technologies, is it not in the manufacturers' common interest to collaborate to demonstrate the safety of the technology? Is there a role for regulators in facilitating such cooperation? Two examples of this approach are:

- In the United States, aviation carriers willingly provide and exchange safety-related data through a third-party entity. This data is anonymized, and regulators have explicitly agreed not to attempt to identify specific companies or enforce regulations based on this information.
- In the Norwegian offshore oil and gas sector, operators, unions, and the authorities have established collaboration efforts to monitor the risk level in the industry and enhance safety in operation (the RNNP project).

In an ever more interconnected global landscape, there is potential for regulatory collaboration spanning international borders. This becomes especially relevant dealing with emerging technologies that lack established regulatory structures and procedures. An example of this approach is the maritime industry's Port State Control regime. The objective for each Port State under the Paris MoU is to control 25% of the foreign flagged ships calling at their ports annually. Based on the inspection findings for the last 3 years, the Paris MOU categorizes the performance of the Flag states into white-, grey- and blacklists, representing quality of flags from high performance to flags with poor performance considered high or very high risk.

## The use of safety cases during system runtime

Safety cases are usually presented within the safety certification process for complex systems. However, safety case development should begin during the early system design stages and not only be incorporated into later stages of implementation and certification. The early integration of safety cases requires the use of integrated tools to keep track of assumptions, evidence, and safety arguments while the system evolves during design and implementation stages. Two key roles for safety cases during system runtime are identified.

The first refers to the potential use of an approved safety case as an online risk assessment tool to monitor the system's operation. It may be possible to monitor the validity of evidence, assumptions, and safety claims based on the current system operation. Potential methods to track the validity of assumptions and safety arguments are through signal-, data-, and model-based approaches. For this purpose, safety performance metrics used to inform about the system's operation must be tied to evidence, assumptions, and arguments used in the safety case. In particular, special emphasis should be placed on properly formulating assumptions so they can be monitored through collected evidence. Further, the concepts of Operational Design

Domain (ODD) or Concept of Operations (CONOPS) can play an important role in setting, tracking, evaluating, and enforcing the operational safety envelope of the system.

The second role refers to the lifecycle of safety cases. From a risk management side, approaches are needed to frequently incorporate new data collected and experience gained during operation in an integrated and optimized way. The criteria for the construction, update and/or upgrade process of safety cases should be defined by the appropriate regulator depending on the domain. These criteria should include the scope and frequency of the safety case modifications, as well as requirements regarding incident investigation and the incorporation of lessons learned. Hence, there is a need to establish data exchange policies between industry and regulatory entities so safety case modification criteria can be founded on changes in system, changes in operational use, unplanned modifications, and extraordinary circumstances.

# Concluding Remarks

## Should Safety Cases be more than a regulative exercise?

This question assumes greater urgency as an increasing number of industries embrace autonomous systems, sparking a renewed interest in the use of safety cases. Regulatory bodies, which have previously been more proactive in certain sectors, currently appear to have assumed a more passive stance, leaving ambiguity regarding the safety goals and principles that will be enforced. The need for well-defined safety criteria is clear.

If we look at the railway and automotive sector, they both have established standards that encompass safety cases, however, only the railway sector relies on clearly defined safety criteria. On the other hand, there are strict risk acceptance criteria in the railway sector, which may prove an obstacle for new technology where the risk and uncertainties are novel. While conventional software can be rigorously verified, the same cannot be said for artificial intelligence and machine learning systems and their interfaces with the broader system, making them focal points in the safety case process.

Good assumptions are a prerequisite for constructing defensible safety cases, and the set of all assumptions plays a pivotal role in delineating the scope of the system's applicability. The emergence of AI and ML-based systems introduces a new layer of assumptions, and it may be the case that we need to include mechanisms to halt system operations if the assumptions are invalidated during runtime. A frequent drawback of current safety case frameworks is the limited incorporation of human errors and organizational factors. While the UK has a strong tradition of task analysis and human reliability within the nuclear and railway sectors, the treatment of human reliability and human factors varies across industries, and in many sectors, there exists a regulatory void concerning human errors and performance. This analysis of human reliability and human factors may become more relevant if regulatory entities are tasked to determine culpability for breaches in the safety case.

## Final Message & Future Directions

Conscious efforts must be made to reach a representative definition of what kinds of autonomous functions and operations exist in autonomous systems, what is the role of the human operator, crew, or user involved in the system's operation, what is implies in terms of safety and safety assessment, among multiple other topics. Throughout decades, the risk communities within different industries have reached sufficient consensus to enable the use of risk assessments - "agreements" that are also required on what constitutes a safe autonomous system. The main discussion points and open questions can be summarized as the following:

- How to determine appropriate safety goals for autonomous systems, which are needed in a safety case approach? From a regulatory and legal perspective, it is perceived that more discussions involving societal risk tolerance are required. In particular, regarding technology that will potentially interact with greater portions of society, such as critical infrastructure and autonomous transportation systems.

- How to establish clear methodologies for demonstrating system safety and safety assurance for autonomous systems? The lack of clear guidelines for formulating, structuring, and implementing safety cases for these emerging technologies remain a significant setback for regulatory authorities to assess the benefits of this approach, outside specific areas of industry whose safety certification is tied to the development of safety cases.

- How do we train the next generation of engineers and analysts to address safety assurance of autonomous systems? Questions regarding how to establish sufficient expertise both to develop and assess autonomous system safety remain. As in the case of other emerging technologies, a challenge is to train analysts that can provide independent assessments on system safety, safety case validity, and contribute to societal risk acceptance and goals.

- How to address the challenges of open-world autonomous system operation? Two different principles arise, either increase the system's functionality (and complexity) or actively seek to reduce their operational environment complexity. While the first approach may be technically feasible, the latter may be a more viable approach to safe systems and operation.

- How to address the inscrutability of autonomous systems' functions? When discussing methodologies to assess the black box behavior of autonomous systems' functions and operations, it is important to consider the interpretability and explainability of the outputs. If interpretable models cannot be used, explainable AI (XAI) techniques should be employed to present understandable explanations for system behavior.

# Organizing Committee

**Christoph A. Thieme, PhD** – SINTEF

Dr. Christoph Thieme is a researcher at SINTEF Digital in Trondheim, Norway, where he applies his knowledge to different research projects related to safety and security of socio-technical systems. He obtained his PhD in Marine Technology from NTNU, with specialization in safety, reliability, and risk assessment for autonomous systems. He has experience with risk assessment of autonomous systems with a focus on software safety and human-machine interaction. Additionally, he is a visiting professor at the University of Toulon lecturing on Risk and Reliability engineering and potential application of AI methods, building on his research and the insights gained at the IWASS workshops within safety, reliability, and security for autonomous systems.

christophthieme.com christoph.thieme@sintef.no

**Marilia Ramos, PhD** - UCLA

Dr. Marilia Ramos is a Research Affiliate at the B. John Garrick Institute for the Risk Sciences, UCLA, and a sessional instructor of Human Reliability Analysis at UCLA. She holds a PhD in Chemical Engineering from the Federal University of Pernambuco, Brazil. She has expertise in Risk Analysis and Human Reliability with extensive experience leading and collaborating with projects for different industries, from the nuclear industry to autonomous cars and wildfire evacuation. She is currently a Business and Research Development Officer at the University of Toronto, where she forges collaborations between academia, industry, not-for-profit, and governmental agencies around cutting-edge research topics.

mariliaramos.net marilia@risksciences.ucla.edu

## **Andrey Morozov, Dr.-Ing** - University of Stuttgart

Andrey Morozov received his diploma in Computer Science and Mathematics from Ufa State Aviation Technical University in 2007 in Ufa, Russia. In 2009 he moved to Germany and, in 2012, got a doctoral degree (Dr.-Ing.) at the Institute of Automation (IfA) of Technische Universität Dresden. After that, being a postdoc researcher, he worked on several R&D projects funded by DLR, ESA, NASA, and DFG. In 2014, Jun.-Prof. Morozov built a research group at IfA focusing on the model-based analysis of safety-critical mechatronic systems. Since 2020, Andrey has been a tenure-track professor for Networked Automation Systems at the Institute of Industrial Automation and Software Engineering of the University of Stuttgart. The research interests of Andrey lie at the intersection of three domains, namely, (i) Cyber-Physical Systems, (ii) Dependability and Risk Analysis, and (iii) Artificial Intelligence (AI).

ias.uni-stuttgart.de/institut/team/Morozov/ andrey.morozov@ias.uni-stuttgart.de

## **Ingrid B. Utne, PhD** – NTNU

Dr. Ingrid Bouwer Utne is a Professor at the Department of Marine Technology, NTNU. Her research is focused on risk assessment and modeling of complex and autonomous marine systems and operations. She is an affiliated Researcher in the Center of Excellence on Autonomous Marine Operations and Systems (NTNU AMOS), and she is a principal investigator and manager of several research/industry projects. One of her main contributions in recent years is supervisory risk control bridging the scientific disciplines of risk management and engineering cybernetics advancing 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.

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

## **Camila Correa-Jullian, M.S.** - UCLA

Camila is a PhD candidate of the Mechanical and Aerospace Engineering department at the University of California, Los Angeles. She obtained her MS in Reliability Engineering at the University of Maryland and her BS in Mechanical Engineering at the University of Chile. Her current research at the B. John Garrick Institute for the Risk Sciences is focused on characterizing, modeling, and simulating operational safety risks of human-system interactions in automated systems and Connected and Automated Vehicles (CAV).

camcorreajullian.github.io/ ccorreaj@ucla.edu

## **Spencer Dugan** - NTNU

Spencer Dugan is a PhD candidate at the Department of Marine Technology, NTNU. He holds a BSc in Naval Architecture and Marine Engineering from Webb Institute, an MSc in Maritime Technology from NTNU, and an MEng. in Mechanical Engineering from DTU. His research is on the design and operation of propulsion systems for autonomous ships.

ntnu.no/ansatte/spencer.a.dugan spencer.a.dugan@ntnu.no

## **Joachim Grimstad, M.Sc.** - University of Stuttgart

Joachim is a PhD candidate at the Institute of Industrial Automation and Software Engineering, University of Stuttgart. He holds a Bachelor of Engineering in Subsea Technology - Maintenance and Operations from the Western Norway University of applied Sciences and a Master of Science in Engineering in Reliability, Availability, Maintainability, and Safety (RAMS) from the Norwegian University of Science and Technology. His current research topic is Model-Based Systems Engineering (MBSE) and the use of Artificial Intelligence (AI) techniques in MBSE.

ias.uni-stuttgart.de/en/institute/team/Grimstad/ joachim.grimstad@ias.uni-stuttgart.de

# Organizers & 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 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) improve 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).

