Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P1181-cd

Evolving Perspective of Safety, Reliability, and Security of Autonomous Systems – Findings from IWASS 2024

Andrey Morozov, Joachim Grimstad

Institute of Industrial Automation and Software Engineering, University of Stuttgart, Germany.
E-mail: andrey.morozov@ias.uni-stuttgart.de, joachim.grimstad@ias.uni-stuttgart.de

Camila Correa-Jullian, Marilia Ramos, Ali Mosleh

B. John Garrick Institute for the Risk Sciences, University of California, Los Angeles (UCLA), United States.
E-mail: ccorreaj@g.ucla.edu, marilia@risksciences.ucla.edu, mosleh@ucla.edu

Spencer August Dugan, Ingrid Bouwer Utne

Department of Marine Technology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, Norway. E-mail: spencer.a.dugan@ntnu.no, ingrid.b.utne@ntnu.no

Christoph Alexander Thieme

Department of Software Engineering, Safety, and Security, SINTEF Digital, Trondheim, Norway.
E-mail: christoph.thieme@sintef.no

This paper summarizes the key insights and discussions from the International Workshop on Autonomous System Safety (IWASS) 2024, held in Krakow, Poland. As the fifth iteration of the IWASS series, the workshop brought together experts from academia, industry, and regulatory bodies to address critical challenges in the safety, reliability, and security (SRS) of autonomous systems. The event highlighted the interdisciplinary nature of SRS, focusing on diverse domains such as automotive, maritime, robotics, and industrial automation. Key themes included human-autonomous system interaction, methods to address the risk of autonomous systems, regulatory challenges, and advancements in sensor technologies. Discussions underscored the need for robust frameworks to ensure safe and reliable system operations, emphasizing the integration of real-time monitoring, explainable AI, and continuous safety assessments. The findings from IWASS 2024 offer a roadmap for future research and industry collaboration, aiming to overcome existing barriers and foster the safe and widespread adoption of autonomous technologies.

Keywords: Autonomous Systems, Safety, Reliability, Security, Risk.

1. Introduction

The International Workshop on Autonomous System Safety (IWASS) 2024 marks the fifth installment of this highly focused series on the Safety, Reliability, and Security (SRS) of autonomous systems. IWASS is a platform for fostering collaboration among experts from academia, regulatory bodies, and industry. The workshop facilitates discussions aimed at addressing challenges and exploring innovative solutions to enhance the SRS of autonomous systems. Building on the success of previous workshops held in Trondheim (Thieme et al., 2019), online (Thieme et al., 2021), Dublin (Thieme et al., 2022), and Southampton (Correa-Jullian et al., 2023), IWASS 2024 (Correa-Jullian et al., 2024) took place on June 23 in the city of Krakow, Poland. Hosted ahead of the 34th European Safety and Reliability Conference (ESREL), the workshop gathered 30 participants from 23 organizations across 11 countries.

IWASS attracts experts from multiple disciplines. One of the main goals is to find synergies in the SRS of autonomous systems across different domains, including the automotive industry, industrial robotics, ships, energy, and mobile robots. Despite apparent differences between these areas, the underlying methods for analyzing and ensuring SRS share numerous similarities. For any socio-technical system with a high level of autonomy, it is crucial to understand and evaluate risks

by identifying hazards and failure scenarios and quantifying their likelihoods and consequences. It is equally important to maintain an acceptable operational risk level by viewing safety assurance as a continuous process, incorporating ongoing safety monitoring and anomaly detection. Finally, learning from these results and implementing regulations and procedures to ensure SRS is vital, as is establishing effective communication to build trust with society.

The 2024 workshop focused on several key themes, including risk assessment methodologies, operational performance and safety challenges, regulatory frameworks, perception and AI, and human-system interaction in autonomous systems. Discussions spanned a range of topics, from the complexities of safety assurance in open environments to the evolving roles of humans in overseeing and collaborating with autonomous systems. Also, as the highlight of the fifth installment of IWASS, we aimed to reassess and reevaluate the state of autonomous systems' deployments: Are the expectations set five years ago about increased system efficiency and safety reasonable? What are the main unresolved issues that challenge the adoption of these technologies? Is a shift in research, industry, and regulatory paradigms required to address these challenges, or are external – business and social – pressures the main forces behind the advances and setbacks of autonomous system deployments?

