A FUNCTIONAL SAFETY APPROACH TO SUPPORT MARITIME AUTONOMY

A Barker MSc CEng FSP, I Groom Cdre MBE (Rtd.) CEng FIMarEST and **T Hussain**, Safeguard Engineering Ltd, UK.

SUMMARY

Maritime autonomy offers significant potential benefits for operational efficiency, safety and strategic advantage across both commercial and defence applications. However, realising maritime autonomy invokes substantial assurance challenges, particularly when integrating artificial intelligence and machine learning into safety-critical functions.

This paper explores a functional safety approach tailored to support maritime autonomy, building upon established safety standards while addressing the limitations of applying these traditional methods to artificial intelligence technology, given the dynamic and emergent behaviour characteristics of autonomous maritime operations. An illustrative example of a collision avoidance system is provided. Through this example, this paper outlines the combination of quantitative technical measures and qualitative organisational and governance considerations for the assurance of maritime autonomy through a lifecycle approach. This approach provides a structured, standards-aligned evidence stream specifically targeting artificial intelligence technology within maritime autonomous systems and complementing existing methods such as Systems Theoretic Process Analysis.

1. INTRODUCTION - An Overview of Maritime Autonomy

Levels of autonomy describe the extent to which an Artificial Intelligence (AI) system functions independently of human supervision and control. In the maritime sector, the International Maritime Organization (IMO) introduced the Degrees of Autonomy (DoA) framework as part of its scoping exercise for Maritime Autonomous Surface Ships (MASS) [1]. This framework categorises maritime levels of autonomy on a scale from fully crewed ships with some automated functions (Degree 1) to completely autonomous vessels operating without crew or remote human intervention (Degree 4).

Following this, the MASS industry has developed complementary Levels of Control (LoC) frameworks as an alternative to DoA [2]. The LoC framework ranges from fully crewed, e.g., 'traditional', vessels (LoC 0) to fully autonomous vessels (LoC 5) and has been adopted by the United Kingdom Defence Maritime Regulator (DMR) through UK defence guidance on regulating maritime autonomous systems [3].

In addition, classification societies have introduced further frameworks to guide certification levels for autonomous platforms [4]. For example, Lloyd's Register (LR) has created an Autonomy Level (AL) scale (AL0 to AL6) [5] while Det Norske Veritas (DNV) have established the Autonomous Readiness Operational Safety (AROS) framework to defines class notations with Modes of Operation (MOO) ranging from Remote Control (RC) to Full Autonomy (FA) [6].

The abundance of similar, but slightly different, frameworks for levels of maritime autonomy reflects the fundamental challenges in achieving widespread acceptance of a single autonomy framework; any chosen framework must be applicable to all stakeholders, and both current and future systems, resulting in constant evolution and adaptation [7]. We recognise that regardless of the framework used, whether defined as a degree, level, or scale, maritime autonomy is widely viewed as a critical enabler for future commercial and defence capabilities:

- The UK's Strategic Defence Review (SDR) highlights autonomy and AI as essential to achieving national defence objectives [8];
- The UK's maritime 2025 roadmap identifies the importance of smart shipping and autonomy in delivering a cleaner, safer, and more efficient maritime industry [9]; and,
- The Centre for Assuring Autonomy (CfAA) at the University of York shows that integrating digital technologies, AI, and autonomous navigation systems promises significant efficiency gains, cost reductions, and safety improvements [9].

Full autonomy then, under whichever framework is applied (e.g., 'Degree 4', 'LoC 5', 'AL6' or 'FA') is not merely a technological ambition. It represents a foundational capability required to support UK's defence and commercial

maritime goals [11]. A natural question arises - if the benefits are so significant, why are Maritime Autonomous Systems (MAS) (Covering MASS and subsurface systems) not yet commonplace, replacing traditional crewed vessels at scale across the UK maritime sector?

Part of the answer lies in the complexity of the assurance challenges; assurance frameworks have not evolved rapidly enough to keep pace with technological advances, and they do not fully address the complex challenges posed by AI, Machine Learning (ML), and cybersecurity unique to autonomy [12, 13]. The existing 'traditional' assurance approaches assume deterministic and repeatable systems, or rely heavily on human operators for risk reduction, and thus are not suited to MAS [14].

While several autonomous vessels have been developed and trialled, most have relied on case-by-case, goal-based approaches within limited operational envelopes. Such bespoke approaches remain necessary in the absence of an overarching, standardised framework. The IMO's MASS Code is still under development, with adoption of a mandatory code expected by 2032 [15]. This paper does not propose abandoning goal-based approaches, which are fundamental for complex systems [16]. Instead, we outline an approach to adapt existing robust standards tailored specifically for autonomous systems. This approach aims to shift away from bespoke cases-by-case approaches towards greater harmony, auditability and scalability; supporting both industry and regulators and filling the current assurance gap.

We are also conscious that, in the near term, most MAS will likely operate in hybrid or transitional modes. These systems will blend supervised autonomy with remote or direct control, adjusting between levels of autonomy even during a single voyage. This, in turn, demands an incremental approach to assurance and a gradual expansion of the operational risk envelope. Accordingly, this paper proposes a pathway to navigate through the levels of autonomy (e.g., the various degrees, levels, and scales) to ultimately enable fully autonomous operations.