Universität Stuttgart

**Institute of Industrial Automation and Software Engineering, University of Stuttgart, Germany**

The Institute of Industrial Automation and Software Engineering looks back at over 80 years of tradition at the University of Stuttgart. We see ourselves as the think-tank, bridge builder and integration hub of a creative environment in the heart of the industrial metropolis Stuttgart. Currently our R&D interest lies at the intersection of three domains, namely, (i) Networked Robotic Systems, (ii) Dependability, and (iii) Artificial Intelligence (AI). Accurate assessment of risk, reliability, safety, and resilience is essential for modern technical systems because of the high cost of downtime and strict safety requirements. However, the analytical capabilities of risk evaluation methods, which are currently applied in the industry, are far behind the technical level of the systems in question. These methods cannot adequately describe sophisticated failure scenarios of highly dynamic and intelligent systems. Besides that, future robotic systems will include more and more AI components. However, the reliability and safety analysis of AI is an entirely open question at the moment. An inevitable revolution in the risk methods is expected in the next few years. So, the main goal is to build a strong research team capable of taking a leading role in the development of the next generation of risk analysis methods for modern and future robotic systems.

**DNV**

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. 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).

# Acknowledgements

The IWASS organizing committee would like to thank the organizing and event committees from the European Conference on Safety and Reliability (ESREL) and the Faculty of Social Sciences in the University of Southampton for the support provided for the development of the 2023 workshop.

**European Conference on Safety and Reliability (ESREL) & European Safety and Reliability Association (ESRA)**

ESREL (European Safety and Reliability) is an annual conference series run under the auspices of the European Safety and Reliability Association (ESRA). The conference has become well established in the international community, attracting a good mix of academic and industry participants that present and discuss subjects of interest and application across various industries.

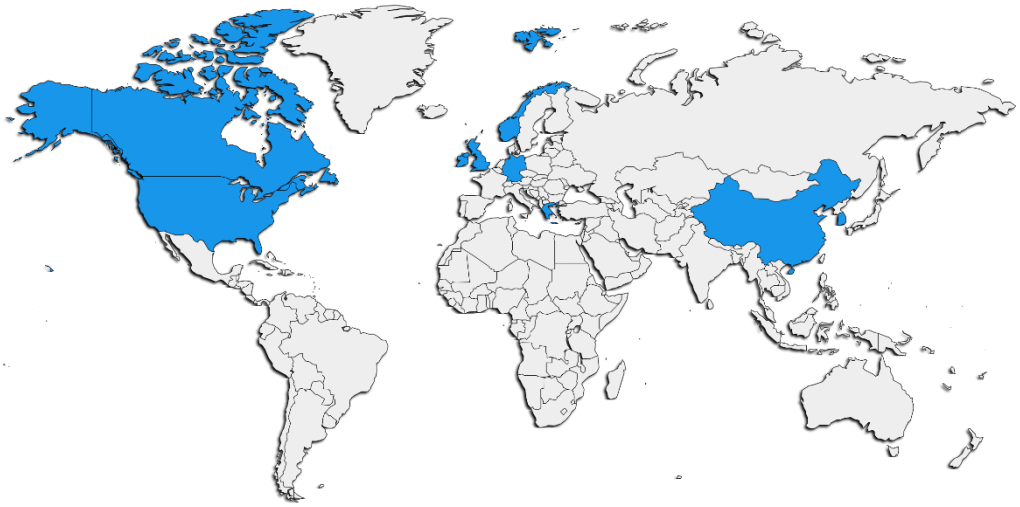
In 2023 the theme of the conference is "The Future of Safety in a Re-Connected World". The conference covers several topics within safety, risk and reliability analysis methods, maintenance, optimization, and risk management. Special focus has been placed on how technological developments and nature induced hazards impact societal safety in a world increasingly reconnected post Covid-19 pandemic.

**Faculty of Social Sciences in the University of Southampton**

The University of Southampton is a public research university located in Southampton, England. It was founded in 1862 as the Hartley Institution, and it received its royal charter in 1952. The university has nine faculties and over 25,000 students.

The Faculty of Social Sciences is one of the largest faculties at the University of Southampton. It offers a wide range of undergraduate and postgraduate courses in social sciences, including anthropology, criminology, economics, geography, politics, psychology, and sociology. The faculty is also home to a number of research centers and institutes, including the Centre for Operational Research, Management Science and Information Systems (CORMSIS) and the Centre for Risk Research (CRR).

# IWASS Participants



## Organizing Committee

| Name | Affiliation |
|---|---|
| **Ali Mosleh** | University of California Los Angeles (UCLA) - USA |
| **Andrey Morozov** | University of Stuttgart - Germany |
| **Camila Correa-Jullian** | University of California Los Angeles (UCLA) - USA |
| **Christoph Thieme** | SINTEF Digital - Norway |
| **Ingrid Bouwer Utne** | Norwegian University of Science and Technology (NTNU) - Norway |
| **Joachim Grimstad** | University of Stuttgart - Germany |
| **Marilia Ramos** | University of California Los Angeles (UCLA) – USA |
| **Spencer Dugan** | Norwegian University of Science and Technology (NTNU) - Norway |

## Keynote Presenters

| Name | Affiliation |
|---|---|
| **Mollie D'Agostino** | University of California Davis, Institute of Transportation Studies - USA |
| **Rasmus Adler** | Fraunhofer Institute for Experimental Software Engineering IESE - Germany |
| **Thor Myklebust** | SINTEF Digital - Norway |

## Participants in Southampton

| Name | Affiliation |
|---|---|
| **Arne Ulrik Bindingsbø** | Equinor - Norway |
| **Berit Schürrle** | University of Stuttgart - Germany |
| **Børge Kjeldstad** | Maritime Robotics/ Norwegian University of Science and Technology (NTNU) - Norway |
| **Chanjei Vasanthan** | DNV- Norway |
| **Chris Harrison** | Rail Safety and Standards Board - UK |
| **Dirk Söffker** | University of Duisburg-Essen - Germany |
| **Hector Diego Estrada Lugo** | TU Dublin - Ireland |
| **Hyungju Kim** | University of South-Eastern Norway - Norway |
| **Jana Price** | National Transportation Safety Board - USA |
| **Jarle Fosen** | Gard - Norway |
| **Jon Arne Glomsrud** | DNV - Norway |
| **Manuelis Annetis** | National Technical University of Athens, School of Naval Architecture and Marine Engineering, Maritime Risk Group - Greece |
| **Marie Farrell** | University of Manchester - UK |
| **Mirko Conrad** | Samoconsult GmbH / TU Dresden - Germany |
| **Ørnulf Jan Rødseth** | MITS Consult - Norway |
| **Salvatore Massaiu** | Institute for Energy Technology, IFE - Norway |
| **Silvia Tolo** | University of Nottingham- UK |
| **Silvia Vock** | German Federal Institute for Occupational Safety and Health (BAuA) - Germany |
| **Sizarta Sarshar** | Institute for Energy Technology, IFE - Norway |
| **Zhang Di** | National Engineering Research Center for Water Transport Safety Director, ITS Research Center, Wuhan University of Technology - China |

# Online Participants

| Name | Affiliation |
|---|---|
| **Niav Hughes Green** | Nuclear Regulatory Commission (NRC) - USA |
| **Philip Koopman** | Carnegie Mellon University - USA |
| **Ruochen Yang** | University of Maryland - USA |
| **Seojeong Lee** | Korea Maritime and Ocean University - Republic of Korea |
| **Sergio Guarro** | ASCA Inc.- USA |
| **Tingting Cheng** | Norwegian University of Science and Technology (NTNU) - Norway |
| **Tunc Aldemir** | Ohio State University - USA |
| **Yan-Fu Li** | Tsinghua University - China |

# Appendix

IWASS 2023 International Workshop on Autonomous Systems Safety

# The Safety Case for Autonomous Systems: An Overview

UCLA ENGINEERING
B. John Garrick Institute for the Risk Sciences

NTNU – Trondheim
Norwegian University of Science and Technology

Universität Stuttgart

# The Safety Case for Autonomous Systems: An Overview

***White paper for the 4th International Workshop on Autonomous System Safety (IWASS 2023)***


**Authors:**

Camila Correa-Jullian, Joachim Grimstad, Spencer August Dugan

**Edited by:**

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



August 2023



UCLA ENGINEERING
B. John Garrick Institute for the Risk Sciences

NTNU – Trondheim
Norwegian University of Science and Technology

Universität Stuttgart

# Summary

The safety assurance of complex systems is an ongoing challenge for both system developers and regulator entities. A common path to demonstrate system safety is through the construction and presentation of safety cases. In general, safety cases not only require developers to provide evidence on regulation compliance, but also on application-specific safety and risk targets. Challenges to develop efficient safety cases to monitor the system's safety throughout its lifecycle are highlighted by the increasing adoption of autonomy and automation technologies in industry.

This whitepaper aims to provide a common understanding of safety cases and their use in industry ahead of the discussions at the 4th International Workshop on Autonomous System Safety (IWASS). IWASS 2023 discussions will focus on addressing the challenges of providing thorough and credible safety assurance of complex systems as automation capabilities increase across multiple industries.