2. Safety, Reliability, and Security

The domains of **reliability**, **risk**, **safety**, and **security**, as shown in Figure 1, are closely interrelated and must be addressed collectively to ensure comprehensive autonomous system safety. **Safety** is commonly defined as the *"state where risk has been reduced to a level that is as low as reasonably practicable and where the remaining risk is generally acceptable"* (Rausand and Haugen, 2020), directly linking it to the concept of risk. According to Kaplan and Garrick (1981) seminal work, **risk** is quantified by addressing three key questions: *"What can go wrong?"*, *"How likely is it to happen?"*, and *"What are the consequences?"* These questions are also explored in other works, including those by Aven (2014, 2012).

The concepts of risk analysis closely align with Laprie's **dependability** theory (Avizienis et al., 2004), which defines key terminologies. Dependability is categorized into three groups: (i) **attributes**, such as availability, reliability, safety, confidentiality, integrity, and maintainability; (ii) **means**, including fault prevention, fault tolerance, fault removal, and fault forecasting; and (iii) **threats**, classified as faults, errors, and failures. The threats category is particularly relevant in this context. A **fault** is a system defect that, when activated, may cause an error. An **error** is an incorrect internal state or a discrepancy between intended and actual behavior. A **failure** occurs when the system's external behavior deviates from its specification. As illustrated in Figure 1, faults in technical systems can arise from various causes, propagate as errors, and ultimately lead to failures, preventing the system from fulfilling its function.

The concepts of **reliability** and **safety**, though related, differ significantly. **Reliability** refers to *the continuity of correct service* or *the system's ability to perform as specified under given conditions for a defined period*. For example, standards define reliability as *"the ability of a machine or its components to perform a required function under specified conditions and for a given period without failing"* (ISO12100, 2010), and *"the ability of an item to perform a required function under given environmental and operational conditions for a stated period"* (ISO26262, 2011). In contrast, **safety** focuses on *the absence of catastrophic consequences* (Avizienis et al., 2004). Standards define safety as *"freedom from unacceptable risk"* (IEC61508, 2010) and *"absence of unreasonable risk due to hazards caused by malfunctioning behavior of Electrical/Electronic systems"* (ISO26262, 2011). A system can be reliable but unsafe, or vice versa. For example, an autonomous vehicle with an empty battery that remains parked in a garage is perfectly safe but not reliable, while an industrial robot operating without safety sensors near human workers may be reliable but poses significant safety risks.

Security refers to protecting autonomous sys-

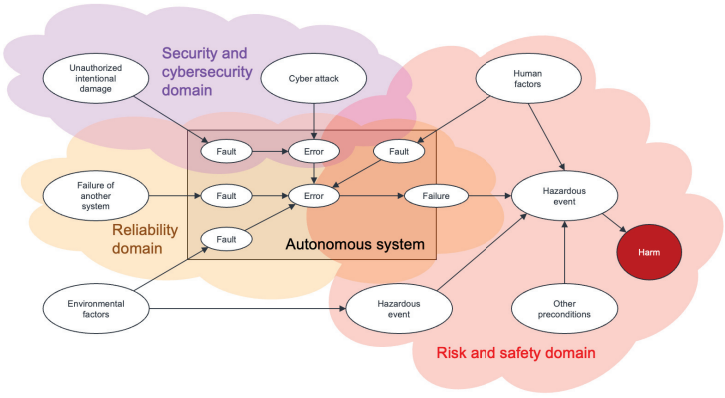

Fig. 1. Interrelation of reliability, security, and safety domains in autonomous systems.

tems from harm, including physical, financial, or informational threats. It focuses on preventing unauthorized access, ensuring resilience against attacks, and maintaining confidentiality, integrity, and availability of resources. ISO/IEC Guide 51 (2011) defines security as *freedom from unacceptable risk*, while Rausand and Haugen (2020) highlights resilience against both intentional and unintentional harm. **Cybersecurity**, as a subset, protects systems and data from digital threats such as hacking, malware, and data breaches through secure design, threat protection, and incident response strategies.

3. Keynotes

IWASS 2024 featured three invited presentations, each offering a unique perspective on the challenges and advancements in autonomous system safety. These presentations set the stage for in-depth discussions.

Human Factors Considerations for Remote and Autonomous Operation of Nuclear Facilities: Niav Hughes Green from the U.S. Nuclear Regulatory Commission discussed the evolving role of human operators in advanced nuclear reactors, focusing on the challenges of remote operation and automation. Key topics included passive safety features, alarm management, and the importance of human factors like situational awareness and communication.

