The Safety Case for Autonomous Systems: An Overview
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
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

Summary .................................................................................................................................................. 1

The International Workshop on Autonomous System Safety ......................................................... 1

Scope and goal of the white paper ..................................................................................................... 2

The safety case ...................................................................................................................................... 4

Background and history ......................................................................................................................... 5

The elements of a safety case .................................................................................................................. 5

Use and acceptance of safety cases today .............................................................................................. 9

Variations of the safety case .................................................................................................................... 12

The safety case in different industries .................................................................................................. 13

Automotive .......................................................................................................................................... 14

Railway ............................................................................................................................................... 14

Oil and Gas, Process industry ............................................................................................................... 15

Industrial Automation ............................................................................................................................ 16

Nuclear industry .................................................................................................................................... 17

Aeronautical and Aerospace ................................................................................................................ 18

Other industries .................................................................................................................................... 18

Conclusion .......................................................................................................................................... 20

References ......................................................................................................................................... 22
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: 2019¹, 2021², 2022³.

IWASS 2023⁴ will take place in Southampton, United Kingdom, on September 2nd and 3rd.

---
⁴ International Workshop on Autonomous Systems Safety 2023. [https://www.risksciences.ucla.edu/iwass-2023-home](https://www.risksciences.ucla.edu/iwass-2023-home)
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 4\(^{th}\) 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.

---


\(^7\) 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, IEC 61511, or ISO 26262. 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

---

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].

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 system 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.
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.
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 be11.

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 practice12 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

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 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 Risk 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].

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\(^\text{14}\).

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\(^\text{15}\) and ANSI/UL 4600\(^\text{16}\) 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\(^\text{17}\) 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


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)\textsuperscript{20}. No offshore installation may be operated without an approved safety case\textsuperscript{21,22}. HSE relies increasingly on the good practice interpretation of ALARP\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{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)\textsuperscript{36}. MHD also promotes the application of single major hazard scenario risk matrices and acceptance criteria. With this

\textsuperscript{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.


\textsuperscript{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 MHI's 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, commonly known as the "Machinery Directive", establishes a standardized level of safety and facilitates the unrestricted movement of compliant machinery across all member nations. 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. 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 published in the Official Journal of the European Union is presumed to conform to the directive. The technical file should also

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
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.

![Robotic gripper with CE conformity mark at the networked automation systems lab, Institute of Industrial Automation and Software Engineering, University of Stuttgart.](image)

**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