# Table of Contents

# The International Workshop on Autonomous System Safety

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-GIRS), the Norwegian University of Science and Technology (NTNU) and the University of Stuttgart.

IWASS is an invitation-only event designed to be a platform for **cross-industrial** and **interdisciplinary effort** and **knowledge exchange** on autonomous systems' Safety, Reliability, and Security (SRS). The workshop gathers experts from academia, regulatory agencies, and industry to discuss challenges and potential solutions for SRS of autonomous systems from **different perspectives**. It complements existing events organized around specific types of autonomous systems (e.g., cars, ships, aviation) or the safety or security-related aspects of such systems (e.g., cyber risk, software reliability). IWASS distinguishes itself from these events by addressing these topics together and proposing solutions for SRS challenges common to different types of autonomous systems.

IWASS previous editions (*2019* - Trondheim/Norway; *2021* - online; *2022* - Dublin/Ireland) successfully assembled a broad and diverse field of experts from different organizations and countries. **IWASS proceedings summarize the discussions held during the events and provide a strong foundation concerning autonomous systems SRS, ranging from risk analysis methods, and cascading failures to "human on the loop" and regulations: 2019[1], 2021[2], 2022[3].**

IWASS 2023[4] will take place in Southampton, United Kingdom, on September 2nd and 3rd.

[1] Proceedings to the 1st International Workshop on Autonomous Systems Safety. Trondheim – Norway, 11-13 March 2019. https://bit.ly/2SsPrLd

[2] Proceedings to the International Workshop on Autonomous Systems Safety 2021. 20, 21 and 28 March 2021. https://www.risksciences.ucla.edu/iwass-2021-proceedings

[3] Proceedings to the International Workshop on Autonomous Systems Safety 2022. Dublin - Ireland, 28 March 2023. https://www.risksciences.ucla.edu/iwass-2022-proceedings

[4] International Workshop on Autonomous Systems Safety 2023. https://www.risksciences.ucla.edu/iwass-2023-home


IWASS
International Workshop on Autonomous Systems Safety
Proceedings
August 28, 2022
Dublin | Ireland
IWASS
International Workshop on Autonomous Systems Safety
Proceedings
April 20, 21 and 28, 2021
Online Event
IWASS
First International Workshop on Autonomous Systems Safety
Proceedings
11th to 13th March, 2019
Trondheim | Norway

## Scope and goal of the white paper

The projected increase of autonomous and automated systems across multiple safety-critical applications raises questions about how developers, operators, and regulators address the safety of these systems' operations. **The concept of safety cases has been central to the regulation of multiple safety-critical systems**, including nuclear, railway, oil and gas, automotive, industrial automation, and aerospace.

In these systems, software plays a key role in assuring the safety-relevant behavior of key components and subsystems. However, many challenges remain to address safety assurance of software based on machine learning (ML) and other artificial intelligence (AI) methods, from both the developers and the regulators' perspective. Their black-box nature and limited interpretability makes it difficult to guarantee that, given a specific context, sufficient assumptions about the data and calculation functions are met such that the safety-related tasks are performed as intended. For instance, it may be possible that the intended safety-related behavior results in unsafe actions given unforeseen edge cases [1]. In particular, when dealing with emergent behavior exposed during system operation and human-machine interaction. Issues of state-space explosion, robustness, system integration, adversarial attacks, as well as setting the requirements and test specifications have also been identified as crucial challenges in the verification and validation of AI/ML-based software systems [2]. **Given the wide range of issues, a multidisciplinary, risk-based approach is required for the development of comprehensive safety assurance tools for autonomous systems.** To date, efforts to address these challenges have led to the development of standards and technical reports focused on AI/ML applications. These include the ISO/IEC TR 29119-11[5] regarding the testing of AI systems and the ISO/IEC TR 24028[6] related to AI system trustworthiness and assessments, among other efforts. Currently, the ISO/IEC TR 5469[7] is under development, aiming to address the functional safety related to AI systems.

This white paper overviews the fundamental concepts concerning safety cases: background and history, elements, and how they are constructed and used in different industries. The whitepaper aims to provide a common understanding ahead of the discussions at the 4th International Workshop on Autonomous System Safety (IWASS). IWASS 2023 discussions will focus on addressing the challenges of providing thorough and credible safety assurance of complex systems as automation capabilities increase across multiple industries.

[5] ISO/IEC TR 29119-11:2020 Software and systems engineering – Software testing – Part 11: Guidelines on testing of AI-based systems.
[6] ISO/IEC TR 24028:2020 Information technology – Artificial Intelligence – Overview of trustworthiness in artificial intelligence.
[7] ISO/IEC DTR 5469 Artificial Intelligence – Functional Safety and AI Systems (in development)

In this whitepaper, we briefly recapitulate the history, motivation, and structure of the safety case. We touch upon different methodologies and tools used to develop and communicate them. We finally summarize challenges of applying safety cases identified in recent literature. **At IWASS 2023, we will explore, together with the workshop participants, what academia, different industries, applied researchers, and policymakers can input to the discussion surrounding the use of safety cases, methods involved in their development, and their future as credible safety assurance frameworks.**

# The safety case

The regulation of system safety has traditionally been based on two different approaches. The prescriptive approach is common in many industries and relies on compliance-based certification of highly specific safety standards, such as IEC 61508[8], IEC 61511[9], or ISO 26262[10]. These standards cover the design, implementation, maintenance, and management of systems during their entire life cycle. An alternative is performance-based approaches, such as safety cases. This performance-based approach relies on certification authorities specifying the threshold of acceptable system performance, usually an acceptable risk target. It is then the role of the system's designers, managers, or operators to provide the necessary evidence to assure the system's safety requirements are achieved, independently from the methods employed to do so [3]. Performance-based approaches still require compliance with applicable standards that are required by the corresponding regulatory authorities [4]. Currently, standards may also require compliance with particular risk thresholds. Yet, safety cases require evidence that a thorough and systematic process to assess and control risks associated with the system has been adopted, going beyond a reactive, standards-based approach to safety management [5].

**The aim of a safety case is to present a structured argument and the corresponding evidence that a system can operate safely for a given context.** Though the concept has been described in several ways, it is formally defined as "*a structured argument, supported by a body of evidence that provides a compelling, comprehensible, and valid case that a system is safe for a given application in a given environment.*" [6]. From a risk perspective, a safety case aims to provide evidence demonstrating that a system's risks have exhaustively been identified, assessed, and are being managed accordingly. That is, appropriate risk controls have been implemented, and their effectiveness is monitored and assessed throughout the system's life cycle to ensure that residual risk remains within acceptable levels.

Safety cases and their variations have been produced, reviewed, and researched for several decades [7]. Their adoption in regulation has become more widespread, coupled in part with the increase in complexity in industries where safety is critical, such as aviation, rail, automotive, oil and gas, industrial robotics, and nuclear power. Safety

[8] IEC 61508:2010 Functional safety of electrical/electronic/programmable electronic safety-related systems.
[9] IEC 61511: 2018 Functional Safety-Safety Instrumented Systems for the Process Industry Sector.
[10] ISO 26262:2018 Road vehicles – Functional Safety.

cases may be derived at the component, system, process, or network level or can focus specifically on events or procedures [8].

## Background and history

The historical development of safety cases has been usually tied to severe industrial accidents since the 1960s. **Safety cases were introduced as tools to comply with legislative modifications introduced to avoid future losses**. Work on the conceptual basis of safety cases was formally established in the 1990s by Kelly, McDermid, Bishop & Bloomfield [9–11]. In the past decades, academic research on safety cases has focused on developing improved notations, integration, and evaluation methods, as well as model-based arguments, and approaches to automate their development and interpretation. Several studies also focus on estimating and propagating the safety argument's uncertainties to obtain the overall confidence on the safety cases' claims. The concept of safety cases was extended to *assurance cases* to address cybersecurity concerns (i.e., security cases) rising through digital infrastructure. Currently, developing safety cases plays an important role in regulations and safety standards across multiple industries and countries.

**Examples of safety-case approaches can be found across multiple industries, regulatory bodies, and countries.** In the United Kingdom, the Aircraft and Armament Evaluation Establishment (A&AEE) requires the development of safety cases for the operation of nuclear, chemical, rail transport, petrochemical, and defense systems. In the EU, the European Organisation for the Safety of Air Navigation (Eurocontrol) developed a Safety Case Development Manual to oversee the civil aviation industry. In the US, the traditional safety driving forces have been the nuclear and space programs, through the Nuclear Regulatory Commission (NRC) and the National Aeronautics and Space Administration (NASA), respectively. While not explicitly named safety cases, NASA has implemented the use of Safety Analysis Reports (SAR) and Mission Safety Evaluations (MSE) for their operations. Similarly, the Federal Aviation Administration (FAA) developed the Aviation Reporting System (ASRS) to support accident investigation and system improvement. Both the US Occupational Health and Safety Agency (OSHA) and the Food and Drug Administration (FDA) have incorporated similar performance-based approaches in specific instances.

## The elements of a safety case

Safety cases are typically created by a team of engineers, scientists, and other experts providing system, software, human factors, risk, and standard compliance perspectives. Given the highly specialized nature of standard compliance, it is usually expected that the system's operators have their own internal safety management team or committee in charge of developing and overseeing the implementation of safety-related

policies. **Safety cases should be clear, comprehensive, compelling, and defensible** [4].

In general, the development of safety cases follows traditional risk assessment framework's structure to present a risk-based argument [4]. This process usually consists of the steps depicted in Figure 1. These consist of identifying the hazards present in the system, assessing their risk, identifying potential risk mitigation measures, and reducing it to an acceptable level. This is followed by verifying the risk has been reduced and that residual risks are acceptable. Finally, this process also may provide means to track the risk throughout the system's life cycle [8].