Autonomous Vehicles – A Perspective of Aspiring Peripheries: Krzysztof Wróbel from Gdynia Maritime University highlighted barriers to autonomous vehicle development in Central Europe, such as limited RD, weak business ecosystems, and societal attitudes. He emphasized the need for global collaboration to support innovation in these regions.

Safety Assurance Through the Lens of Continuous Operations: Ryan Yee from Zoox outlined the company's safety assurance approach for autonomous vehicles, emphasizing continuous testing, feedback, and operational safety processes to balance innovation with public safety.

4. Discussions

The IWASS participants were divided into two groups for an in-depth examination of the five topics listed below. Figure 2 maps these five topics (T1-T5) onto the layout of key autonomous system components.

4.1. *Topic 1: Successful Deployment and Incidents of Autonomous Systems*

The first discussion session focused on the progress and current state of autonomous system deployment across different industries. Participants examined whether the development of autonomous systems had met the expectations set five years ago. There was a consensus that certain sectors, such as warehousing and logistics, have

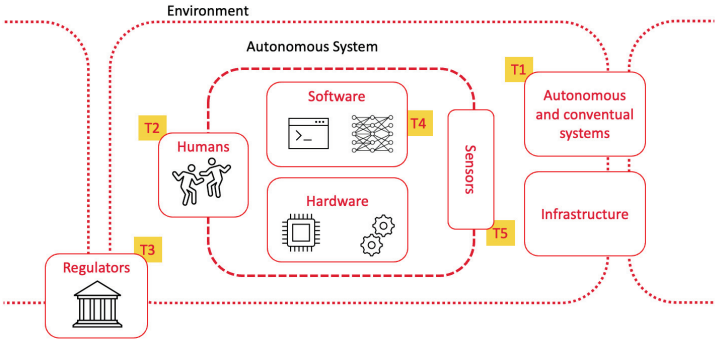

Fig. 2. Five discussion topics of IWASS 2025.

made significant strides in deploying autonomous mobile robots (AMRs). However, in spite of notable exceptions, the transportation and maritime industries lag behind, primarily due to regulatory and operational challenges. The discussion then shifted to safety, security, and reliability concerns. Attendees noted that while these aspects have been addressed to varying degrees, significant gaps remain in building robust safety frameworks. Participants identified automated vehicles as a domain that requires more comprehensive safety measures due to their deployment on public roads and interaction with unpredictable environments. Finally, the groups debated the transferable lessons learned from different industries.

4.2. *Topic 2: Human-Autonomous System Interaction*

This session revolved around the evolving role of humans in autonomous systems. Participants discussed how the relationship between human operators and autonomous systems has changed over time. The group highlighted a shift toward more remote supervisory roles, particularly in maritime and nuclear industries. The debate also touched on the adequacy of current methods for assessing human factors in system design. Several participants argued that existing human error assessment models need to be adapted to account for the complexities of autonomous operations. Concerns about misuse, abuse, and potential malicious intent were also raised, emphasizing the need for robust safeguards and monitoring mechanisms. In conclusion, the discussion underscored the importance of designing systems that support human operators through intuitive interfaces and real-time information.

4.3. *Topic 3: Continuous Safety Assessments*

The third discussion focused on transitioning from design-focused safety assurance to operational safety. Participants debated the role of surrogate safety metrics in real-time deployment decisions. It was agreed that developing metrics to monitor safety performance proactively is critical to mitigating risks and enhancing system reliability. Near-miss analysis emerged as a key topic. The group explored ways to incorporate lessons from near-misses into safety frameworks, emphasizing their potential to prevent future incidents. Participants advocated for a proactive approach to safety, incorporating continuous monitoring and learning from operational data. The conversation also covered the challenges of over-the-air software updates. While these updates offer the advantage of continuous system improvement, they also introduce risks related to system stability and security. Attendees called for rigorous and automated testing protocols and robust continuous verification processes to ensure that updates do not compromise safety.