Finally, we wish to stress an important point. Our approach does not suggest an alternative to developing a safety case to provide overall assurance of 'safe' through a goal-based approach. Safety cases remain the foundational means of demonstrating system safety across many critical sectors, including nuclear, defence and aerospace [18]. There is growing interest in safety cases for AI systems, although much of the existing work focuses on preventing large-scale risks in general-purpose AI rather than safety-critical, integrated applications [19]. We recognise this and therefore present our approach as a method for developing the specific evidence required for the autonomous, AI, aspects of a maritime platform. This AI-related evidence will form an essential part of the overall safety case.

2. ASSURANCE APPROACHES FOR MARITIME AUTONOMY

2.1 DEFINITIONS AND EXISTING 'TRADITIONAL' APPROACHES

Autonomy is defined as a system capable of "modifying its intended domain of use or goal without external intervention, control, or oversight" [20]. Achieving autonomy requires AI technology, being the "technology that implements or enables an AI system" [20], consisting of:

- AI components, defined as "functional elements that construct an AI system" [20]; and
- AI systems, defined as "engineered systems that generate outputs" [20].

Assuring a MAS therefore requires assurance at multiple levels: the individual AI components, the integrated AI systems, and the resulting AI technology providing the autonomous behaviours.

In MAS, these AI components and AI systems will not function in isolation. By design, they will be required to operate in harmony with traditional hardware and software and interact with human operators throughout different operational stages (such as launch, recovery, and maintenance) and across the entire lifecycle. As a result, assurance depends not only on the technical properties of individual components (AI or otherwise), but also on the interactions, emergent behaviours, and broader social and ethical implications across the entire system lifecycle.

There is already a robust body of knowledge on how to avoid failures and demonstrating 'safe' in traditional hardware and software (e.g., non-AI) [21], and many established frameworks and standards are available for the assurance of complex systems. For example, the UK Ministry of Defence's Defence Standard (Def Stan) 00-056 [22] provides a comprehensive structure for integrating safety engineering into defence equipment and services [23]. This standard encourages the use of civil (non-defence), open, or other recognised good practice standards and promotes a goal-based approach, justified through a robust safety case.

A widely referenced open standard is International Electrotechnical Commission (IEC) 61508 [24], the international standard for the functional safety of electrical, electronic, and programmable electronic (E/E/PE) safety-related systems. IEC 61508 is broadly adopted as a sector-agnostic safety lifecycle framework. It introduces Safety Integrity Levels (SILs), which set derived safety targets depending on the level of risk. As SILs increase, so does the evidence and assurance activities that are required, using both qualitative and quantitative methods for software and hardware.

The framework is often illustrated using a traditional 'V' lifecycle model. On the left side, activities include hazard and risk analysis and allocation of functional safety requirements. On the right side, activities focus on realisation through design, installation, validation, operation, and maintenance. Figure 1 shows this V model hierarchy as applied to a maritime vessel, while Figure 2 provides an example flowchart demonstrating how IEC 61508 is implemented across its various parts.