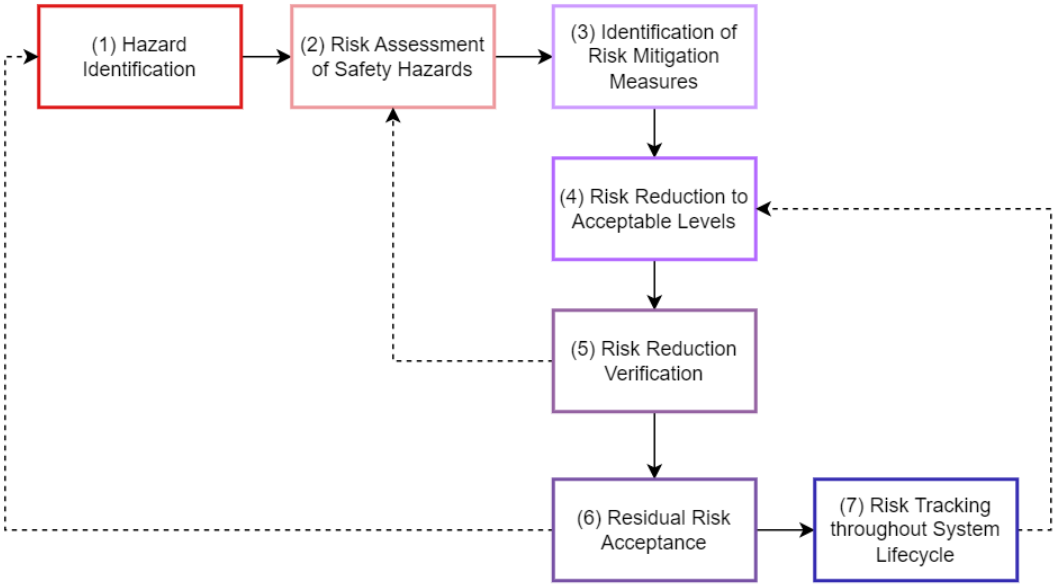



*Figure 1: General Risk Assessment Framework for Safety Case Argument Development.*

The contents of a safety case vary depending on the specific system and the industry in which it is being used. However, as a result of the risk assessment-based procedure, most safety cases include the following elements:

- A description of the system and safety boundaries.
- A description of the operational environment, context, and conditions.
- A description of the hazards and risks associated with the system, and how these have been identified and assessed.
- A description of the controls and mitigations in place to reduce the identified risks.
- A description of the evidence that supports the safety of the system.
- A justification of the acceptability of the residual risk.

The safety case is usually presented in the form of a report, describing the assumptions made about the system's functions and boundaries, the methods employed

to assess risk, a justification of how the evidence was collected, and what deductions may be extracted from the evidence. The purpose of the report is to explicitly present the safety argument, i.e., demonstrate that the process or system meets the required regulations, the hazards have been comprehensively identified and mitigated, that key safety responsibilities have been defined, and that the level of residual risk is acceptable.

However, the use of natural language to present safety cases can lack adequate clarity and structure and can be difficult to comprehend. Thus, structured approaches focused on the development and presentation of safety arguments have received significant attention from researchers. Currently, most safety cases are based on the use of two notations and their derivatives [4,12]: Claims, arguments, and evidence (CAE) [10,13] and Goal Structuring Notation (GSN) [14], both based on classical set theory, graph theory and relation algebra [15].

The CAE notation is built on block structures as shown in Figure 2. These blocks consist of three basic elements:

- **Claims:** Statements about a systems or sub-system's properties to be demonstrated through safety arguments and evidence. Claims may be hierarchically constructed through sub-claims, reaching a level of decomposition until assumptions (claims asserted without justification) are explicitly identified.

- **Arguments:** Statements that link the evidence to the safety claim. Arguments are built using inference rules and argue the trustworthiness of the evidence's implications, as well as the scientific or engineering laws used. Bloomfield and Netkachova [13] defined five basic building blocks representing types of arguments. These are: decomposition, substitution, concretion, calculation/proof, and evidence incorporation. Arguments are supported by different side warrants depending on the type of argument used. These blocks may be combined depending on the safety argument.

- **Evidence:** A documented basis for the safety argumentation or justification of the claim. Sources of evidence may include the design, the development process, prior field experience, testing, source code analysis, or formal analyses, which demonstrate the achievement or non-achievement of safety-related goals. Industry-specific safety performance indicators (SPI) may be used to track and present evidence of the system's safety.

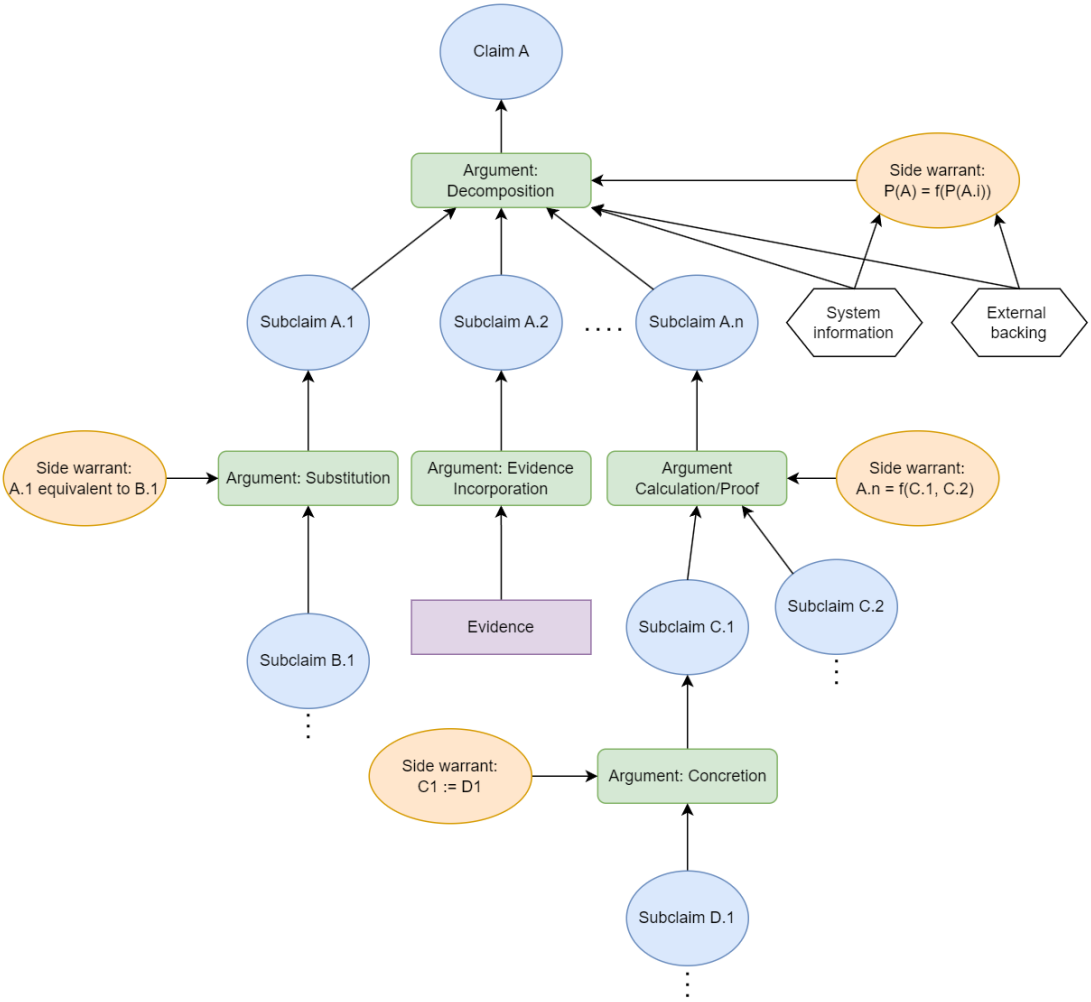



*Figure 2: Example CAE block structure. Adapted from Bloomfield and Netkachova [13].*

GSN is a graphical tool used to build structurally cohesive arguments based on elements analogous to the CAE notation. GSN operates through goals (analogous to claims), an argumentation strategy, and a solution (analogous to evidence) based on assumptions about the system. Goals may also be decomposed into sub-goals to be hierarchically organized, as in the case of CAE [12]. However, GSN also introduces contextual elements to set different goals depending on the operational conditions. An example is presented in Figure 3 [16], where the top goal is "Control System is safe to operate (G1)". This goal is supported by subgoals, and solution strategies established along with a set of assumptions, justifications, and contextual information. Hence, to ensure the safety goals are met, the validity of the justification, assumptions, and solutions must be continuously monitored. Given its graphical interface, GSN has been adopted in many domains in presenting and communicating safety cases [17]. However, as noted by Langari and Maibaum [17], argument fallacies may easily creep into a safety case, as the semantics of arguments are not well defined, and pose additional challenges for reviewers, as they may often be hard to discover.

