Operability and Functionality within Advanced Cyber Systems

Anthony R. Valiaveedu[a]
[a] Massachusetts Institute of Technology, Cambridge, United States, arv7@mit.edu

Abstract: The current push towards increasingly digital ecosystems with the advent of Artificial Intelligence has challenged existing practices for risk and reliability analysis writ-large. Existing regulations are currently being scrutinized for applicability and ensuring a level of safety equivalence towards prior technology. One large push by multiple industries has been the significant investment into developing complex algorithms as assurance measures for AI-integrated cyber systems. Though useful, such approaches face challenges as the AI field is rapidly evolving in terms of architecture, infrastructure, and integration prospects. Instead, a potential starting point may be better focused on questioning the fundamentals of operability and functionality for such systems. This approach can provide risk-informed flexibility in engineering decision-making while maintaining a level of safety equivalency. This paper explores the differences between functionality and operability as it relates to varying technologies and the implication of AI as a dependent failure mode, rather than existing paradigms of independence. Finally, the implication of these results will be drawn to AI systems.

1. INTRODUCTION

The rapid advancement of artificial intelligence and cyber technologies has significantly transformed safety-critical industries in recent years. As novel cyber systems become increasingly integrated into high-stakes operational environments, they are adopted under the premise of enhanced reliability, improved performance, and reduced deployment costs. Industries ranging from nuclear power generation and aerospace to medical devices and autonomous transportation systems have begun embracing these technologies at an accelerating pace, fundamentally reshaping how engineers and operators approach system design, deployment, and oversight.

However, this technological evolution has not come without consequence. The increasing complexity and interconnectedness of modern cyber systems has placed growing and unprecedented demands on engineers tasked with reviewing these systems for mission assurance. Traditional review methodologies, developed largely in an era of analog and rules-based systems, are being challenged to keep pace with the rapid iteration cycles, opaque decision-making processes, and emergent behaviors that characterize modern cyber and AI-enabled systems. As a result, the engineering community faces an urgent need to develop robust, adaptable, and rigorous frameworks capable of addressing these challenges.

To support the review efforts of this technology, examination of *functionality* and *operability* can provide mission critical insights towards realizing the larger safety question in operations. These two concepts provide mission-critical insights that ultimately inform broader safety determinations in operational contexts. Together, they form the analytical foundation upon which engineers can systematically evaluate whether a given system is not only capable of performing its intended purpose, but also whether it can do so safely, consistently, and within acceptable risk thresholds. Understanding the distinctions and interplay between these two concepts is therefore critical to any meaningful safety assurance effort in the modern technological landscape.

1.1. Operability

Operability provides a deterministic definition of safety. In a systematic sense, the term reviews the *operational* profile of the system in question by meeting specific goals. Aligning closely with

deterministic profiles, operability enables the engineer to quickly assess and define clear prescriptive requirements when encountering technology development. By establishing explicit thresholds and requirements, operability assessments enable engineers to make confident, well-supported determinations about a system's readiness for deployment and continued operation.

This deterministic foundation is particularly well-suited for evaluating highly reliable system components, such as software architectures and physical infrastructure, where consistent and predictable performance is well-established. In the context of software systems, for example, operability assessments might evaluate whether a program executes in its intended operation profile within acceptable parameters. For physical systems, such as pipelines or structural components, operability may involve verifying that materials and configurations meet prescribed design standards and can withstand anticipated operational stresses without degradation or failure.

Operability assessments are typically structured around the development and evaluation of an operational profile, which defines the conditions, inputs, and performance expectations that characterize normal and abnormal system operation. This profile serves as the benchmark against which system behavior is measured, providing a systematic and repeatable basis for evaluation. Developing a comprehensive operational profile requires close collaboration between engineers, operators, and stakeholders, as it must accurately reflect the full range of conditions the system is expected to encounter throughout its operational life.

This goal-oriented conception of operability closely mirrors modern assurance mechanisms being advanced across multiple industries. International standards committees and various national regulatory bodies have increasingly advocated for goals-based design and assurance approaches to be adopted broadly. These frameworks argue that outcomes-oriented approach is necessary to achieve meaningful safety assurance. This push has largely been witnessed recently with the International Maritime Organization’s development of a standard for autonomous maritime surface ships. [1]

Within this broader regulatory context, operability serves as a critical bridge between high-level safety goals and the specific technical requirements that govern system design and operation. By translating abstract safety objectives into concrete, measurable operational criteria, operability assessments provide engineers and regulators with a practical and defensible mechanism for evaluating system safety. This is particularly valuable in the context of emerging technologies such as AI and advanced cyber systems, where the absence of established precedent and standardized evaluation methodologies can make safety assessments especially challenging.