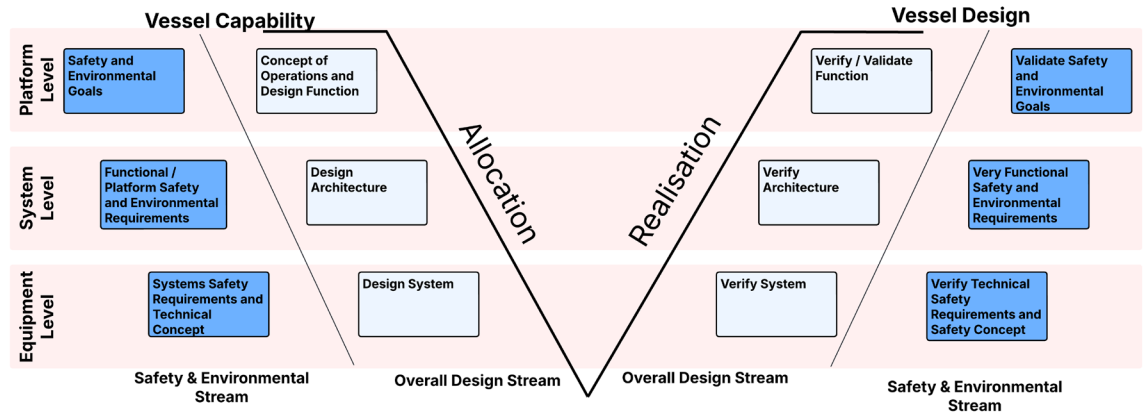

Figure 1: Systems V Model

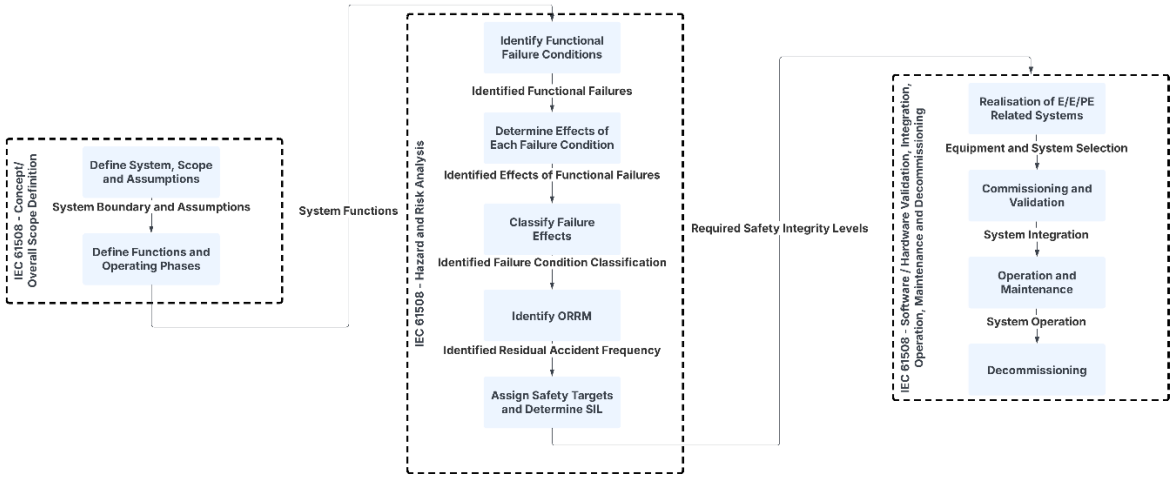

Figure 2: Example Flowchart Showing the Application of IEC 61508 (With Reference to Applicable IEC 61508 Parts)

However, traditional safety standards like IEC 61508 [24] were developed mainly for deterministic systems, where hazards are relatively stable and well understood. In contrast, MAS that incorporate AI and ML present new assurance challenges that these standards do not currently fully address [25, 26]. For example:

- Non-deterministic behaviours, continuous learning, and emergent properties make assurance significantly more complex;
- Traditional standards aren't suited to the agile development environment typically used for AI and ML systems; and,
- Unlike conventional systems, AI and ML based systems often involve complex hierarchies and iterative development across multiple organisations.

Despite these challenges, AI components and AI systems are increasingly used to perform safety-related functions in AI technology, replacing traditional software and hardware in maritime systems to enable true autonomy. For

example, an autonomous vessel might rely on an AI system to integrate radar and camera data to make real-time collision avoidance decisions. The perception and decision-making algorithms in these systems can vary significantly depending on training data, algorithm weights, environmental conditions, unexpected objects, or emergent behaviours [26].

2.2 EMERGING ASSURANCE APPROACHES FOR AUTONOMY

In recognition of the challenges posed by AI components and AI systems, new approaches have been developed, and existing approaches adapted, to better address the complexity and associated assurance requirements:

- The latest version of Def Stan 00-056 [22] does not address AI but acknowledges the need for modern approaches to safety engineering to address increasingly complex systems. It highlights systems engineering-based techniques, such as Systems-Theoretic Process Analysis (STPA), which can be used to analyse complex systems where potential failures may not be immediately obvious and may result in emergent hazards, such as for autonomous systems;
- The IEC is currently drafting a new international standard titled Functional Safety – Framework for Safety-Critical E/E/PE Systems for Defence Industry Application [27]. Whilst this framework will not address AI directly, it will build on IEC 61508 [24], extending it to apply to defence applications and incorporating systems engineering principles to manage safety aspects in highly complex systems, such as autonomous platforms [27];
- UK Joint Service Publication (JSP) 936 [28] has been introduced as the principal policy framework for the safe and responsible adoption of AI within the UK Ministry of Defence. This policy sets out high-level overarching requirements for governance, ethical considerations, and safety of AI to provide an initial step towards integrating AI assurance into defence procurement and operational practices;
- The Safety-Critical Systems Club (SCSC) has published guidance on the assurance of autonomous systems incorporating AI and ML [26]. This guidance proposes tailoring assurance activities to system criticality and highlights the importance of evidence-based safety arguments.

While these initiatives provide a comprehensive overarching framework for the assurance of complex and autonomous systems, there remains a need to formalise specific assurance methods and techniques for the individual AI components of the AI system. Practical and detailed methodologies for conducting such hazard and risk analyses and deriving qualitative and quantitative safety requirements for AI components and AI systems remains underdeveloped.

This gap has been recognised across both industry and academia. In response, the Assurance of Machine Learning for Autonomous Systems (AMLAS) framework, developed by the CfAA at the University of York [29], provides a structured approach for integrating ML assurance into systems engineering lifecycles through hazard analysis, data management, model verification and validation, and evidence generation to support assurance cases. Although AMLAS is applicable to potentially safety-critical contexts, its methods are still evolving and are not yet tailored to the highest levels of assurance demanded by highly safety-critical environments and high SILs, such as MAS. The current focus of AMLAS has largely been on mitigating risks associated with ML-specific failures, for example, unintended behaviours arising from poor data quality or inadequate training and Large Language Models (LLMs), rather than on addressing broader systematic risks that may arise across the lifecycle of a system in safety-critical domains. There is a need for domain-specific and supplementary guidance for safe implementation in safety-critical systems [30].

2.3 ADAPTING A FUNCTIONAL SAFETY APPROACH FOR AUTONOMY

In recognition of the challenges posed by integrating AI components and AI systems in safety-critical systems, the IEC published Technical Report (TR) 5469 [31] in 2024. This report provides guidance on incorporating AI components and AI systems into functional safety frameworks. Rather than replacing established standards, such as IEC 61508 [24], IEC TR 5469 extends them to enable the inclusion of AI-specific characteristics within traditional safety lifecycles.

IEC TR 5469 [31] aligns with AI lifecycle concepts described in IEC 5338 [32] (AI System Life Cycle) and IEC 23053 [33] (Framework for AI Systems Using ML), integrating these as informative references to guide the extension of established standards such as IEC 61508, and the associated lifecycle standards of IEC 12207 for software development [34] and IEC 15288 for system lifecycle processes [35]. As a result, it provides additional assurance methods and techniques to address the unique properties of AI components and AI systems.