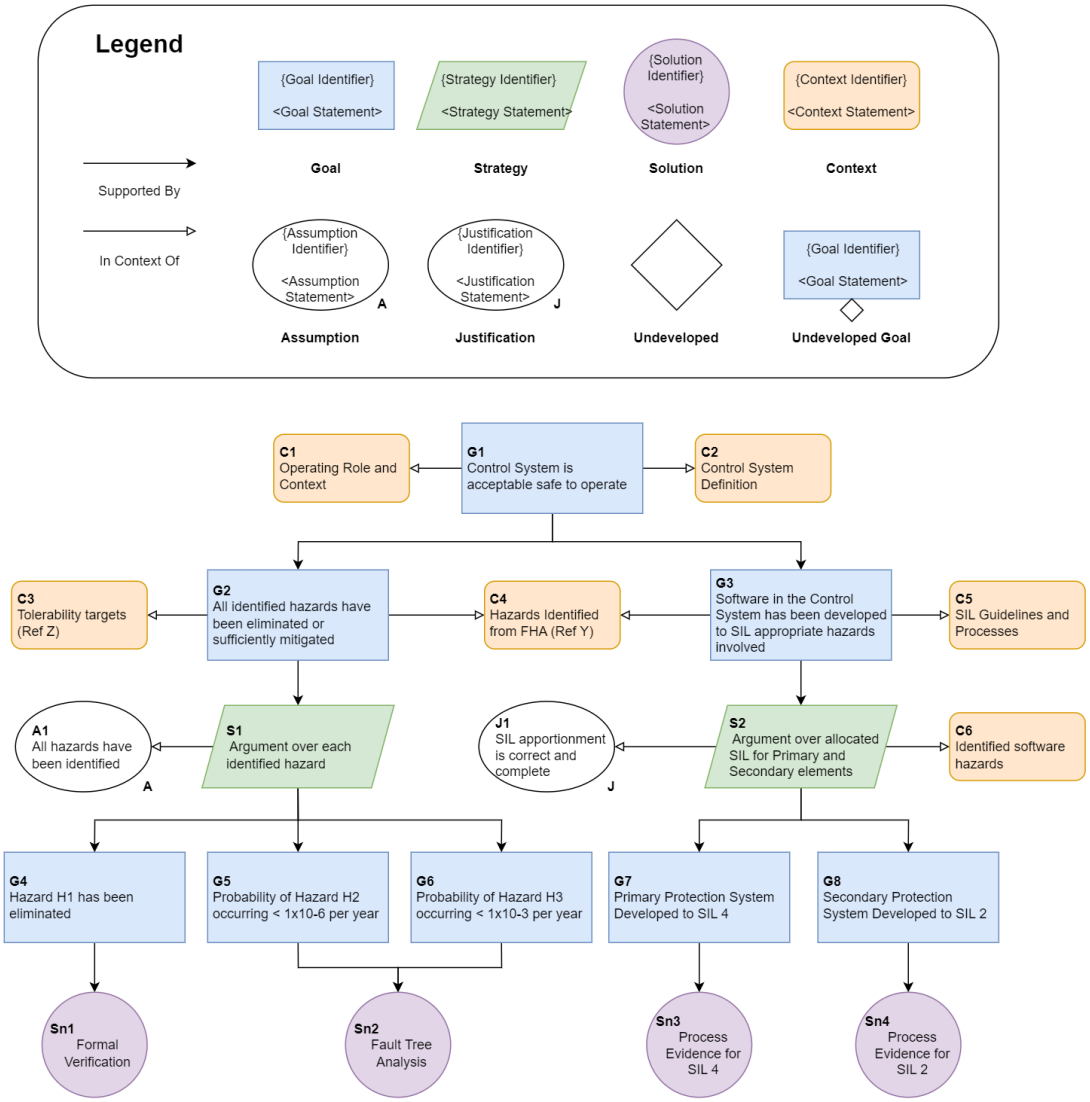



*Figure 3: An example of a goal structure. Adapted from Wei et al. [16].*

## Use and acceptance of safety cases today

Currently, the main use of safety cases in the industry is to demonstrate compliance with standards and regulations relevant to the system. In some areas, safety cases also serve to fill regulatory safety gaps while standards are under development. In this regard, the use of safety cases may lead to significant benefits during the design and operation of multiple systems, including:

- **Increased safety:** The structured process to demonstrate compliance with multiple safety regulations may help identify hazards and lead to the development of adequate risk mitigation measures.

- **Improved risk management:** Safety cases can help to improve risk management by providing a framework for identifying, assessing, and controlling risks.

Likewise, safety cases may be used to record residual risk and as a management tool during system modifications. However, **several challenges associated with the development and effectiveness of safety cases are frequently cited in recent literature**. These include:

- **Complexity:** Safety cases and resulting documentation can be complex and time-consuming to create and review. As the complexity of systems increases, so does the documentation volume, hindering readability [18].

- **Uncertainty:** Safety cases are based on evidence, but methods to propagate and validate uncertainty estimations remain an open research topic [19].

- **Resources:** The cost and time consumption of creating and maintaining a safety case can be significant. Safety cases and resulting documentation can be difficult to maintain and update as systems, procedures, and operational conditions change. With each system update, underlying assumptions about the system's environment, functions, and performance require revision.

- **Variability:** Different industries have distinctive styles and use different graphical or written structures to build and present safety cases. Likewise, a variety of evidence types and sources employed by different industries (e.g., testing, simulation) and a lack of formal theories to combine them, makes it hard to compare their use and effectiveness across industries [17].

- **Development and assessment team assembly:** It may be difficult to ensure that a varied group of experts within each organization is available to conduct the analysis required for the safety cases. Although outsourcing safety cases may lead to unrepresentative analysis, questions remain on how independent verification may be conducted effectively [20]. Similarly, these issues also affect the regulatory entities expected to assess and evaluate the safety cases.

The intent of building safety cases is the construction of *sound* safety arguments. While this is expected to rely on risk-based arguments, no specific methodology or approach is expressly required. This flexibility is explained by the heterogeneity of industries that employ safety cases and how safety-related best practices may differ. Traditional risk assessment methods are frequently used [8], such as Hazards and Operability Study (HAZOP), Structured What-If Technique (SWIFT), Fault Tree Analysis (FTA), Event Tree Analysis (ETA), Failure Mode and Effect (Criticality) Analysis

(FMEA/FMECA), multiple Human Reliability Analysis (HRA) models, as well as approaches designed specifically to analyze software sub-systems. On the other hand, methods to assess the confidence of a safety case are usually based on either Dempster–Shafer (D–S) theory or the use of Bayesian Networks [19].

An important aspect present in the discussion of safety cases is the underlying challenge of determining *how safe is safe enough*. Naturally, several scenarios involving systems ranging from transportation to the energy field may lead to some type of loss or injury to humans, i.e., the system may behave unsafely in its lifetime (due to inherent or external factors). The discussions about a safe system concern, thus, an acceptable level of risk of the system's operation, or how safe autonomous systems should be[11].

Traditionally, demonstrating safety and reducing risks arising from processes or systems' operations leans on the concept of reducing these risks to As Low As Reasonably Practicable (ALARP) [8]. This term carries challenges on its own, given the practical limitations of identifying and controlling all hazards to an acceptable level, for both individuals and societies. Another consideration is the inherently unclear definition of ALARP. A common interpretation is the cost-benefit perspective, a widely used approach that involves quantifying and comparing the unmitigated risks, with costs and benefits of risk control measures. It is however controversial to attribute monetary value to human health losses or irreparable damage to the environment [21]. From a legal perspective ALARP is also potentially problematic, with the leading court case on the subject being from Edwards v The National Coal Board from 1949 [22].

> "*Reasonably practicable' is a narrower term than 'physically possible' and seems to me to imply that a computation must be made by the owner in which the quantum of risk is placed on one scale and the sacrifice involved in the measures necessary for averting the risk (whether in money, time or trouble) is placed in the other, and that, if it be shown that there is a gross disproportion between them - the risk being insignificant in relation to the sacrifice - the defendants discharge the onus on them.*" – Asquith LJ as cited in [22]

However, courts have often turned to industry *good practice*[12] when determining what can be considered reasonably practicable [21], consequently, this view has been gaining popularity. In some jurisdictions, certain *good practices* are recognized and benefit from a special legal status [23,24]. It is generally accepted that completeness of hazard scenarios cannot be guaranteed, and setting ALARP thresholds for new systems and technologies is particularly difficult [3]. Additionally, safety case procedures are developed in a success-oriented manner. This has been noted to lead to confirmation bias issues, as negative evidence may be ignored or discarded prematurely [3]. Further, the premise of

[11] Proceedings to the International Workshop on Autonomous Systems Safety 2022. Dublin – Ireland, 28 March 2023. https://www.risksciences.ucla.edu/iwass-2022-proceedings

[12] Good Practice - Established, proven or accepted industry specific practice that meets legal or regulatory requirements. Not to be confused with Industry *best practice*, which can be considered practice above and beyond legal or regulatory requirements.

tracing evidence back to the safety claims may not be enough to demonstrate the soundness of the arguments.

**One of the biggest issues of safety cases resides in the lack of empirical evaluation, validation, and inspection mechanisms for assessing their impact throughout the life cycle of complex systems.** Usually, more importance is given to safety cases during the design, deployment, and certification of these systems and are not actively integrated into safety assessments during operation [3,4,9]. In addition to the challenges in developing, regulators face similar challenges in revision and quality assessment. In this regard, the development of safety cases may be seen more as a means to communicate the safety culture surrounding the system or process rather than a functional document to assess the system's safety.

## Variations of the safety case

As safety cases have been developed through different approaches and with different perspectives, alternative methods to demonstrate other properties of the systems have evolved. Some of these methods are also centered in safety or quality, while others may focus on communicating the trustworthiness of the safety case. Safety cases may also be referred to as *assurance cases*, which includes all types of structured argumentation demonstrating the system will operate as intended. The transition from safety cases to more general assurance case notation is in part due to the rise of security cases dedicated to digital infrastructure safety.

Security Cases or Security Assurance Cases (SACs) focus on the security of software systems. Notably, the standard ISO/SAE 21434[13] requires the development of cybersecurity cases for road vehicles in order to demonstrate that risks are not unreasonable. SACs have not been applied universally as safety cases and are considered much less mature in comparison. Furthermore, there is no standard for the required documentation or suggested development techniques [52].

**Different interpretations of safety cases have been explored to address some of the shortcomings mentioned in the previous section.** For instance, the modularization of safety cases into types has been proposed as a method to clarify the intent of the safety case. Under this umbrella term, design cases, confidence cases, and operational cases may be developed with their corresponding goals, metrics, and internal logic [17,25]. A focus on risk management led to the proposition of R*isk cases*, aiming to demonstrate that appropriate controls and mitigations have been put in place to address a system's hazards [26]. Alternative approaches have also included setting an opposite goal for the safety cases: to demonstrate the system is unsafe. By considering the worst-case scenarios, incorrect assumptions may be more easily identified [3].