Furthermore, the application of operability frameworks to cyber systems introduces unique considerations that must be carefully addressed. Unlike traditional mechanical or electrical systems, cyber systems can exhibit complex, non-linear behaviors that may not be fully anticipated during the design phase. Software bugs, cybersecurity vulnerabilities, and unexpected interactions between system components can all undermine operability in ways that are difficult to detect and characterize using conventional assessment methods. As a result, operability reviews for cyber systems must be designed with sufficient rigor and depth to account for these unique challenges, incorporating techniques such as formal verification, model-based testing, and adversarial analysis to provide comprehensive coverage of potential failure modes.

1.2 Functionality

Unlike operability, functionality provides a probabilistic definition of safety. A definition can similarly be obtained for functionality as meeting requirements for achieving the *functional* profile of the system by evaluating the probabilistic metrics associated with completing a function. This aligns closely with efforts of risk-based practices that establish probabilistic profiles for meeting design requirements.

The probabilistic nature of functionality assessments reflects a fundamental recognition that real-world systems operate in environments characterized by inherent uncertainty. No system can be designed to perform perfectly under every conceivable condition, and the goal of functionality assessment is not to eliminate risk entirely but rather to characterize and manage it in a systematic and defensible manner. By establishing probabilistic performance profiles for key system functions, engineers can identify areas of elevated risk, prioritize mitigation efforts, and demonstrate that residual risks have been reduced to acceptable levels.

Functionality assessments are typically grounded in the development of a functional profile, which defines the specific tasks and objectives that the system is required to accomplish, along with the performance metrics and acceptance criteria by which success is measured. This profile must account for the full range of behaviors the system may encounter, including normal operations, off-normal conditions, and potential failure scenarios. Developing a robust functional profile requires extensive engagement with stakeholders, operators, and subject matter experts, as well as access to relevant operational data and historical performance records. Engineers and regulatory authorities have strongly championed the functionality-based methodology as a means of developing a more comprehensive and realistic understanding of system design and performance.

Achieving meaningful assurance through a functional lens requires substantial engagement in requirements definition and data-driven analysis. The quality of a functionality assessment is fundamentally dependent on the quality and completeness of the data and models used to characterize system performance. This places significant demands on organizations developing and deploying safety-critical cyber systems, requiring them to invest in robust data collection and management capabilities, rigorous modeling and simulation tools, and systematic processes for updating and refining functional assessments as new information becomes available.

The application of functionality frameworks to AI and cyber systems presents both significant opportunities and notable challenges. On one hand, the data-driven nature of AI systems makes them well-suited to probabilistic performance characterization, as large volumes of operational data can be used to develop detailed statistical models of system behavior. On the other hand, the complexity and opacity of many AI algorithms make it difficult to fully characterize their functional profiles, particularly in edge cases and novel operational scenarios not well-represented in available training data. Addressing these challenges requires the development of new methodological approaches specifically tailored to the unique characteristics of AI and cyber systems, including techniques for uncertainty quantification, robustness testing, and interpretability analysis.

1.3 Advanced Cyber Systems

Efforts to develop assurance mechanisms in the cyber-domain have largely pushed towards the use of deterministic efforts due to the inherent design attribute of software as a static system. This assumption has been increasingly challenged at the present with the incorporation of AI-incorporated cyber systems. Especially with the proliferated use of Neural Network logics, the causal interaction within software is a key concern in the validation of such systems. Efforts have been developed to describe such system either as “Black Boxes” or “White Boxes”.

Black box analysis for such systems creates focus areas based around the perturbing the system’s features to assess the level of certainty the system undergoes. [2] Similar to Schrodinger’s cat, the use of the black box methodology treats such cyber system to be controlled through indirect efforts (i.e., evaluating the input and output operations). These efforts retain the level of uncertainty within the system but provides opportunity for enabling complex system operations. Black box usage can be witnessed with recent efforts on the implementation of RTA systems and other variable-structure controllers to enable assurances within operational envelopes, while enabling reliability. [3] However, the emphasis of Black-box operations lends itself into a slippery-slope in terms of complete assurance due to potential of edge cases and unrecognized operational interactions.