It is important to note that IEC TR 5469 [31] does not replace the requirement to address random hardware faults and systematic faults as outlined in IEC 61508 [24]. Instead, it supplements these requirements by introducing aspects necessary for assuring AI technology. From this perspective, the assurance approach follows a traditional functional safety lifecycle which is adapted to include AI components and systems. Through this approach:

- The allocation phase (e.g., hazard and risk analysis and safety target allocation - the left side of the V-model, as shown in Figure 1) remain guided by IEC 61508 [24]; however,
- The realisation phase (e.g., verification and validation through realisation, installation, validation, operation, and maintenance - the right side of the V-model, as shown in Figure 1) is modified to account for AI components and AI systems.

An updated example flowchart illustrating how IEC TR 5469 [31] interacts with the various parts of IEC 61508 [24] is presented in Figure 3.

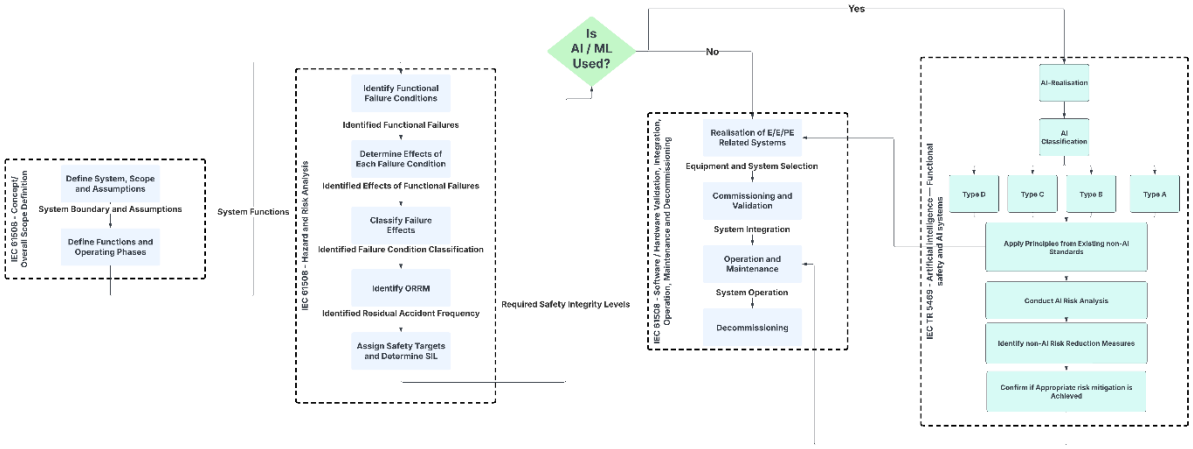

Figure 3: Example Flowchart Showing the Application of IEC 61508 / IEC TR 5469 (With Reference to Applicable IEC 61508 Parts and IEC TR 5469)

While IEC TR 5469 [31] is currently only a Technical Report and not a formal standard, the methods and techniques it outlines can be used to evaluate the suitability of AI components and systems for meeting safety targets derived from IEC 61508 [24]. This forms a necessary part of evidence to support the safety case for MAS that use AI. According to the Norwegian Research Centre for AI Innovation [36], IEC TR 5469 achieves this through:

- Describing the use of AI and ML in safety-related E/E/PE systems;
- Providing a classification scheme for the applicability of AI in safety-related E/E/PE systems;
- Explaining the properties of AI components and how they relate to safety within a functional safety context; and
- Detailing verification and validation methods and techniques tailored to AI and ML.

To support this, IEC TR 5469 aligns with the three-stage realisation principle of AI derived from IEC 22989 [20]:

1. Data acquisition - gathering representative, high-quality data;
2. Knowledge induction - learning from data, including model training;
3. Processing and output generation - producing decisions or actions from learned models.

This three-stage model provides a generalised process framework for AI. In some examples, this model has been further expanded into more detailed conceptual frameworks, such as the Sense-Understand-Decide-Act (SUDA) model proposed by the CfAA [18], but we choose not to do so in this work, to retain alignment with IEC TR 5469 [31]. The three-stage realisation principle from IEC 22989 [20] and the SUDA model are compared and shown to be compatible in Figure 4.

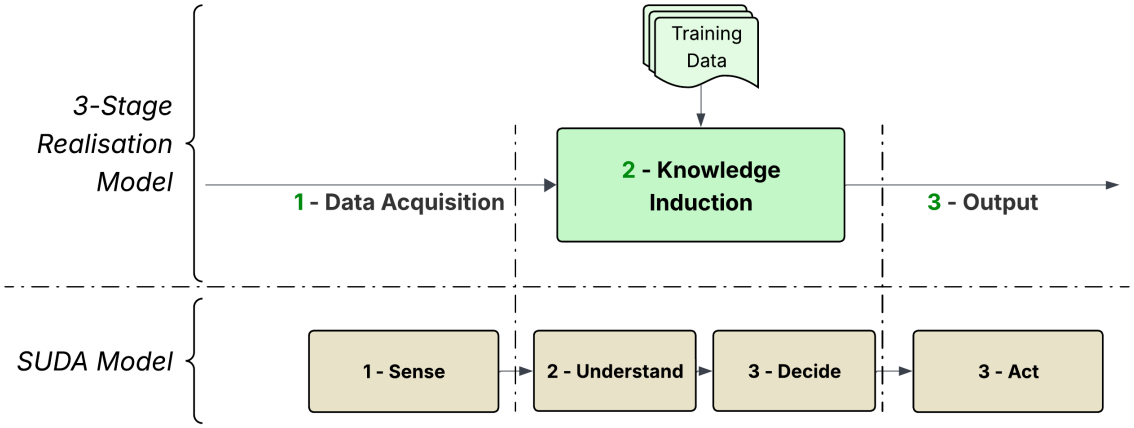

Figure 4:Three Stage Realisation Principle of AI / SUDA Model Comparison

3. APPLYING A FUNCTIONAL SAFETY APPROACH FOR AUTONOMY

3.1 INTRODUCTION

The three-stage realisation model can be used to derive acceptance criteria, assurance methods, and validation techniques for AI components and AI systems. This is achieved by considering:

- The desirable properties for each stage of the realisation model;
- The methods and techniques needed to achieve and verify those properties; and
- The acceptance criteria for assessing the adequacy of those methods and techniques.

An overarching 'AI safety' argument can then be constructed for the AI technology, covering each stage and demonstrating that all AI components and systems are safe for their intended use, through achieving the acceptance criteria for the desirable properties using the defined methods and techniques. This includes both quantitative (technical, numerical, performance-based) measures for AI elements and components but also qualitative measures at the organisational, governance, and system management levels in line with IEC 61508's [24] broader lifecycle framework fundamentals.

The rigour of the properties, methods, techniques and organisational arrangements increases in line with both:

- The safety targets (e.g., SIL's) derived from hazard and risk analysis on the left side of the V-model, guided by IEC 61508 [24]; and
- The AI technology 'Level' and 'Class', as defined in IEC TR 5469 [31].

3.2 AI TECHNOLOGY LEVEL

The AI technology 'Level' defines how AI is used and the extent of control within the system, as follows:

- Level A1 - The AI technology is used in a safety-relevant E/E/PE system, and automated decision-making of the system function using AI technology is possible;
- Level A2 - The AI technology is used in a safety-relevant E/E/PE system, but no automated decision-making of the system function using AI technology is possible; and,
- Level C - The AI technology is not part of the safety function but can have an indirect impact on system behaviour.

Additional levels (Level B1, B2) are defined where AI is used during the development of safety-relevant E/E/PE systems (e.g., as offline analysis or support tools) or where it is not part of the safety function and cannot have any impact on system behaviour (Level D). Given the focus on AI technology performing safety-related functions within MAS, these levels are not discussed further in this paper.

Because AI technology levels define the degree of control and influence AI technology exhibits, they can be directly mapped to frameworks such as the previously introduced DMR LoC, as illustrated in Figure 5.

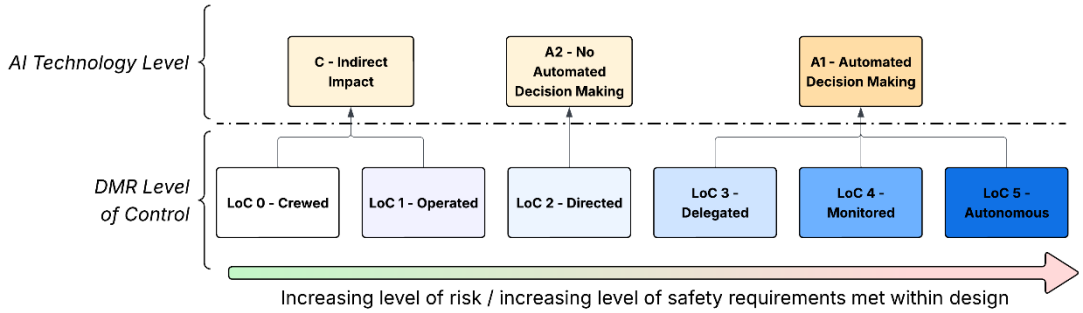

Figure 5: AI Technology Level / DMR Level of Control Mapping

3.3 AI TECHNOLOGY CLASS

The AI technology 'Class' categorises the feasibility of developing and assuring AI using existing functional safety standards, as follows:

- Class 1 - The AI technology is fully developed and reviewed using existing functional safety standards, such as IEC 61508 [24]. Established risk reduction concepts and techniques can be applied, leading to predictable and fully specified behaviour. This typically refers to 'traditional' AI technology, such as rule-based expert systems with no adaptive learning during operation;
- Class 2 - The AI technology cannot be fully developed and reviewed using existing functional safety standards alone. However, it is possible to identify additional complementary requirements, methods, and techniques (as outlined in IEC TR 5469 [31]). This typically applies to partially adaptive systems, such as ML-based perception modules, where risk can be managed through redundancy, operational restrictions, and fallback mechanisms;
- Class 3 - The AI technology cannot be adequately developed or reviewed using existing functional safety standards, and it is not possible to define sufficient complementary requirements, methods, or techniques to ensure acceptable risk reduction. This typically includes fully autonomous, adaptive, and opaque AI systems used in isolation, such as reinforcement learning (RL) controllers that learn online in open environments and continually adapt behaviour during operation.

Class 3 AI technologies are characterised by the absence of any currently known or accepted methods to achieve risk reduction to acceptable levels. Therefore, we would suggest they are not used in isolation in safety-related or safety-critical functions. This is not to say that Class 3 AI technology should be avoided altogether, rather efforts should be made to "reduce" the AI technology class through methods such as:

- Constraining the operational envelope to restrict online learning and adaptation in in-service environments;
- Introducing redundancy, where safety-critical decisions are overseen or validated by lower-classified (Class 1 or 2) systems or traditional E/E/PE; and
- Incorporating fallback mechanisms to ensure safe shutdown or transition to controlled degraded states in the event of unexpected behaviour (e.g., from LoC 5 to LoC 4 / 3).