[13] ISO/SAE 21434:2021 Road vehicles – Cybersecurity engineering

# The safety case in different industries

The use of safety cases has become common in many industries. However, as it has been adopted under different circumstances, the focus of the safety cases may differ for different systems. Safety cases developed for traditional sectors, such as oil and gas, railway, and automotive, are now complemented by safety cases developed specifically for rapid development of software [27], and, more recently, **autonomous systems**. The most general Assurance Case guidance for system and software engineering are given in the ISO/IEC 15026 standard[14].

However, as detailed in [1], **it is challenging to transfer safety concepts from functional safety standards to other systems in which autonomy plays an important role and employ it for certification purposes.** One of the main reasons expressed is the difficulty in identifying that all assumptions are valid under changing contextual conditions, given that the safety functions vary as well, and hence, that the specified behavior under some situations may be unsafe under unforeseen scenarios [28]. The importance of this issue is underscored by the lack of interpretability and *explainability* of AI/ML models. In this regard, safety standards ISO/PAS 21448[15] and ANSI/UL 4600[16] provide concepts more applicable to autonomous system safety, addressing performance limitations and the use of Safety Performance Indicators (SPI) to track the validity of safety claims. The sub-sections below overview the use of safety cases in different industries.

## Automotive

Technical regulations and standards provide a regulatory framework in the automotive industry [5]. These regulations and standards are complemented by multiple best practices developed by multiple organizations, such as the U.S. National Highway Traffic Safety Administration (NHTSA) and Federal Motor Vehicle Safety Standards (FMVSS), as well as the EU and UN-ECE. One of the most relevant safety standards for automotive systems is the ISO 26262[17] standard. This document describes different safety argumentation tiers: safety goals, functional safety requirements, technical safety requirements. This covers aspects from design, operation, and incident response stages, and uses an Automotive Safety Integrity Level (ASIL) risk categorization. In general, compliance to the ISO 26262 standard is considered to be similar to the construction of

[14] ISO/IEC/IEEE 15026:2019 Systems and software engineering – Systems and software assurance.
[15] ISO/PAS 21448:2019 Road Vehicles – Safety of the Intended Functionality.
[16] ANSI/UL 4600:2022 Standard for Evaluation of Autonomous Products.
[17] ISO 26262:2018 Road vehicles – Functional Safety.

a safety case. Similarly, ISO/NP PAS 8800[18] presents the derivation of evidence required to support assurance argumentation for AI/ML-based functionalities [1].

Regarding commercial applications of Automated Driving Systems (ADS), Waymo recently published a report detailing the strategy and systematic approach towards the creation of safety cases [29]. This approach consists of three main elements: a layered approach to safety to determine Absence of Unreasonable Risk (AUR) based on acceptance criteria measured by SPI, a dynamic approach to safety determination lifecycle, and a credible approach to safety through case credibility assessment framework. This report joins the growing number of safety reports published by developers such as Cruise, Aurora, and Zoox. As the regulatory environment surrounding the use of Automated Driving Systems evolves, questions remain regarding whether simulation outputs may be considered sufficient evidence of safe design, or to what extent real-world road tests need to be designed to support risk arguments. From a policy perspective, performance-based safety certification would also present challenges for regulators to maintain an appropriate level of knowledge, capability, and capacity to both provide guidance and conduct safety case audits [28].

## Railway

In the UK, safety cases are required to be developed for railway operations. Standard EN 50129[19] states the safety acceptance conditions for the railway and calls for the safety case to demonstrate that the conditions are fulfilled. EN 50129 provides hierarchical safety objectives and recommendations for techniques to demonstrate compliance. The safety case must be developed by the manufacturer and assessed by an independent third party before system commissioning.

Myklebust and Stålhane [27] have used the structural requirements for safety cases from EN 50129 to propose a framework for the development of agile safety cases. Agile safety cases are constructed during system development, thereby continuously introducing new functionality, and shortening the time to market. Agile safety cases adapt to continuous project developments and consider the entire system lifecycle.

Other research on safety cases for the railway industry include Wang et al. [30], who propose two techniques for improving the requirements from EN 50129 by visualizing the safety case using GSN, and to include a framework to quantify confidence in the evidence. The GSN has also been proposed and used for structuring safety cases for autonomous trains [31].

[18] ISO/AWI PAS 8800 Road Vehicles – Safety and artificial intelligence (under development).
[19] EN 50129:2018 Railway applications - Communication, signalling and processing systems – Safety related electronic systems for signalling.

## Oil and Gas, Process industry

On the United Kingdom Continental Shelf (UKCS) the workplace health and safety in the offshore industry is regulated by the Health and Safety Executive (HSE) [20]. No offshore installation may be operated without an approved safety case[21,22]. HSE relies increasingly on the good practice interpretation of ALARP [21]. In the UK, good practices in the Approved Codes of Practice (ACOP) as determined by the Health and Safety Commission (HSC) enjoy special legal status. For these practices it is sufficient to demonstrate compliance with an approved practice to demonstrate compliance with the law.

In Australia, the offshore regulator is National Offshore Petroleum Safety and Environmental Management Authority (NOPSEMA). The Offshore Petroleum and Greenhouse Gas Storage (Safety) (OPGGS(S)) regulations of 2009, requires all operators of offshore installations in commonwealth waters to submit a safety case to NOPSEMA. Safety cases must facilitate workforce[23] participation and educate the workforce on the relevant hazards, control measures and any potential vulnerabilities. All offshore activities are required to adhere to an NOPSEMA approved safety case [34]. OPGGS(S) requires an operator to ensure that the health and safety related risks are reduced to ALARP, and the onus is on the operator to demonstrate that ALARP is achieved. NOPSEMA (2022) states that in many cases referring to industry good practice may be sufficient [23].

In Singapore, installations involved in producing, processing, manufacturing, or storing substantial quantities of flammable or toxic materials are identified as Major Hazard Installations (MHIs). The Major Hazards Department (MHD) serves as the competent authority, overseeing these installations under a safety case regulatory framework modeled on Australia and European safety case frameworks [35]. MHD acknowledges that demonstrating control measures as ALARP may necessitate a combination of different approaches such as the good practice approach, hazard and risk criteria or cost benefit analysis. Although no source for good practice provides special legal status, the following order of precedence is provided: legislation, regulatory guidance, standards by standard-making organizations, guidance by industry representative organization, and industry good practice. MHD requires new MHI to be compliant to IEC 61511 regarding Safety Instrumented Functions (SIFs) [36]. MHD also promotes the application of single major hazard scenario risk matrices and acceptance criteria. With this

[20] HSE defines operation as all activities related to exploration and production, including design, planning, construction, operation, and decommissioning. But excluding activities related to transportation of petroleum products.

[21] Offshore Installations (Safety Case) Regulations 2005 Regulation 7(1)(a) & 8(1)(a). Accessed [online] at: https://www.legislation.gov.uk/uksi/2005/3117 on 12.04.23.

[22] Offshore Installations (Offshore Safety Directive) (Safety Case) Regulations 2015. Accessed [online] at: https://www.legislation.gov.uk/uksi/2015/398/ on 12.04.23.

[23] Workforce is defined in OPGGS(S) to include any member who is identifiable prior to the formulation of a safety case and is working, or likely to be working on the installation in question.

approach, the residual risk for comparative scenarios at newer MHIs should be less than or equal to that of older installations [35,37].

The cost benefit approach is suggested for instances where it is difficult to determine whether the cost of a measure is justified. A comparative assessment of possible control measures and their respective cost benefits may also be utilized. MHD refers to the UK HSE on the uses and limitations of a cost benefit approach but does emphasize that determination if a control measure is "reasonably practicable" should not be done in isolation, since the cost may be distributed across multiple major accident scenarios [35,36]. Furthermore, any safety critical events that require human intervention are defined as safety critical tasks and require a Human Reliability Analysis using recognized approaches such as human-HAZOP [38].

## Industrial Automation

In the European Union (EU), Directive 2006/42/EC[24], commonly known as the "Machinery Directive", establishes a standardized level of safety and facilitates the unrestricted movement of compliant machinery across all member nations[25]. Each individual member nation is responsible for enacting legislation and establishing a competent authority in accordance with the directive. Additionally, member nations are tasked with overseeing the market surveillance, verifying compliance of machinery, and implementing appropriate measures in cases of non-compliance.

The directive mandates that all machinery must adhere to the essential health and safety requirements specified, however, it also emphasizes the importance of considering state-of-the-art practices. Consequently, an iterative risk assessment must be conducted during the design and construction of machinery to determine which health and safety requirements are applicable[26]. Prior to its introduction into the EU market, the manufacturer is obliged to perform a conformity assessment and affix the "CE" conformity mark to the machinery as seen in Figure 4. Moreover, the directive mandates the manufacturer to create and maintain a technical file, where the manufacturer demonstrates the conformity of the machinery in question. Although the use of harmonized standards is not mandatory for demonstrating conformity, machinery that complies with relevant harmonized standards[27] published in the Official Journal of the European Union is presumed to conform to the directive. The technical file should also

[24] Directive 2006/42/EC - new machinery directive. European Agency for Safety and Health at Work, 2021.

[25] Machinery – "an assembly, fitted with or intended to be fitted with a drive system other than directly applied human or animal effort, consisting of linked parts or components, at least one of which moves, and which are joined together for a specific application." – Directive 2006/42/EC

[26] Guide to application of the Machinery Directive 2006/42/EC. European Commission, 2022.

[27] Harmonized standard – standard developed by CEN, CENELEC, or ETSI.

encompass aspects that are not easily inspectable by the competent authority, such as the design and manufacturing processes of the machinery.