White box analysis, in contrast, attempts to deconstruct the AI system to capture the full internal interactions by the cyber system. This operation is common in the case of traditional cyber system architectures due to the deterministic attribute of software. Advanced cyber interactions introduce a layer of non-determinism which has introduced new technologies like "Explainable AI". White boxing requires transforming the non-deterministic software design into a largely deterministic design. Aside from the significant cost burden associated with its transformation, it also may require the loss of non-deterministic attributes within the system and, as a result, the system's flexibility. Non-deterministic system attributes may be beneficial in the context of system performance measures and thereby be hindered due to such requirements.

In addition, cyber systems include aspects of *communication* and *actions*. These items can influence the level of assurance necessary for reliable systems. Communication pathways include the physical domain of Instrumentation and Controls, but also the non-physical and more emerging network system that uses signal-based communication pathways. Due to the level of independence that such systems contain Actions consider the outcome from the advanced cyber system. It can be collision avoidance (in the context of transportation) or even control rod movements (in the context of nuclear power).

2. EXISTING EFFORTS

Due to the prevalent nature of development many safety critical domains have developed guidance or, in certain cases, rulemaking to define assurance requirements for advanced cyber systems.

2.1 Automotive

The automotive industry has recently published ANSI/UL-4600 that establishes safety for AI-incorporated autonomy through using safety cases to justify the system's design, coupled with Safety Performance Indicators to monitor deviations from "safety". [4] The safety case contains the following topics for incorporation:

- Safety Case and Arguments
- Risk Assessment
- Interactions with Human and Road Users
- Autonomy Functions and Support
- Software and System Engineering Processes
- Dependability
- Metrics and SPIs
- Assessment of Conformance

More interestingly is the view towards automation by the industry. From the US regulatory side, the National Highway Traffic Safety Administration isolates Automated Driving Systems to improve safety by reducing/eliminating human error processes. [5] Specific focuses are set upon defining and bounding an operational design domain, while evaluating system response in event of excursions (i.e., fallbacks).

2.2 Nuclear

The nuclear community has been comparatively quiet in terms of deployment of advanced cyber systems. Recently this focus has been on the application of Digital Instrumentation and Control (I&C) Systems and Digital Twins, but deployment has been sparse in the industry. Recent advanced reactor systems have incorporated digital systems within nuclear systems and modernization efforts are becoming more prevalent within the existing fleets. Digital twins are also becoming incorporated for the purposes of preventive maintenance procedures and development.

For emerging technology concepts, this consideration is included in the Nuclear Regulatory Commission's Part 53 rulemaking. Though not focused on cyber, the "Self-Reliant-Mitigation Facility"

process is created to focus on the development of reduction of operator action and reliance on internal system automation. Likewise, the UK Office for Nuclear Regulation moved towards sandboxing such activities for implementation of AI system integration. [6]

2.3 Maritime

Maritime domain activities have looked at advanced cyber systems largely in the context of Maritime Autonomous Surface Ships (MASS). Regarding this topic, the International Maritime Organization has been developing regulations focused on the developing a code on assuring safety of such system. [1] This development has also been done in parallel with domestic developments, such as Det Norske Veritas which has created test cases to evaluate autonomous navigation systems. [7]

2.4 Aerospace

Aerospace has updated its existing DO-178 standard with version C to consider the use of fallback and deterministic mechanisms to maintain reliable operations. [8] In addition, Run-Time Assurance methods have been researched upon for developing a separate "safety monitor" system to operate on top of black-box systems. [3] This focus has been developed to separate the question of safety to reliability, as the safety monitor creates a bounded operational zone for optimization.

3. DESIGNING SAFETY

Given the current state of development, re-evaluating operability and functionality in such systems require reviewing what it means for such systems to achieve operability and functionality. While traditional deterministic principles have been largely deployed in cyber systems, advanced cyber systems involve higher degrees of non-deterministic design development and less reliable operations from solely rules-based software assurance mechanism.

3.1 Operability vs Functionality in Advanced Cyber Systems

Moving back to the question of operability vs functionality of advanced cyber systems, current methodologies target that such systems have only a defined operability standard. Considering the modern shift into non-deterministic operations, this view should be appropriately shifted. This interaction of the cyber system is illustrated in Figure 1.