3.4 DEMONSTRATING A FUNCTIONAL SAFETY APPROACH FOR AUTONOMY – QUANTITATIVE CONSIDERATIONS

To illustrate the practical application of a functional safety approach to maritime autonomy, we consider the example of an autonomous vessel's collision avoidance system, where tasks related to the three-stage realisation principle include perception (data acquisition), prediction (knowledge induction), and planning (processing and output generation) [37]. In this example, we focus on the prediction (knowledge induction) task, which is responsible for identifying objects to then generate a safe trajectory (through processing and output generation) that satisfies multiple objectives; advancing towards a destination, complying with navigational rules and avoiding collisions.

In this hypothetical example, we can assume that the system processes data, from the data acquisition stage, from onboard cameras using a perception algorithm based on a Deep Learning (DL) model, such as a Deep Neural Network (DNN) per the knowledge induction stage.

If we further assume that the system does not adapt online and that other inputs (e.g., radar) are utilised for redundancy and input fusion then, under IEC TR 5469 [31], this AI technology would be classified as Class 2. Based on this AI technology class and using the aforementioned definitions, we would apply established risk reduction concepts and techniques per IEC 61508 [24], supplemented by additional complementary requirements, methods, and techniques as outlined in IEC TR 5469 [31].

If the vessel operates in full autonomy (e.g., LoC 5), it would be categorised as AI Level A1, indicating automated decision-making without human intervention. Using the definitions of AI Level A1 and AI Technology Class 2, specific quantitative properties, methods and techniques, and acceptance criteria can be derived per IEC TR 5469 [31]. While full details are beyond the scope of this paper, a worked example focusing on the property of 'Specifiability' is presented in Table 1, following the format provided in IEC TR 5469.

Table 1. AI Technology Specifiability Properties, Methods and Acceptance Criteria

Stage	Desirable Property	Requirements	Properties	Details (Acceptance Criteria)	Methods for Compliance
2 – Knowledge Induction	Specifiability	Correctly identify diverse object types, unpredictable behaviours, and variable weather and sea states.	Specification of the Data Set	Amount of data (data set coverage)	- Manual creation - Active learning -Requirement traceability - Test plans - Compliance with IEC TR 24029 (Methods for Testing and Validating Robustness of AI/ML models)
			Type of Data	Object classes, object definitions, weather conditions, geographic domains, background environments	
			Specification of the labelling policy	Data annotation (labelling quality)	- Defined labelling guidelines - Random sampling and independent review of labels
			Number of Annotators	Use of independent annotators, random checks to confirm consistency and reduce bias	- Minimum number of annotators specified

3.5 DEMONSTRATING A FUNCTIONAL SAFETY APPROACH FOR AUTONOMY – QUALITATIVE CONSIDERATIONS

In addition to quantitative methods and techniques for individual AI elements and components, IEC TR 5469 [31] emphasises the importance of qualitative measures that address the organisational and managerial dimensions of AI safety, beyond just technical, quantitative validation of AI components. It formalises the requirement to establish a tailored AI management system, as specified in IEC 42001 [38], to ensure comprehensive oversight of governance, risk assessment, ethics and continuous learning throughout system operation aligned with the overall risk presented. To achieve this through the established management system, IEC TR 5469 [31] references, or signposts to, a set of qualitative standards and guidance documents to be followed for the example collision avoidance system at the identified Class and Level including:

- IEC 38507 (Guidelines for governance, oversight, accountability and transparency in AI systems) [39] to demonstrate that adequate organisational structures and decision-making frameworks are in place to oversee AI technology development;
- IEC 42005 (AI System Impact Assessment) [40] to address the non-functional risks associated with AI, including:
    - Model bias;
    - Inconsistent behaviour; and,
    - Fail-unsafe scenarios and potential unintended system-level effects.
- IEC TR 24028 (Trustworthiness in AI) [41] to demonstrate that AI systems are trustworthy, explainable, and reliable and that identified biases have been systematically mitigated.

IEC TR 5469 [31] highlights the importance of including structured feedback mechanisms such as post-deployment monitoring, incident reporting and stakeholder feedback loops within the operational framework. These mechanisms support the ongoing validation of safety assumptions and enable the system to adapt to emerging behaviours or changes in the environment. Their inclusion ensures that safety assurance evolves in response to operational experience, reinforcing the integrity of the safety case over time.

3.6 DEMONSTRATING A FUNCTIONAL SAFETY APPROACH FOR AUTONOMY – A TRANSITIONAL APPROACH

As highlighted earlier in this paper, most MAS are expected to operate in hybrid or transitional modes, blending full autonomy (LoC 5) with remote (LoC 2) or direct human control (LoC 1) at different stages of operation.

As illustrated in Figure 5, increasing the LoC towards full autonomy expands the operational risk envelope and reduces external, outside of design, risk controls, such as those provided by human intervention and external Other Risk Reduction Measures (ORRM). From an AI technology perspective, the required safety target will correspondingly increase when transitioning from AI Level A2 (human-authorised decision-making) to AI Level A1 (fully automated decision-making), which necessitates higher levels of technical assurance and greater rigour in both qualitative and quantitative safety measures.