While the directive does not explicitly mention safety cases, the technical file can be considered akin to a safety case, as it serves as the primary document through which the manufacturer demonstrates compliance with the directive.


*Figure 4: Robotic gripper with CE conformity mark at the networked automation systems lab, Institute of Industrial Automation and Software Engineering, University of Stuttgart.*

## Nuclear industry

Nuclear plants in the UK must be justified by a safety case. In the nuclear industry, safety cases have two core focuses: 1) a deterministic hazard analysis, and 2) demonstrating that the provisions to prevent these hazards are sufficient and adequate [10]. Nuclear safety cases often describe and provide evidence for the layered defense provisions as the "defense in depth" of the system. In cases where defense in depth is not feasible, the "incredibility of failure" has been used to provide evidence, either probabilistically or deterministically [39]. Experiments of passive safety systems are also common to demonstrate evidence [40].

Nuclear safety cases are often difficult to read or disassociated from the operational aspects of the plant. The Office for Nuclear Regulation (ONR) has identified a number of common problems with safety cases, including intelligibility, completeness, and validity. In response to these concerns, ONR has issued clearer instructions on the contents that should be included in a nuclear safety case [41]. Furthermore, research on safety cases in the nuclear domain has recently aimed at writing "usable" safety cases. This stems from the acknowledgement that safety cases in the nuclear industry are overly complex [41].

## Aeronautical and Aerospace

Safety cases are an alternative method of demonstrating safety in aviation. The development, research, and applications of safety cases have primarily been driven by NASA, including the technical report "Considerations in Assuring Safety of Increasingly Autonomous Systems" [42] and the European Union Aviation Safety Agency (EASA) concept paper "First usable guidance for Level 1 machine learning applications" [43]. The wide majority of safety cases in aviation have been applied to Unmanned Aircraft Systems. The NASA System Safety Handbook explains the use and structure of the Risk Informed Safety Case (RISC) [44]. The RISC specifically refers to the totality of safety-related documentation that must be submitted for technical reviews. The guidelines specifically state that the evaluation of the RISC should be aimed at discovering flaws within the safety argument, rather than justifying it as proof.

Research within the aviation domain has aimed to improve methods for generating and structuring evidence. Formal methods have demonstrated suitability for automated generation of evidence [45]. The concept of safety architecture extends from the bow tie diagrams to demonstrate overall safety assurance [46].

## Other industries

In the UK, safety cases are used to justify safety "for a given application in a given environment" for all defense Products, Services, and Systems (PSS) [6]. This wide umbrella includes individual pieces of equipment, as well as large-scale systems such as ships and air defense systems. Safety cases must be managed throughout the duration of the PSS life – they must demonstrate how safety "will be, is being and has been, achieved and maintained." Thus, *safety case reports* are used to summarize the arguments and evidence of the safety case at a given time. The UK defense standard provides non-mandatory guidelines on the topics that should be addressed in a safety case report. These are the scope; identified hazards and accidents; assumptions, dependencies, and limitations; the context of use; unusual aspects of the design; and safety justification. The justification must be accompanied by a search and treatment of potential counter evidence. Similarly, in the U.S, the DARPA Assured Autonomy Program aims for continuous safety and functional assurance [47].

Safety cases have been proposed and studied by the UK HSE for patient safety management in healthcare. The review by the Health Foundation [5] offers a unique look at the opportunities for safety cases in a novel application setting. The benefits include integrated evidence, added communication among stakeholders, explicit documentation, and aided safety management. However, three risks were also identified with the use of safety cases: 1) that the safety case becomes a paper exercise, 2) the separation of the safety case from actual operation, and 3) production of the safety case by external parties.

The review concludes by recommending the use of safety cases along with a call for increased guidance and training in the techniques for their development [5]. These concerns are echoed by research [48,49]. After repeated safety issues with infusion pumps, devices that require accurate and precise control for drug delivery, the United States Food and Drug Administration (FDA) introduced safety assurance cases as a requirement for their regulatory review. These cases are used to demonstrate that new or modified infusion pumps are as safe and effective as the previous device. To introduce the requirement, the FDA initiated a Safety Assurance Case Pilot Program, in which industry voiced concerns on safety cases and feedback for the FDA to provide clear requirements for safety case development. The safety case requirement resulted in more approvals of infusion pumps and no significant change in FDA review time.

The maritime industry has not adopted any formal regulations for the provision or contents of safety cases. The International Maritime Organization (IMO) has instead developed the formal safety assessment (FSA) to assess risk on a fleet or ship type level [53], whereas safety cases investigate a specific design or operation. Nevertheless, safety cases have been identified as a promising tool to assess the operations of individual autonomous ships [32]. The GSN proved especially compatible for an autonomous ship prototype [33]. Challenges remain for the widespread adoption of a novel method in the historically conservative maritime domain.

Additionally, safety cases have been applied to an assortment of other complex and safety-critical systems. These include high-rise construction to manage the risks of fire and structural failures [50], as well as the use of digital twins to generate evidence for safety cases of collaborative robotics [51].

# Conclusion

The concept of safety cases plays a significant role in system safety regulation and certification in multiple industries. Different versions and interpretations have been developed to address the particularities of the different systems in which it is applied, such as in the automotive, nuclear, railway, industrial automation, marine, and oil and gas industries.

The main strength of a safety case is the underlying structure required for the construction of the safety arguments, based on evidence to provide safe assurance within certain levels of confidence [4]. Even if safety cases play an important role in the regulation and certification of many safety-critical industries, to date, there has been little empirical research on their use and efficiency [4], [9]. A main concern about the use of safety cases as guiding documentation to assure safety is that this could result in complex and costly processes purely focused on the design stages and not fully integrated into the life cycle of the processes of systems that they analyze [3], [4]. In this regard, a safety case in itself is not intended to be an end goal, but a means of achieving safety for which the necessary provisions need to be taken, such as adequate team confirmation, clear use of safety metrics and communication strategies. Further, the use of safety cases for certification purposes implies that regulatory bodies would require enough knowledge, capability, and capacity to both provide guidelines and to conduct audits. This may lead to regulatory gaps, in which safety cases are put forward as acceptable evidence of system safety but lack external validation or oversight.

**The challenges concerning the use of safety cases become more important as systems are increasingly supported or operated by autonomous agents.** In particular, a critical issue is how the validity of system assumptions is challenged by the **"explosion" of possible scenarios** under which autonomous systems are expected to function. Coupled with the **lack of interpretability and *explainability*** of AI-based functionalities prevalent in autonomous systems, the collection of evidence would then rely on designing representative virtual and real-world tests as well as the construction and tracking of SPIs throughout the system's life cycle. While the effort in building a strong safety case cannot be underscored sufficiently, the most difficult task may be for regulators to determine **if a safety case is safe enough**.

**As we come close to the 4th IWASS, we invite participants to reflect on the topics discussed in this white paper:**

- How can safety cases be used to improve system design and safety assurance and not become *only* a certification requirement?

- What should a regulatory entity accept as a minimum in a safety case? How to demonstrate sufficient completeness without converting it into a standard compliance checklist?
- How to track and ensure assumptions are valid in changing environments, as the foundations of the safety arguments? What happens when a safety argument is invalidated? How are these monitored?

# References

[1] R. Adler and M. Klaes, "Assurance Cases as Foundation Stone for Auditing AI-Enabled and Autonomous Systems: Workshop Results and Political Recommendations for Action from the ExamAI Project", *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 13520 LNCS, pp. 283–300, 2022, doi: 10.1007/978-3-031-18158-0_21.

[2] M. Borg *et al.*, "Safely Entering the Deep: A Review of Verification and Validation for Machine Learning and a Challenge Elicitation in the Automotive Industry", *J. Automot. Softw. Eng.*, vol. 1, no. 1, pp. 1–19, Jan. 2019, doi: 10.2991/JASE.D.190131.001.

[3] N. Leveson, "The Use of Safety Cases in Certification and Regulation", *ESD Work. Pap. Ser.*, no. November, pp. 1–12, 2011, [Online]. Available: https://dspace.mit.edu/bitstream/handle/1721.1/102833/esd-wp-2011-13.pdf?sequence=1&isAllowed=y

[4] T. Kelly, "Safety Cases", *Handbook of Safety Principles*, 10.1016/B978-1-4377-3524-6.00006-X, Ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2018, pp. 361–385. doi: 10.1002/9781119443070.ch16.

[5] The Health Foundation, "Using safety cases in industry and healthcare", no. December. 2012. [Online]. Available: http://www.health.org.uk/publications/using-safety-cases-in-industry-and-healthcare/

[6] Defence Standard 00-56 "Part 1: Safety Management Requirements for Defence Systems – Requirements", *(UK MoD Def Stan 00-56 Part 1, Issue 7, 2017)*.

[7] Rinehart, David J., J. C. Knight, and J. Rowanhill, "Understanding What It Means for Assurance Cases to" Work"", *No. NF1676L-26066. 2017.*

[8] R. Maguire, "Safety cases and safety reports: Meaning, motivation and management", CRC Press, 2006.

[9] I. Habli, R. Alexander, and R. Hawkins, "Safety cases: an impending crisis", *Safety-Critical Syst. Symp.*, 2020, [Online]. Available: https://eprints.whiterose.ac.uk/169183/

[10] R. Bloomfield and P. Bishop, "Safety and Assurance Cases: Past, Present and Possible Future – an Adelard Perspective", in *Making Systems Safer*, London: Springer London, 2010, pp. 51–67. doi: 10.1007/978-1-84996-086-1_4.

[11] T. P. Kelly, "Arguing safety-a systematic approach to safety case management", DPhil Thesis York University, Department of Computer Science Report YCST, 1999.

[12] V. Sklyar and V. Kharchenko, "Assurance Case For Safety And Security Implementation: A Survey Of Applications", *Int. J. Comput.*, vol. 19, no. 4, pp. 610–619, Dec. 2020, doi: 10.47839/ijc.19.4.1995.