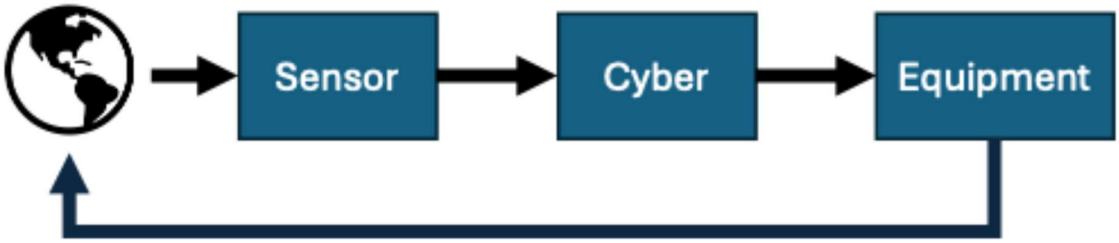

Figure 1. Connection between Cyber System to other sub-systems

Questions of operability will likely retain its concept of input/output interface used in traditional system design. In a control's mindset, it targets the focus to be on the aim to be on how the cyber operation signal the output equipment given various input feed. These requirements can be outlined and defined through the following format in Table 1. Developing a detailed operability list can support developers in assuring their system is well defined and controlled for system.

Table 1. Software Operability Requirements

Input	Sensor	Output	Equipment
Temperature above 70	Readout from Thermistor A	Temperature below 70	Energize Ventilation System

Similarly, a functional aspect can be considered within the system. Profile analysis of the cyber system's input and output feedback can be incorporated to understand the nuanced operation for a *functionally* comprehensive safety case. Consider the attributes of a static Deep Neural Network (DNN) system. While cyber system operates as a non-linear function, attributes can still be placed on the system for mapping its profile. Bounded *input* evaluations across appropriate value frames can be evaluated based on distribution curves that lends itself well for incorporating into PRA models. In this sense, a beta distribution can still be used for depending on variation of input. A benefit of this approach is that predictive software performance can still collected en masse providing potentially better insights than conducting reliability analysis of the physical components. The reliability of sensor can then be used to develop the testing conditions for the DNN. An example of this is tabulated in Table 2.

Table 2. Failure Probability Collection

Input Feature	Operating Cycles	α	β	Error Factor
70-72				
72-75				
75+				

However, another attribute for potential advanced cyber architecture includes non-static architectures that adjust over operational life. Described as "Self-Learning", certain models contain a feedback loop to assist in re-calibration to account for sensor and equipment drift, as well as adjusting internal functions for reducing output error. In such systems, the evaluation of operability will likely remain unchanged, dependent on the self-learning architecture, but it will be increasingly important to map out the adjustment during each iteration of self-learning compared to the initial operation.

It is especially prevalent in non-static cyber architectures that the concept of "Guardian Angels" to be especially of value. In such system design, the development of a parallel deterministic system can reduce the burden of detailed evaluation of the "disruptive" portion of the system. Such design is similar to a Variable Structure Control system, where the system has a switch (either physical or software based) that adjusts between the Guardian Angel system and the operating system. Figure 2 illustrates this process through a behavior-based approach. Depending on the behavior of the system, normal operations can continue based on the certainty of the operating regime. However, once the level of risk has increased based on the system's design, various Guardian Angels may active (e.g., emergency diesels, Reactor Protection System, etc.). In this process, the level of assurance can be boosted by the system so that the operating cyber system can be optimized for utility, while issues of safety can be maximized within traditional systems.

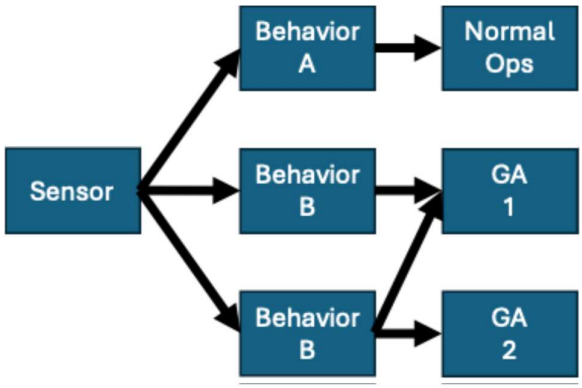

Figure 2. Behavior Based Approach for Guardian Angels

3.2 Fault Areas

To evaluate the risk to a given deployment, understanding the specific areas through which failures can manifest is essential for engineers conducting safety assessments. Using a command-and-control perspective, five critical performance areas are identified as the primary fault domains for such systems operations: Prediction, Detection, Monitoring, System Health, and Human Factors. Each of these areas contains unique failure modes that must be systematically evaluated when assessing deployments in safety-critical contexts. Special attention is given to the inherent trade-offs between these performance measures, as these conflicts often represent the most significant latent risks in deployment.