4.4. *Topic 4: AI and Trust*

The fourth session delved into the complexities of building trust in AI-driven autonomous systems. Participants discussed the need for transparent decision-making processes to foster user

trust. Several attendees pointed out that trust can be built either through the system's adherence to established standards or through trust in the organization developing the system. The groups also debated the balance between trust and situational awareness. While high levels of automation (when correctly implemented) can reduce operator workload, they can also lead to over-reliance on the system, potentially compromising safety in critical situations. Participants recommended integrating explainable AI (XAI) to enhance user understanding of system behavior and foster an optimal level of trust. Finally, the role of AI in safety, reliability, and security was examined. Participants emphasized the need for AI design and certification processes to address potential risks comprehensively. The session concluded with a call for collaborative efforts to develop guidelines and standards for AI-driven systems.

4.5. *Topic 5: Perception and Sensors*

The final discussion session focused on the role of sensor technologies in autonomous systems. Participants examined the critical types of sensors required for safe and reliable operations such as cameras and LIDARs. Advances in sensor fusion and multi-modal perception were highlighted as key enablers for improving system awareness and decision-making capabilities. The conversation then turned to the evolution of sensor technologies over the past five years. While there has been significant progress, challenges remain in ensuring the accuracy and reliability of sensor data. Participants stressed the importance of validating sensor performance under real-world conditions. Finally, the groups discussed methods for detecting and compensating for sensor failures in real-time. Adaptive algorithms and redundancy mechanisms were identified as potential solutions to enhance system resilience.

5. Outcomes

5.1. *Examples of SRS-critical Autonomous Systems*

Before the discussions, several examples from different domains highlighted the current state and challenges of autonomous systems. In the **energy sector**, Japan's Tomoni Point project is developing the world's first autonomous combined cycle power plant, using AI for maintenance, anomaly detection, and operational planning. In the **automotive sector**, Zoox is building fully driverless vehicles for urban mobility, Waymo operates autonomous cars in several U.S. cities using real-time sensor data, and Tesla applies deep learning in its Autopilot system to enable features like automatic steering and smart parking. In the **maritime sector**, Stockholm's Estelle electric ferry combines solar-powered propulsion with advanced navigation and collision avoidance, aiming to transition from operator-assisted to fully autonomous operation, supporting the city's goal of all-electric maritime traffic by 2030. In **robotics**, general-purpose robots like Figure 01 are being developed for industrial tasks, OpenAI is advancing AI for robotic manipulation and automation, and NVIDIA is applying digital twins and AI simulations to optimize industrial workflows and accelerate deployment. These examples informed the following discussions, which are summarized into three main groups of challenges in the subsections below.

5.2. *Operational Challenges*

Autonomous systems face varying levels of adoption across industries, each presenting specific operational challenges. In controlled environments like warehousing, autonomous mobile robots (AMRs) have been rapidly deployed due to clearly defined tasks (Keith and La, 2024; Grover and Ashraf, 2023). In contrast, open environments—such as transportation and maritime sectors—pose significant obstacles, including complex operational conditions, regulatory barriers, and difficulties in achieving seamless human-robot collaboration.

A key challenge across domains is the integration of autonomous systems into environments shared with humans and existing infrastructure. Overly conservative operational rules, while ensuring safety, can limit efficiency and scalability. For example, AMRs often halt when humans are nearby, disrupting workflows, while human-robot interaction in multi-robot industrial environments

remains underexplored (Mehak et al., 2024). In land transportation, inconsistent regional regulations slow the deployment of autonomous vehicles (AVs), particularly in Europe, where stricter legal frameworks and heightened public scrutiny of AV incidents further complicate progress.

The maritime sector faces similar issues, including the absence of standardized communication protocols and regulatory frameworks for crewless vessels (Issa et al., 2022; Osaloni and Ayeni, 2022). Projects like Yara Birkeland and MF Estelle highlight the need for improved satellite communication, automated navigation, and carefully designed human-autonomy control transitions to maintain situational awareness and safety.

Cross-domain insights reveal shared challenges in latency, environmental structure, and infrastructure support: Maritime operations face high system latencies between control actions and observable effects, demanding foresight in decision-making. Infrastructure support is limited, making autonomous perception and planning more difficult. Rail systems also face significant stopping latency, but their structured environments and proactive infrastructure support help reduce the uncertainty involved in making predictions over extended time horizons. Road environments are highly dynamic and less structured, requiring real-time perception and rapid response to unpredictable scenarios. The coexistence of manual-, assisted-, and autonomous driving adds complexity, and uncertainty remains about the scalability and economic viability of full autonomy, particularly regarding the funding of necessary infrastructure. These examples highlight the need for a collaborative, cross-domain approach to address common challenges and adapt operational frameworks to enable the safe and efficient deployment of autonomous systems.