Rather than attempting to adopt AI Level A1 or LoC 5 from the outset, the mapping of LoC to AI Levels (as shown in Figure 5) demonstrates how the level of assurance rigour of the AI technology can and therefore should incrementally increase in parallel with the level of autonomy. This staged approach enables the demonstration of an acceptable level of safety at each step, tailored to the level of risk presented at that stage of operational maturity.

By applying a functional safety approach, incorporating standards such as IEC 61508 [24] and IEC TR 5469 [31], both qualitative (e.g., governance, organisational readiness, management systems) and quantitative (e.g., verification, validation, robustness testing) methods can be progressively identified, demonstrated and strengthened. This supports not only immediate operational safety but also the development of a documented strategy and roadmap to achieve future acceptance criteria required for increased SIL, AI technology level and LoC as the MAS advances toward fully sustained autonomous operations.

In simple terms, this approach can be summarised using a “crawl, walk, run” strategy; vessels initially operate at low LoC and in well-defined operational envelopes, supported by robust external risk reduction measures for AI technology and conservative assumptions. As assurance evidence is generated and confidence in the AI technology increases, these constraints can be systematically relaxed, to expand the AI technology level and LoC, thereby enabling the gradual expansion of the operational scope of a vessel to become a MAS.

4. CONCLUSION

This paper has examined the integration of a functional safety approach to support the assurance of MAS, with a particular focus on adapting established standards such as IEC 61508 [24] supported by IEC TR 5469 [31]. Through the proposed mapping of LoC to AI technology levels, it is possible to incrementally demonstrate acceptable safety at each developmental and operational stage for the associated AI technology, including AI components and AI systems.

This staged strategy supports a "crawl, walk, run" transition model, enabling conservative initial operations within bounded risk envelopes and allowing for gradual expansion as confidence in AI technologies grows and reliance on ORRM decreases. Crucially, this paper highlights how assurance for maritime autonomy must encompass not only quantitative engineering measures but also qualitative organisational and governance dimensions of a lifecycle approach. In doing so, this approach does not propose abandoning existing goal-based safety case methodologies. Instead, it aims to provide a structured, standards-aligned evidence stream that addresses the unique challenges presented by AI and autonomy, thereby complementing and reinforcing the evidence within the overall safety cases for MAS.

4.1 RECOMMENDATIONS FOR FUTURE WORK

While this paper argues that the proposed functional safety approach can provide a robust foundation for AI assurance in maritime autonomy, further work is needed to fully integrate these methods within comprehensive safety case frameworks. What is clear from our work, is that there is much work involved in the assurance of MAS; the structured evidence stream presented here offers specific support for AI aspects of MAS, but it is not a

light touch. Given this, we recommend future work must focus on providing tailoring and practical guidance usable by both industry and regulators. This can be achieved by exploring how this AI technology focused evidence stream can be integrated with system-level analyses derived from STPA. In our view, while STPA can be used to support identifying hazards from complex interactions and emergent behaviours, it often lacks mechanisms to convert these insights into verifiable requirements for AI systems and components. The methods outlined in this paper are proposed to fill this gap by defining specific properties, methods, and acceptance criteria for AI technologies following the application of STPA, thereby complementing STPA outputs and strengthening the overall argument within safety cases.

Additionally, there is significant potential to further align an approach with the UK DMR's guide to the regulation of MAS. The DMR advocates a tailored, goal-based approach that supports gradual progression from lower to higher LoC levels. By combining a functional safety strategy for the supporting AI technology that scales the assurance rigour with increasing autonomy, and integrating it with STPA and systems-engineering approaches, industry can provide traceable, auditable evidence to meet DMR and emerging international regulatory expectations.

Finally, we identify that this further work should also focus on developing detailed guidance to unify these approaches into a single safety case framework tailored to MAS aligned with DMR policy, Defence Standard 00-056, and evolving international standards such as the forthcoming IMO MASS Code whilst building on academic frameworks. Extending these methods to develop bespoke assurance frameworks for Class 3 AI technologies, which currently lack robust verification pathways, is critical to support the maritime industries autonomous aims. This includes investigating adaptive operational constraints, real-time monitoring, fail-safe mechanisms, and hybrid assurance architectures that combine deterministic supervisory controls with adaptive, learning-based components for those online dynamic learning AI systems and components.

5. ACKNOWLEDGEMENTS

The authors wish to thank Dr Paul Hogan, Safeguard Engineering Ltd, for their contribution and support in the development of this work.

6. AUTHOR BIOGRAPHIES

Alexander Barker holds the current position of Principal Consultant at Safeguard Engineering Ltd. The author has a Masters in Safety, Risk and Reliability Engineering, is a Chartered Engineer and is an accredited Functional Safety Professional. He is currently responsible for leading safety, risk and reliability projects, with a particular focus on digital technology. The Author's previous experience includes safety and assurance for UK Ministry of Defence uncrewed surface and subsurface vessels, and commercial remotely operated and autonomous surface vessels.

Ian Groom MBE holds the current position of Principal Consultant at Safeguard Engineering Ltd. The author is a Chartered Marine Engineer responsible for technical support and advice to defence projects and programmes, with a particular focus in the maritime domain. The Author's previous experience includes an extensive career as a decision maker in the UK Royal Navy, latterly as the Defence Maritime Regulator (DMR) for the UK MOD, where he supported the registration of defence autonomous vessels.

Thayyab Hussain holds the current position of Graduate Consultant at Safeguard Engineering Ltd. The author has a Masters in Physics and is an accredited AI Risk Management Practitioner. He is currently responsible for supporting safety and environmental projects, with a particular focus on software safety, cybersecurity, AI risk management and legislative compliance.