Prediction measures describe the capability of cyber system to generate estimations for physical conditions and future system states based on the various sensor measurements integrated with the system. Two primary fault modes are identified for prediction functions: over-prediction and under-prediction. These fault modes manifest primarily as bias measurements within the system output. Careful calibration procedures are necessary to establish performance distributions appropriate to the specific operational environment, analogous to those developed for nuclear components in established regulatory practice.

Each of these fault modes can result in a detection failure with potentially significant safety consequences. Where a hazard is not identified or signaled, the system fails to respond to an underlying risk entirely, effectively eliminating the protective function. Where a hazard is misclassified, the detection function provides erroneous information that can lead to a worsened corrective response. For example, in a maritime context, if an iceberg is misclassified as floating debris, the system may increase speed rather than initiate avoidance maneuvers. The use of detection capabilities has been particularly prevalent in machinery fault detection and building energy systems, providing a reference basis for evaluating detection fault modes in deployments.

1. Failure to identify a hazard — the system does not recognize the presence of a condition requiring response.
2. Failure to correctly classify a hazard — the system identifies a condition but assigns it an incorrect classification, leading to an inappropriate corrective response.
3. Failure to signal — the system correctly identifies and classifies a hazard but fails to generate or transmit the appropriate alert or command.

Failure to establish operational limits may occur either statically, as in the case of pre-defined Limiting Conditions of Operation analogous to those used in nuclear plant technical specifications, or dynamically, where limits must be adjusted in response to varying input conditions such as sea state in maritime operations. Failure to track system changes concentrates fault modes related to the integration of the various data feeds throughout the system, including feedback pathways such as motor speed adjustments driven by object detection outputs and general integration functions such as the fusion of data from multiple sonar detectors. Outlining monitoring functions as a discrete

subsystem assists in mapping the command-and-control network for deployments and provides engineers with a structured basis for identifying integration-level fault modes.

1. Failure to establish operational limits — the system does not define adequate static or dynamic thresholds governing its operational boundaries.
2. Failure to track system changes — failures in the integration of various data feeds throughout the cyber system, including feedback mechanisms and multi-sensor integration, result in an inability to accurately characterize the current system state.
3. Failure to maintain limits — the system fails to prevent exceedance of established operational boundaries, independent of whether those limits were correctly established.

System health fault modes are particularly consequential in fully autonomous deployments where no human operators are present to provide corrective intervention. Repair and maintenance strategies can be conducted via human operators in lower autonomy configurations, but efforts toward fully autonomous systems will require these functions to be performed without human involvement. A relevant example can be found in autonomous cybersecurity infrastructure, where AI-driven techniques may devise and execute strategies to counter intrusions without human agent involvement. Emergency operating procedures within this fault domain include responses designed to bring the system to a defined safe state through the activation of additional equipment or protective mechanisms, including kill switches and fallback control modes. The wide capture area of system health fault modes enables this domain to account for a diverse set of failure scenarios across the full cyber system's operational lifecycle.

Likewise, human factors remain a consequential aspect of operations. As these cyber systems advanced towards higher complexities and non-linearities, the opaqueness of understanding can negatively affect system reviews and consequences of manual actions. Clearly developing system Input/Output diagrams (i.e., Table 1), while creating explicit interlocks can assist with mapping out system interactions.

3.3 Cross-Cutting Attributes

Cross-cutting attributes differ from performance-area fault modes in that they are not confined to any single operational function but instead reflect systemic qualities of the cyber system that, when deficient, can compromise the integrity of the entire system. Identifying such attributes is important because a failure in a cross-cutting area will require a deeper evaluation process of the cyber system, rather than a targeted remediation of a specific subsystem. Cross-cutting attributes are organized into two primary categories reflecting the fundamental components of advanced cyber system: Datasets and Models. This division is grounded in the recognition that AI development can be generally classified as a data-driven modeling scheme, and that both the data and the model must be evaluated as an integrated unit rather than in isolation.

Datasets represent a critical foundation of any AI system, as deficiencies in the data used to develop and operate the AI can threaten the integrity of the entire AIAS deployment. The absence of a data-centric approach in AI design has been commonplace in governance and evaluation schemes, representing a significant gap that engineers must actively address. Two primary areas of concern are identified within the dataset category.