5.3. *Risk Perception and Societal Acceptance*

The perception and tolerance of risk play a pivotal role in the adoption and regulation of autonomous systems. Participants highlighted that societal expectations for the safety of these technologies often exceed those for human-operated systems, creating significant barriers to wider acceptance (Kenesei et al., 2025). This discrepancy underscores the need for transparent communication about both the risks and benefits of autonomy, supported by rigorous safety demonstrations that address not only technical compliance but also societal concerns around fairness and accountability (Naiseh et al., 2024).

A core question emerges: *who decides what constitutes acceptable risk for autonomous systems?* Risk perception is highly subjective, shaped by individual traits, professional backgrounds, and lived experiences. For instance, astronauts—often test pilots or engineers—are trained to manage extreme risks and embrace danger in pursuit of exploration, developing a high tolerance for uncertainty. In contrast, safety officers or risk managers typically take a more conservative stance, prioritizing caution and minimizing exposure to even low-probability hazards. These contrasting perspectives raise the issue of whether acceptable risk should reflect public opinion, expert judgment, political considerations, or the views of those directly engaging with the systems.

Aligning these diverse perceptions with the realities of autonomous system performance remains a key challenge. While technological advancements continually reduce risk, some degree of failure is inevitable. Without a shared understanding of acceptable risk levels, public trust and regulatory approval may lag behind technical progress. This is particularly problematic in reactive regulatory environments, where high-profile incidents drive fragmented and inconsistent policy responses. The maritime industry exemplifies this, with slow adaptation to the unique needs of autonomous vessels limiting scalability and operational scope.

Public perception, amplified by media attention to system failures, further influences regulatory priorities and acceptance. Addressing these challenges requires proactive engagement with stakeholders, clear communication strategies, and educational initiatives that demystify autonomous technologies. Only by fostering a nuanced understanding of risk—one that balances expert knowl-

edge with societal values—can we build the trust necessary for the widespread adoption of autonomous systems.

5.4. *Human-System Interaction*

The evolving relationship between humans and autonomous systems is central to their successful implementation. Participants emphasized the need for intuitive human-machine interfaces (HMIs) that support situational awareness and enable effective decision-making, particularly in high-pressure or time-sensitive scenarios (Hery et al., 2024). In particular, as automation takes over routine tasks, maintaining operator proficiency becomes a growing concern, with risks of skill degradation and disengagement that can compromise safety during manual interventions.

A key solution to these challenges is the integration of **explainable AI (XAI)**. Unlike human errors, which are often accepted as inevitable, mistakes made by autonomous systems face greater scrutiny and lower tolerance. This heightened expectation makes explainability essential not only for user trust and system acceptance but also for ensuring accountability and preventing misconceptions. Operators must understand how AI-driven systems make decisions, particularly in complex environments like maritime and road transport, where these systems are expected to demonstrate ”good seamanship”, exercise ”sound judgment”, and determine an appropriate ”safe speed” based on the specific circumstances. Providing clear justifications for system actions—particularly when deviating from standard procedures to avoid hazards—is crucial for maintaining trust and meeting legal and ethical standards.

Shared control between humans and machines also requires careful design. Balancing human cognitive strengths with machine-driven data processing can optimize system performance but demands thoughtful task allocation, timing, and safeguards against automation complacency. This becomes even more critical in remote operations, where delayed responses and reduced situational awareness can hinder effective decision-making. Systems must therefore deliver real-time, context-rich information to keep operators engaged and prepared to intervene when needed.

6. Conclusion

IWASS 2024 highlighted the growing complexity of ensuring the safety, reliability, and security of autonomous systems across industries. The discussions emphasized the need for collaborative, interdisciplinary efforts, supported by adaptive safety assurance frameworks, continuous learning, real-time monitoring, and anomaly detection.

The role of human operators in the oversight of autonomous systems remains a topic of significant debate. While automation offers the potential to reduce human error and enhance efficiency, maintaining a balance between human and machine contributions is crucial. This requires the design of intuitive interfaces and the provision of comprehensive training programs to ensure that operators can effectively intervene when necessary.

Furthermore, the workshop emphasized the need for clear and consistent regulatory frameworks to support the safe deployment of autonomous systems. Policymakers must work closely with researchers and industry stakeholders to establish guidelines that promote innovation while safeguarding public trust and safety.