7. REFERENCES

1. IMO, 'MSC.1/Circ.1638 – Regulatory Scoping Exercise for the Use of Maritime Autonomous Surface Ships (MASS)', *International Maritime Organization*, 2021.

2. SMI, 'Maritime Autonomous Ship Systems (MASS)', *UK Industry Conduct Principles and Code of Practice, Version 8*, November 2024.

3. Defence Safety Authority, DSA03-DMR, Guide to Regulation of Maritime Autonomous Systems, Version 3.1, August 2024.

4. RODSETH, O et al., 'A Criticism of Proposed Levels of Autonomy for MASS', *Proceedings of the 33rd European Safety and Reliability Conference* (ESREL 2023), 2023.

5. Lloyd's Register, 'Out of the box – Implementing Autonomy and Assuring AI', 2023.

6. Det Norske Veritas, 'Ensuring the Safety of Autonomous Shipping', 2025.

7. HOWARD, G., CHILCOTT, J., and PARKIN, P., 'Unmanned and autonomous maritime – the challenges of assurance', Conference Proceedings of INEC, 2020.

8. UK Government, 'The Strategic Defence Review 2025 – Making Britain Safer: Secure at home, strong abroad', 2025.

9. Department for Transport, Maritime 2050: Navigating the Future, UK Government, 2019.

10. Centre for Assuring Autonomy, University of York, Future of Autonomous Maritime Operations Report, 2023.

11. DE VOS, J., HEKKENBERG, R G., and VALDEZ BANDA, O A., 'The Impact of Autonomous Ships on Safety at Sea – A Statistical Analysis', *Reliability Engineering and System Safety*, 2021.

12. PAYNE, A., and STEHR, A., 'Is Regulation really the barrier? Exploring the Opportunities and Challenges in Certifying Maritime Systems with Increased Automation and Autonomy', *Conference Proceedings of INEC*, 2024.

13. IMO, 'MSC-FAL.1/Circ.3/Rev.1 - Guidelines on Maritime Cyber Risk Management', 2021.

14. DIERMET, S., MILLET., GROVES, J., and JOYCE, J., 'Safety Integrity Levels for Artificial Intelligence', *Critical Systems Labs Inc*., 2023.

15. IMO, Development of a Non-Mandatory MASS Code and Roadmap, IMO Documents MSC 107/5/1 and MSC 107/5/2, 2023.

16. LEVESON, N G., 'Engineering a Safer World: Systems Thinking Applied to Safety', *MIT Press*, 2012.

17. Ministry of Defence, Defence Standard 00-056: Safety Management Requirements for Defence Systems, UK MOD, 2022.

18. Centre for Assuring Autonomy, University of York, 'The BIG Argument: Assuring Safety of AI Systems', 2022.

19. Safety-Critical Systems Club, 'Guidance on the Assurance of Machine Learning in Safety-Related Applications', Version 2, 2021.

20. IEC 22989, Information Technology — Artificial Intelligence — Concepts and Terminology, 2022.

21. HOBBS, C., 'Embedded Software Development for Safety-Critical Systems', 2nd ed.

22. Ministry of Defence, Defence Standard 00-056: Safety Management Requirements for Defence Systems, UK MOD, 2022.

23. HAMILTON, V., 'A New Concept in Defence Safety Standards: The Revised UK Defence Standard 00-56' *ACS Workshop on Tools and Standards*, 2005.

24. IEC, IEC 61508: Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related

25. Systems, International Electrotechnical Commission, 2010.

26. ISO, ISO 21448: Safety of the Intended Functionality (SOTIF), International Organization for Standardization, 2019.

27. Safety-Critical Systems Club, 'SCSC Guidance on the Assurance of Autonomous Systems', Version 1, 2022.

28. INGE, J., et al., 'Engineering Safety into Complex Defence Systems', *Proceedings of the 2023 International Conference on System Safety*, 2023.

29. UK Ministry of Defence, JSP936 – Dependable Artificial Intelligence (AI) in Defence, Part 1 – Direction, 2024.

30. HABLI., et al., "The BIG Argument for AI Safety Cases", *Centre for Assuring Autonomy'*, 2025.

31. LAHER, et a., 'Review of the AMLAS Methodology for Application in Healthcare', 2022.

32. IEC, IEC TR 5469: Functional Safety – Guidance for AI and ML Elements, International Electrotechnical Commission, 2024.

33. IEC, IEC 5338, Information Technology – Artificial Intelligence – AI System Life Cycle Processes, 2023.

34. IEC, IEC 23053, Framework for Artificial Intelligence (AI) Systems Using Machine Learning.

35. IEC, IEC 12207, Systems and Software Engineering – Software Life Cycle Processes, 2017.

36. IEC, IEC 15288, Systems and Software Engineering – Systems Life Cycle Processes, 2023.

37. Norwegian Research Centre for AI Innovation, Review of AI Regulations and Governance, 2022.

38. Stanford University, "AI-Based Vehicle Motion Planning: Foundations and Challenges", Stanford AI Lab, 2022.

39. IEC, IEC 42001, Information Technology – Artificial Intelligence – Management System, 2023.

40. IEC, IEC 38507, Information Technology – Governance of IT – Governance Implication of the Use of AI by Organizations, 2022.

41. IEC, IEC 42005, Information Technology – AI System Impact Assessment, 2025

42. IEC, IEC TR 24028, Information Technology - Artificial Intelligence – Overview of Trustworthiness I Artificial Intelligence, 2020.