[13] R. E. Bloomfield and K. Netkachova, "Building Blocks for Assurance Cases", *International Symposium on Software Reliability Engineering (ISSRE)*, 2014. [Online]. Available: http://openaccess.city.ac.uk/5121/

[14] The Assurance Case Working Group (ACWG), "Goal Structuring Notation Community Standard Version 3 The Assurance Case Working Group (ACWG)", 2021, [Online]. Available: https://scsc.uk/scsc-141C

[15] V. Sklyar and V. Kharchenko, "Assurance Case Driven Design for Computer Systems: Graphical Notations versus Mathematical Methods", pp. 308–312, Jan. 2017, doi: 10.1109/MCSI.2016.063.

[16] R. Wei *et al.*, "Model based system assurance using the structured assurance case metamodel", *J. Syst. Softw.*, vol. 154, pp. 211–233, Aug. 2019, doi: 10.1016/j.jss.2019.05.013.

[17] Z. Langari and T. Maibaum, "Safety cases: A review of challenges", *1st International Workshop on Assurance Cases for Software-Intensive Systems (ASSURE)*, May 2013, pp. 1–6. doi: 10.1109/ASSURE.2013.6614263.

[18] P. Wilkinson, "Safety Cases: Success or Failure?", 2002.

[19] D. Nešić, M. Nyberg, and B. Gallina, "A probabilistic model of belief in safety cases", *Saf. Sci.*, vol. 138, no. February, p. 105187, Jun. 2021, doi: 10.1016/j.ssci.2021.105187.

[20] J. Rushby, "Formalism in Safety Cases", *Making Systems Safer*, C. Dale and T. Anderson, Eds. London: Springer London, 2010, pp. 3–17. doi: 10.1007/978-1-84996-086-1_1.

[21] A. Hopkins, "The need for a general duty of care", *Houst. J. Int. Law*, vol. 37, pp. 841–848, 2015.

[22] P. Ho, "Do safety cases demonstrate risks have been reduced so far as is reasonably practicable? an Australian study examining the methods of presenting safety cases", *Saf. Sci.*, vol. 159, no. October 2021, p. 106042, Mar. 2023, doi: 10.1016/j.ssci.2022.106042.

[23] NOPSEMA, "ALARP guidance note." https://www.nopsema.gov.au/search?keys=N-04300-GN0166+A138249 (accessed Apr. 12, 2023).

[24] Health and Safety Executive (HSE), "Assessing compliance with the law in individual cases and the use of good practice", 2003.

[25] R. Hawkins, I. Habli, and T. Kelly, "Principled Construction of Software Safety Cases", *SAFE- COMP 2013 - Workshop SASSUR (Next Generation of System Assurance Approaches for Safety- Critical Systems) of the 32nd International Conference on Computer Safety, Reliability and Security*, 2013.

[26] C. Haddon-Cave, "The Nimrod Review", HC 1025, London: The Stationery Office Limited, Oct. 28, 2009.

[27] T. Myklebust and T. Stålhane, "The Agile Safety Case", Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-319-70265-0.

[28] S. Ballingall, M. Sarvi, and P. Sweatman, "Safety assurance for automated driving systems that can adapt using machine learning: A qualitative interview study", *J. Safety Res.*, vol. 84, pp. 243–250, 2023, doi: 10.1016/j.jsr.2022.10.024.

[29] F. Favarò *et al.*, "Building a Credible Case for Safety: Waymo's Approach for the Determination of Absence of Unreasonable Risk", no. March. 2023. [Online]. Available: www.waymo.com/safety

[30] R. Wang *et al.*, "Modelling confidence in railway safety case", *Saf. Sci.*, vol. 110, no. November 2017, pp. 286–299, Dec. 2018, doi: 10.1016/j.ssci.2017.11.012.

[31] M. Chelouati *et al.*, "Graphical safety assurance case using Goal Structuring Notation (GSN) — challenges, opportunities and a framework for autonomous trains", *Reliab. Eng. Syst. Saf.*, vol. 230, no. November 2022, p. 108933, Feb. 2023, doi: 10.1016/j.ress.2022.108933.

[32] J. Montewka *et al.*, "Challenges, solution proposals and research directions in safety and risk assessment of autonomous shipping", I*n PSAM 14 - Probabilistic Safety Assessment and Management*, 2018. [Online]. Available: http://psam14.org/proceedings/paper/paper_426_1.pdf

[33] E. Heikkilä *et al.*, "Safety Qualification Process for an Autonomous Ship Prototype – a Goal-based Safety Case Approach", *TransNav 2017 - 12th International Conference on Marine Navigation and Safety of Sea Transportation*, 2017.

[34] NOPSEMA, "The safety case in context: An overview of the safety case regime", https://www.nopsema.gov.au/offshore-industry/safety/safety-cases-and-validation (accessed Apr. 12, 2023).

[35] J. Lim *et al.*, "Safety Case Technical Guide", *Ministry of Manpower, Singapore*, 2016. https://www.mom.gov.sg/workplace-safety-and-health/major-hazard-installations/preparing-for-safety-case (accessed Apr. 26, 2023).

[36] J. Lim *et al.*, "Guidelines on Safety Instrumented Systems in Major Hazards Installations", *Ministry of Manpower, Singapore*, 2020. https://www.mom.gov.sg/workplace-safety-and-health/major-hazard-installations/preparing-for-safety-case (accessed Apr. 26, 2023).

[37] J. Lim *et al.*, "ALARP Demonstration Guidelines: Single Scenario Risk Tolerability Target and Adequacy of Barriers", *Ministry of Manpower, Singapore*, 2020. https://www.mom.gov.sg/workplace-safety-and-health/major-hazard-installations/preparing-for-safety-case (accessed Apr. 26, 2023).

[38] J. Lim *et al.*, "Guidelines on Managing Human Factors in Major Hazard Installations", *Ministry of Manpower, Singapore*, 2022. https://www.mom.gov.sg/workplace-safety-and-health/major-hazard-installations/preparing-for-safety-case (accessed Apr. 26, 2023).

[39] R. Bullough *et al.*, "The demonstration of incredibility of failure in structural integrity safety cases", *Int. J. Press. Vessel. Pip.*, vol. 78, no. 8, pp. 539–552, 2001, doi: 10.1016/S0308-0161(01)00070-9.

[40] T. Mull *et al.*, "Safety cases for design-basis accidents in LWRs featuring passive systems", *Nucl. Eng. Des.*, vol. 387, no. January 2021, 2022, doi: 10.1016/j.nucengdes.2021.111095.

[41] F. I. Pérez, "Writing 'usable' Nuclear Power Plant (NPP) safety cases using bowtie methodology", *Process Saf. Environ. Prot.*, vol. 149, pp. 850–857, May 2021, doi: 10.1016/j.psep.2021.03.022.

[42] E. E. Alves *et al.*, "Considerations in assuring safety of increasingly autonomous systems", 2018.

[43] European Union Aviation Safety Agency, "First usable guidance for Level 1 machine learning applications", *EASA Concept Paper - A deliverable of the EASA AI Roadmap*, 2021. https://www.easa.europa.eu/en/newsroom-and-events/news/easa-releases-its-concept-paper-first-usable-guidance-level-1-machine-0 (accessed May 08, 2023).

[44] H. Dezfuli *et al.*, "NASA system safety handbook." No. NASA/SP-2010-580/Version 1.0, 2011.

[45] E. Denney, G. Pai, and J. Pohl, "Heterogeneous aviation safety cases: Integrating the formal and the non-formal", *Proc. - 2012 IEEE 17th Int. Conf. Eng. Complex Comput. Syst. ICECCS 2012*, no. 3, pp. 199–208, 2012, doi: 10.1109/ICECCS.2012.20.

[46] E. Denney, G. Pai, and I. Whiteside, "The role of safety architectures in aviation safety cases", *Reliab. Eng. Syst. Saf.*, vol. 191, no. May, p. 106502, Nov. 2019, doi: 10.1016/j.ress.2019.106502.

[47] Defense Advanced Research Projects Agency (DARPA), "Assured Autonomy Program", https://www.darpa.mil/program/assured-autonomy (accessed Jan. 19, 2022).

[48] M. A. Sujan *et al.*, "Safety cases for medical devices and health information technology: Involving health-care organisations in the assurance of safety", *Health Informatics J.*, vol. 19, no. 3, pp. 165–182, Sep. 2013, doi: 10.1177/1460458212462079.

[49] M. A. Sujan *et al.*, "Should healthcare providers do safety cases? Lessons from a cross-industry review of safety case practices", *Saf. Sci.*, vol. 84, pp. 181–189, Apr. 2016, doi: 10.1016/j.ssci.2015.12.021.

[50] Health and Safety Executive (HSE), "Safety cases and safety case reports", https://www.hse.gov.uk/building-safety/safety-cases-reports.htm (accessed Apr. 26, 2023).

[51] J. A. Douthwaite *et al.*, "A Modular Digital Twinning Framework for Safety Assurance of Collaborative Robotics", *Front. Robot. AI*, vol. 8, p. 758099, 2021, doi: 10.3389/frobt.2021.758099.

[52] M. Mohamad, J.P. Steghöfer, and R. Scandariato, "Security assurance cases—state of the art of an emerging approach". Empir Software Eng 26, 70 (2021). https://doi.org/10.1007/s10664-021-09971-7https://doi.org/10.1007/s10664-021-09971-7

[53] J. Wang, "Offshore safety case approach and formal safety assessment of ships". *J Safety Res*, vol. 33, no. 1, pp. 81–115, 2002.


https://www.risksciences.ucla.edu/iwass-2023-home

UCLA ENGINEERING
B. John Garrick Institute for the Risk Sciences

NTNU – Trondheim
Norwegian University of
Science and Technology

Universität Stuttgart

**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**

DOI: 10.34948/G4MW2N
ISSN: 2995-8709