Looking forward, key priorities include advancing real-time risk assessment, improving explainable AI, and strengthening international cooperation to address regulatory challenges.

Declaration

The authors used ChatGPT 4o and Grammarly in order to improve parts of the text for better readability. After using these tools/services, the authors reviewed and edited the content and are taking full responsibility for the publication.

References

Aven, T. (2012). The risk concept—historical and recent development trends. *Reliability Engineering & System Safety 99*, 33–44.

Aven, T. (2014). What is safety science? *Safety Science 67*, 15–20. The Foundations of Safety Science.

Avizienis, A., J.-C. Laprie, B. Randell, and C. Landwehr (2004). Basic concepts and taxonomy of dependable and secure computing. *IEEE Transactions on Dependable and Secure Computing 1*(1), 11–33.

Correa-Jullian, C., J. Grimstad, S. A. Dugan, M. Ramos, C. A. Thieme, A. Morozov, I. B. Utne, and A. Mosleh (Eds.) (2023). *Proceedings of the International Workshop on Autonomous Systems Safety (IWASS 2023)*, Southampton, UK.

Correa-Jullian, C., J. Grimstad, S. A. Dugan, M. Ramos, C. A. Thieme, A. Morozov, I. B. Utne, and A. Mosleh (Eds.) (2024). *Proceedings of the International Workshop on Autonomous Systems Safety (IWASS 2024)*, Cracow, Poland.

Grover, A. K. and M. H. Ashraf (2023). Leveraging autonomous mobile robots for industry 4.0 warehouses: a multiple case study analysis. *The International Journal of Logistics Management 35*(4), 1168–1199.

Hery, A., O. Joseph, O. Femi, and H. Luz (2024). Human-machine interface and decision-making in autonomous driving.

IEC61508 (2010). Functional safety of electrical/electronic/programmable electronic safety-related systems.

ISO12100 (2010). Safety of machinery — general principles for design — risk assessment and risk reduction.

ISO26262 (2011). Road vehicles – Functional safety.

ISO/IEC Guide 51 (2011). Guide 51: Safety Aspects - Guidlines for their inclusion in standards.

Issa, M., A. Ilinca, H. Ibrahim, and P. Rizk (2022). Maritime autonomous surface ships: Problems and challenges facing the regulatory process. *Sustainability 14*(23), 15630.

Kaplan, S. and B. J. Garrick (1981). On the quantitative definition of risk. *Risk analysis 1*(1), 11–27.

Keith, R. and H. M. La (2024). Review of autonomous mobile robots for the warehouse environment. *arXiv preprint arXiv:2406.08333*.

Kenesei, Z., L. Kökény, K. Ásványi, and M. Jászberényi (2025). The central role of trust and perceived risk in the acceptance of autonomous vehicles in an integrated utaut model. *European Transport Research Review 17*(1), 8.

Mehak, S., I. F. Ramos, K. Sagar, A. Ramasubramanian, J. D. Kelleher, M. Guilfoyle, G. Gianini, E. Damiani, and M. C. Leva (2024). A roadmap for improving data quality through standards for collaborative intelligence in human-robot applications. *Frontiers in Robotics and AI 11*, 1434351.

Naiseh, M., J. Clark, T. Akarsu, Y. Hanoch, M. Brito, M. Wald, T. Webster, and P. Shukla (2024). Trust, risk perception, and intention to use autonomous vehicles: an interdisciplinary bibliometric review. *AI & SOCIETY*, 1–21.

Osaloni, O. S. and V. O. Ayeni (2022). The development of maritime autonomous surface ships: regulatory challenges and the way forward. *Beijing L. Rev. 13*, 544.

Rausand, M. and S. Haugen (2020). *Risk assessment: theory, methods, and applications*. John Wiley & Sons.

Thieme, C. A., A. Morozov, I. B. Utne, and A. Mosleh (Eds.) (2019). *Proceedings of the First International Workshop on Autonomous Systems Safety (IWASS 2019)*, Trondheim, Norway.

Thieme, C. A., M. A. Ramos, I. B. Utne, and A. Mosleh (Eds.) (2021). *Proceedings of the International Workshop on Autonomous Systems Safety (IWASS 2021)*, Online.

Thieme, C. A., M. A. Ramos, I. B. Utne, and A. Mosleh (Eds.) (2022). *Proceedings of the 3rd International Workshop on Autonomous Systems Safety (IWASS)*, Dublin, Ireland.