Data Pipeline addresses the selection and appropriateness of data sources used in developing the AI model. The pipeline must remain representative of operational conditions to ensure system effectiveness throughout the deployment lifecycle. Data Integration encompasses the handling, cleaning, and structuring of data as it is prepared for use within the model, including considerations of dataset sizing, feature representation, and the management of training, testing, and validation splits. Integration failures can result in models that perform well under controlled evaluation conditions but degrade significantly when exposed to the full range of operational scenarios encountered in deployment.

Separate from data considerations, model-level cross-cutting attributes focus on the assumptions, architecture, and operational outputs of the cyber system itself. The design philosophy presents differing goals compared to traditional machine learning design.

Table 2: Cross-Cutting Aspects for Datasets and Models

CODE	DESCRIPTION
D.1	Dataset is inclusive of operational parameters.
D.2	Dataset relates to the physical design of operations.
D.3	Data preparation sets are representative of the system.
D.4	Dataset is inclusive of rare events.
D.5	Self-Learning reinforces safety.
M.1	Model is appropriate for application.
M.2	Metrics represent the efficacy of the model.
M.3	Model is evaluated for operational conditions.
M.4	Architecture has established model limits.
M.5	Model performs in integrated ecosystem.

4. CONCLUSION

The rapid proliferation of AI-incorporated advanced cyber systems across safety-critical industries has exposed meaningful gaps in existing deterministic assurance frameworks that were developed for an era of static, rules-based software architectures. This paper has argued that a re-examination of the fundamental concepts of operability and functionality provides a practical and risk-informed starting point for addressing these gaps. By distinguishing operability as a deterministic evaluation of system input/output behavior from functionality as a probabilistic characterization of system performance across operational profiles, engineers and regulatory reviewers gain complementary analytical tools that, together, can support comprehensive safety cases for non-deterministic AI-incorporated systems. The identification of five critical performance fault areas — Prediction, Detection, Monitoring, System Health, and Human Factors — and ten cross-cutting attributes spanning dataset integrity and model design provides a structured framework through which the command-and-control disruptions introduced can be systematically evaluated. Notably, the cross-cutting attributes that underscore AI cannot be assessed in isolation from the data that drives it, and that deficiencies at the systemic level will propagate across all performance domains regardless of the quality of any individual subsystem. Taken together, these frameworks offer engineers and regulators a flexible, scalable foundation for conducting safety evaluations of advanced cyber systems during early-stage development, supporting risk-informed decision-making while preserving the safety equivalency demanded by the operational environments in which these technologies are increasingly being deployed.

Acknowledgements

The author would like to acknowledge Dr. Curtis Smith for his insights and review of the paper. This paper draws from and builds off of the author's upcoming release "Designing for Certainty in an Uncertain World: A Framework for Evaluating AI-incorporated Autonomy and Other Unknowns".

References

[1] Maritime Safety Committee. "Proposed Base Text for the MASS Working Group", MSC 111/5/3, International Maritime Organization, (Jan. 28, 2026)
[2] A. K. Kalusivalingam, et al., "Leveraging SHAP and LIME for enhanced explainability in AI-driven diagnostic systems," *International Journal of AI and ML*, vol. 2, no. 3, (2021).
[3] P. Nagarajan, S. K. Kannan, C. Torens, M. E. Vukas, and G. F. Wilber. "ASTM F3269 - An Industry Standard on Run Time Assurance for Aircraft Systems", in *AIAA Scitech 2021 Forum*, American Institute of Aeronautics and Astronautics, (Jan. 2021). doi: 10.2514/6.2021-0525.

[4] ANSI/UL Standards Technical Panel 4600. “ANSI/UL 4600: Standard for Safety for the Evaluation of Autonomous Products”, UL Standards & Engagement, (2023).
[5] NHTSA. “Automated Driving Systems 2.0: A Vision for Safety”, Tech. Rep., Department of Transportation, (Sept. 2017).
[6] UK Office for Nuclear Regulation. “Regulators’ Pioneer Fund (Department for Science, Innovation and Technology): Pilot of a regulatory sandbox on artificial intelligence in the nuclear sector”, ONR/Environment Agency Report, (2023).
[7] DNV Maritime Schema. “Maritime Schema: Open Formats for Maritime Collision Avoidance Testing”, DNV Open Source, (2025), available at https://dnv-opensource.github.io/maritime-schema/.
[8] L. Rierson. *Developing Safety-Critical Software: A Practical Guide for Aviation Software and DO-178C Compliance*, CRC Press, Hoboken, (2013). ISBN: 978-1-4398-1369-0.